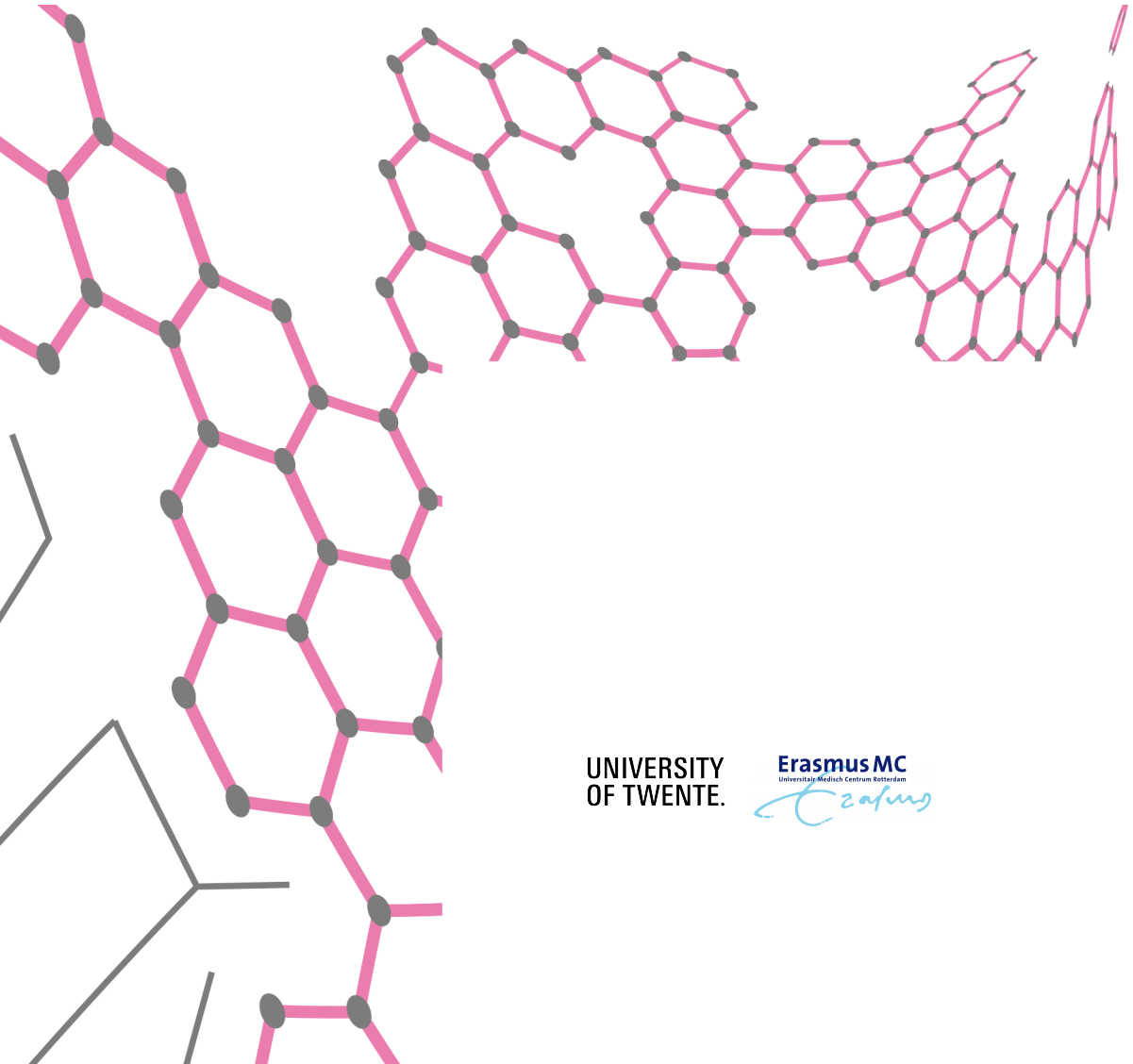
Predicting Survival and Recurrence-free Survival in GIST Patients Using Deep Learning

J.J. Slettenhaar

Supervised by:
D. Spaanderman, dr. M.P.A. Starmans, dr. S. Klein, dr. D.J. Grünhagen (Erasmus MC)
& dr. L. Alic (University of Twente)



UNIVERSITY OF TWENTE.

Erasmus MC
Universitair Medisch Centrum Rotterdam

1 Samenvatting

Inleiding Gastro-intestinale stromale tumoren (GISTs) ontwikkelen zich in het maag-darmkanaal. Ondanks chirurgische resectie en tyrosinekinaseremmers zoals Imatinib, varieert de 5-jaarsoverlevingskans tussen 30-65%. Nauwkeurige risicostratificatie en prognose zijn essentieel vanwege de variabele kwaadaardigheid van GISTs. Traditionele methoden zoals de NIH-classificatie gebruiken factoren zoals tumorgrootte en mitotic count, maar zijn beperkt omdat ze niet alle prognostische features bevatten. Computertomografie (CT) kan aanvullende features bevatten voor risicostratificatie. Geavanceerde technieken, zoals deep learning, bieden veelbelovende mogelijkheden om beeldvormings- en klinische gegevens te integreren om de prognose voorspelling te verbeteren. Wij onderzoeken of het gebruik van CT naast klinische gegevens de voorspelling van survival en recurrence-free survival (RFS) bij GIST patiënten kan verbeteren.

Methoden Patiënten van het Nederlandse GIST Consortium van vijf klinische centra in Nederland werden retrospectief geïncludeerd. Vier modellen werden ontwikkeld: twee op basis van klinische kenmerken, waarbij 1531 patiënten werden geïncludeerd, en twee op basis van beeldvormingsgegevens, waarbij 159 patiënten werden geïncludeerd. De klinische modellen omvatten een deep learning survival model met een Cox-gebaseerd feed-forward neural network en een Cox Proportional Hazard (CPH) model ter vergelijking. Beide modellen voorspelden zowel survival als RFS. Beide modellen gebruikten een minimale set van features met alleen klinische gegevens en een uitgebreide set van features die ook features op basis van pathologie omvatte. Hyperparameters van het deep learning model werden geoptimaliseerd met behulp van Optuna en de modellen werden geëvalueerd met de Concordance-index (C-index) met nested 5-fold cross-validatie. Hoge en lage risicogroepen werden gedefinieerd en vergeleken met een log-rank test. De interpreteerbaarheid werd beoordeeld met een feature exploratory analysis (CPH model) en Shapley analyse (deep learning model). De imaging modellen, een deep learning survival model en een classificatiemodel, maakten gebruik van een DenseNet 121 model architectuur. Het deep learning survival model voorspelde survival en RFS, terwijl het classificatiemodel patiënten indeelde in een lage of hoge risicogroep. Het imaging survival model werd geëvalueerd met de C-index en het classificatiemodel werd geëvalueerd met accuracy, area under the curve (AUC), sensitiviteit en specificiteit met behulp van 5-fold cross-validatie. De interpreteerbaarheid van het classificatiemodel werd beoordeeld met behulp van een Grad-CAM analyse.

Resultaten Het deep learning model op basis van klinische features (uitgebreide set, C-index: 0.64 (95% CI: 0.61 - 0.68)) behaalt een vergelijkbare performance als het CPH-model (uitgebreide set, C-index: 0.64 (95% CI: 0.60 - 0.70)) voor het voorspellen van survival. RFS voorspelling is ook vergelijkbaar tussen het deep learning (uitgebreide set, C-index: 0.62 (95% CI: 0.58 - 0.65)) en CPH model (uitgebreide set, C-index: 0.63 (95% CI: 0.60 - 0.66)). Er worden significante verschillen (P: <0.05) gevonden in survival kansen tussen lage en hoge risicogroepen voor beide modellen voor zowel survival als RFS voorspelling voor de uitgebreide feature set. Bij het CPH model zitten er echter meer patiënten (n_{laag}=92) in de lage risicogroep dan bij het deep learning model (n_{laag}=81) voor survival voorspelling. De feature exploratory analysis laat voor de uitgebreide feature set zien dat in het CPH-model leeftijd en mitotic count significant van invloed zijn op survival (P: <0.05). In het deep learning model worden deze factoren echter slechts gerangschikt als respectievelijk de 2e en 9e belangrijkste voorspellers van survival. Voor RFS voorspelling heeft de primaire tumorgrootte ook een significante invloed in het CPH model (P: <0.05), maar het is slechts de 5e belangrijkste factor in de RFS voorspelling van het deep learning model. Het deep learning imaging survival model heeft een C-index van 0.5, wat dat betekent de performance van het model gelijk is aan gokken. Hoewel de accuracy van het classificatiemodel hoog is voor zowel overlijden of recurrence (0.77 (95% CI: 0.74 - 0.79)) als alleen overlijden (0.89 (95% CI: 0.86 - 0.91)), zijn de bijbehorende AUC-waarden relatief laag, respectievelijk 0.59 (95% CI: 0.51 - 0.67) voor overlijden of recurrence en 0.57 (95% CI: 0.46 - 0.68) voor alleen overlijden. Het model heeft een specificiteit van 1.00 en sensitiviteit van 0.00 voor zowel overlijden als overlijden of recurrence. Grad-CAM toont aan dat het model zich focust op sommige abdominale regio's, maar er zijn ook hoge activaties op de achtergrond van de beelden.

Conclusie Deze studie laat de potentie zien van een deep learning model op basis van klinische kenmerken om survival en RFS te voorspellen, vergelijkbaar met traditionele methoden zoals het CPH model. Echter, verdere verbetering is nodig om het vermogen in risicostratificatie van het deep learning model te verhogen. Het gebruik van imaging bij prognosevoorspelling blijft uitdagend, maar het gebruik van grotere datasets en verbeterde modelarchitecturen kan helpen dit te verbeteren. Ook extra pre-processing, class balancing en data augmentatie kunnen bijdragen aan het verbeteren van de modelprestaties.

2 Abstract

Introduction Gastrointestinal stromal tumors (GISTs) develop in the gastrointestinal tract. Despite surgical resection and tyrosine kinase inhibitors (TKIs) like Imatinib, the 5-year survival rate ranges from 30-65%. Accurate risk stratification and prognosis are essential due to GISTs' variable malignant potential. Traditional methods like the NIH classification consider factors such as tumor size and mitotic count, but are limited since they do not include all prognostic features. Computed Tomography (CT) imaging may also provide additional features for risk stratification. Advanced approaches, including deep learning, offer promising avenues for integrating imaging and clinical data to enhance predictive accuracy. We hypothesize that using CT imaging alongside clinical data can improve the prediction of survival and recurrence-free survival (RFS) in GIST patients.

Methods Patients from the Dutch GIST Consortium from five clinical centers in the Netherlands were retrospectively included. Four models were developed: two based on clinical characteristics in which 1531 patients were included and two on imaging data in which 159 patients were included. The clinical models included a deep learning survival model with a Cox-based feed-forward neural network and a Cox Proportional Hazard (CPH) model for comparison both predicting survival and RFS. Both models used as input a minimal feature set containing only clinical features and an extensive feature set that additionally included features based on pathology. Hyperparameters of the deep learning model were optimized using Optuna and the models were evaluated using the Concordance-index (C-index) with nested 5-fold cross validation. High- and low-risk groups were defined and compared using a log-rank test. Interpretability was assessed with feature exploratory analysis (CPH model) and Shapley analysis (deep learning model). The imaging models, a deep learning survival model and a classification model employed a DenseNet 121 model architecture. The deep learning survival model predicted survival and RFS, while the classification model classified patients into a low- or high-risk group. The imaging data-based survival model was evaluated using the C-index, while the classification model was evaluated with the accuracy, area under the curve (AUC), sensitivity and specificity using 5-fold cross validation. Interpretability of the classification model was assessed using a Grad-CAM analysis.

Results The deep learning model based on clinical characteristics (extensive set, C-index: 0.64 (95% CI: 0.61 - 0.68)) achieves a similar performance to the CPH model (extensive set, C-index: 0.64 (95% CI: 0.60 - 0.70)) for predicting survival. RFS prediction is also comparable between deep learning (extensive set, C-index: 0.62 (95% CI: 0.58 - 0.65)) and CPH (extensive set, C-index: 0.63 (95% CI: 0.60 - 0.66)). Significant differences (P: <0.05) in survival probabilities are found between low and high-risk groups for both models for both survival as RFS prediction for the extensive feature set. However, the CPH model includes more patients ($n_{low}=92$) in the low-risk group than the deep learning model ($n_{low}=81$) for survival prediction. The interpretability analysis reveals for the extensive feature set that in the CPH model, age and mitotic count significantly influence survival (P: <0.05). However, in the deep learning model, these factors ranked only as the 2nd and 9th most important predictors of survival. For RFS prediction, primary tumor size also has a significant influence in the CPH model (P: <0.05), yet it is only the 5th most important factor in the deep learning model's RFS prediction. The deep learning imaging-based survival model has a C-index of 0.5, indicating chance-level performance. Although the accuracy of the classification model is high for both death or recurrence (0.77 (95% CI: 0.74 - 0.79)) and death (0.89 (95% CI: 0.86 - 0.91)), the corresponding AUC values are relatively low, at 0.59 (95% CI: 0.51 - 0.67) for death or recurrence and 0.57 (95% CI: 0.46 - 0.68) for death alone. The model exhibits a specificity of 1.00 and sensitivity of 0.00 for both death and death or recurrence. Interpretability analysis using Grad-CAM shows that the model focuses on some abdominal regions, but also exhibits high activations at the background of the images.

Conclusion In conclusion, our study demonstrates the potential of a deep learning model based on clinical features to predict survival and RFS, comparable to traditional methods like the CPH model. However, further refinement is needed to enhance the deep learning model's risk stratification capability. The use of imaging data in prognosis prediction remains challenging, highlighting the need for larger datasets and improved model architectures. Enhanced pre-processing, class balancing, and data augmentation techniques could be instrumental in improving model performance.

3 Acknowledgements

I would like to start by thanking my daily supervisor, Douwe Spaanderman from the Erasmus MC, for his great guidance and support throughout the duration of this project. Your willingness to always answer my questions, no matter how trivial they seemed, has been greatly appreciated. You would always make the time to think along, which really helped me.

I would also like to thank dr. Martijn Starmans, my second supervisor from the Erasmus MC, whose insightful feedback and criticism have greatly contributed to the development and execution of this project. Your input has helped me a lot and taught me to always be critical on my own work. I would also like to thank dr. Dirk Grünhagen for his help, especially for his valuable clinical knowledge. In addition, I would like to thank dr. Lejla Alic, my supervisor from the University of Twente, for her expertise and guidance throughout this project. Thank you for making it possible to do my master assignment externally at the Erasmus MC, this was something I really appreciated and enjoyed! Furthermore, I would also like to thank dr. Jelmer Wolterink and prof. Bernard ten Haken for their willingness to be part of the master's assignment committee. Your time and effort in evaluating my work are greatly appreciated.

Lastly, I am grateful to all individuals, colleagues, friends, and family members who have supported me while writing this thesis. The many coffee breaks and interesting discussions or just support really helped and provided some much-needed motivation and encouragement.

Contents

4 Introduction

Gastrointestinal stromal tumors (GISTs) are a type of tumor that can develop in the gastrointestinal (GI) tract, which includes the stomach, intestines, and other digestive organs. The 5-year survival rate of patients with GISTs after complete resection of the tumor is only 30 - 65% [1]. Therefore, understanding GISTs is important for accurate prognosis and effective treatment strategies.

GISTs can originate from various anatomical sites, with the stomach (60%), small intestine (25%), rectum (5%), esophagus (2%) and other locations (5%) such as the appendix, gallbladder, pancreas, mesentery, omentum, and retroperitoneum [2]. An annual global incidence rate is estimated to vary from 4.3 to 22 cases per million, with significant discrepancies across different countries [3]. Sporadic GISTs are more common than familial GISTs, which involve a germline mutation in the KIT gene [4].

The treatment approach for GISTs confirmed through immunohistochemistry on a biospy includes surgical removal and tyrosine kinase inhibitors (TKIs) [5]. Chemo- and radiation therapy rarely play a role in the treatment of GISTs, since they are generally ineffective [5]. Despite surgery, a study showed that patients with localized GISTs who underwent complete resection demonstrated a 5-year recurrence-free survival (RFS) rate of 63% [6]. The introduction of Imatinib, the first-line TKI, has transformed GIST management [5, 6]. Imatinib has increased 1-year RFS by 15% after surgery and extended overall survival in cases where tumors are metastatic or unresectable [6]. Adjuvant therapy, additional therapy given after primary treatment, with Imatinib for three years is recommended for patients with high-risk GISTs who have undergone complete tumor resection (R0 and R1), enhancing both overall survival and RFS [5]. Thus, performing accurate risk classification and prognosis prediction is crucial for optimizing treatment strategies and improving patient outcomes.

The development of GISTs is influenced by various factors, leading to a spectrum of malignant potential, ranging from virtually benign tumors to aggressive sarcomas [7]. Some patients face a significantly high risk of tumor recurrence and metastasis even after complete tumor removal [7]. The prognosis of GIST patients is commonly stratified according to the National Institutes of Health (NIH) consensus classification system, which considers tumor size and mitotic count as crucial parameters [7] (see Table 1 [3]).

Table 1: NIH criteria for risk assessment GIST. HPF, high power field. [3]

Risk	**Tumor size (cm)**	**Mitotic count (HPF)**
Very low risk	<2	<5/50
Low risk	2.1 - 5	<5/50
Intermediate risk	<5	6 - 10/50
	5 - 10	<5/50
High risk	>5	>5/50
	>10	Any mitotic count

Although initially based on consensus rather than clinical data, the NIH classification has demonstrated significant prognostic value. Studies analyzing small GIST cohorts have affirmed the association between large tumor size, high mitotic rate, and poor prognosis post-surgical removal [7]. However, the NIH classification has limitations. It does not consider other prognostic factors [7] like tumor site and tumor rupture, and lacks specific guidelines for determining mitotic count or tumor size measurement. Additionally, it fails to define the classification method when a mitotic count falls precisely on the boundary between two risk categories [7].

Several other prognostic factors have been identified [7]. Patients displaying symptoms tend to experience worse outcomes compared to asymptomatic counterparts. Tumor histology also plays a significant role; patients with GISTs featuring mixed spindle cell/epithelioid cell or pure epithelioid cell morphology exhibit poorer 5-year RFS than those with pure spindle cell histology. Other factors like high cellularity, tumor ulceration, and mucosal invasion have been associated with adverse outcomes in smaller patient groups [7]. Moreover, specific types of mutations can influence outcomes [7]. For example, the presence of KIT mutations independently predicts poorer 5-year RFS after surgical removal of localized GISTs [7]. Several tumor histopathologic and biologic factors have been explored as prognostic indicators, such as tumor necrosis, cell

atypia, expression of various markers, growth factors, and genetic factors [7]. More in depth information on prognostic factors and clinical background is given in chapter 5. Determining the significance of these factors in patient outcomes is challenging due to small study cohorts and the exploratory, retrospective nature of the analyses [7]. In most cases, these factors were associated with other high-risk characteristics, reducing sensitivity in detecting their independent prognostic value [7]. Nevertheless, since the introduction of the classification system in 2002, additions have been suggested, including anatomical location and tumor rupture during surgery [8].

Computed Tomography (CT) is the primary imaging modality for GISTs, used for initial diagnosis, surgical planning, postsurgical surveillance, and monitoring therapy response. CT provides morphological details that help predict high-grade GISTs and poor prognosis [8]. For GISTs, critical risk stratification factors include tumor size and rupture. However, other significant CT features also require evaluation. A study conducted by Zhou et al. in 2015 [9] highlighted the importance of considering not only tumor size but also growth patterns and the presence of enlarged vessels feeding or draining the mass (EVFDM) in risk stratification. GISTs displaying large size, a mixed growth pattern, or EVFDM are indicative of high-risk cases [9]. Studies focused on feature selection for risk stratification typically compare the chosen features with the established NIH criteria [9]. Consequently, the importance of CT imaging features, particularly size, in risk stratification is well-supported [9]. A comprehensive model incorporating multiple CT imaging features could enhance prognosis predictions and provide a more detailed perspective on GIST risk stratification.

Machine learning represent advanced approaches to the analysis of medical imaging data, offering significant potential in the risk stratification of GISTs. More in depth information on these methods is given in chapter 6. Radiomics involves extracting a large number of quantitative features from medical images, such as texture, shape, and intensity, to characterize tumors and predict outcomes. These features are then analyzed using statistical methods and machine learning algorithms to develop predictive models [10]. The detailed feature extraction process allows for an analysis of the tumor's characteristics, which can be crucial in assessing risk and treatment response. Deep learning, a subset of machine learning, involves training neural networks to automatically learn features from raw data. Unlike traditional radiomics, which relies on extracting pre-defined features from images, deep learning models learn to identify patterns directly from the images, potentially capturing more complex and subtle variations associated with different risk levels [11]. Although both radiomics and deep learning model can leverage large datasets, deep learning has a distinct ability to automatically learn and optimize feature representations directly from the raw image data, often leading to superior performance in capturing complex patterns improving accuracy and performance of the model [11].

The application of these technologies in GIST risk stratification is promising. For instance, a multicenter study demonstrated the effectiveness of a quantitative CT-based deep learning approach in accurately predicting GIST risk classification based on imaging data [12]. This indicates the potential for these advanced methods to provide more precise and comprehensive tools for clinical decision-making.

Because of the complex nature and varied prognostic factors associated with GISTs, there is a growing need for precise risk stratification and prognostic modeling to guide clinical decision-making and improve patient outcomes. While traditional methods, such as the NIH consensus classification system, have provided valuable insights, the integration of imaging and molecular biomarkers holds promise for enhancing predictive accuracy. Therefore it would be useful to develop a deep learning model utilizing both imaging data and clinical features. We hypothesize that using CT imaging alongside clinical data, such as age, mitotic count, and mutation status, can improve the prediction of survival and RFS in GIST patients.

5 Clinical background

Several prognostic clinical features for GISTs have been identified. In this section, the clinical background of GISTs and some prognostic features, specifically mutations will be discussed more in depth.

5.1 Historic overview

GISTs were initially thought to originate from mesenchymal cells or smooth muscle in the GI tract [2, 3]. However, in the 1980, immunohistochemistry revealed that many GISTs lacked typical smooth muscle markers like actin and desmin, but expressed antigens associated with neural crest cells [2, 6]. Mazur et al. in 1983 and Schaldenbrand et al. in 1984, introduced the term "stromal tumors" to describe these distinct tumors in the GI tract [6]. Despite this, confusion about their origin and behavior persisted [6]. A breakthrough in 1994 showed that most stromal tumors were CD34 positive, suggesting they were related to the interstitial cells of Cajal (ICC), which are spindle cells in the gut wall involved in peristalsis [2, 6]. ICCs express KIT, a receptor tyrosine kinase also known as CD117 or SCFR, which is strongly expressed in most GISTs, supporting the idea that GISTs arise from or share common stem cells with ICCs [2]. Additionally, a mutation in the KIT proto-oncogene's intracellular domain was identified [6]. This discovery significantly advanced the understanding of GIST biology and cancer research [6].

5.2 Symptoms

GISTs are typically asymptomatic until they increase in size. When clinical signs emerge, they may include nonspecific symptoms such as pain, fatigue, nausea, dyspepsia, weight loss, fever, and obstruction [6]. The size of the tumor, which can range from a few millimeters to up to 40 cm, is a determining factor in the appearance of clinical symptoms [6].

5.3 Epidemiology

Determining the incidence of GISTs has been challenging due to the lack of a consistent definition until the late 1990s. Estimates suggest annual global incidence rates vary from 4.3 to 22 cases per million, with significant discrepancies across different countries [3]. High incidence rates are reported in regions such as Hong Kong, Shanghai, Korea, and Norway (19 - 22 cases per million), while lower rates are noted in Shanxi province, the Czech Republic, Slovakia, and North America (4.3 - 6.8 cases per million) [13].

The prevalence of GISTs is difficult to determine as they are often incidentally discovered during radiologic imaging, autopsies, or gastrectomies for unrelated issues [6]. GISTs can occur in individuals from their teenage years to their 90s, with most diagnoses in people in their 60s [2]. Although there is no clear gender preference, some studies suggest a slight male predominance. GISTs in children and young adults, while rare, show distinct characteristics compared to adult cases, including a higher prevalence in females [6].

Sporadic GISTs are more common than familial GISTs, which involve a germline mutation in the KIT gene. Familial GISTs are rare but well-studied, typically presenting with multiple GISTs and cutaneous hyperpigmentation. GISTs are seldom associated with other syndromes and rarely co-occur with other tumors [4]. When they do, it is usually with colorectal carcinomas or adenomas, with documented instances of GISTs colliding with other tumors being exceptionally rare [6].

5.4 Treatment of GISTs

The treatment approach for GISTs that are confirmed through immunohistochemistry on a biopsy comprises surgical removal and TKIs [5]. Chemo- and radiation therapy rarely play a role in the treatment of GISTs [5]. Radiation therapy may occasionally be used for palliation of painful metastases or for patients with unresectable bleeding tumors [14].

5.4.1 Surgical Removal

Complete surgical removal remains the principal curative intervention for GISTs, particularly when the tumor is locally or marginally resectable. The goal of surgery is to achieve R0 resection, meaning no microscopic traces of the tumor remain at the surgical margins. Lymph node involvement is rare in GISTs,

rendering regional lymph node dissection generally unnecessary [6]. Organ-sparing resections, such as segmental resection, are considered oncologically appropriate [6]. Despite surgery, a study showed that patients with localized GISTs who underwent complete resection demonstrated a 5-year RFS rate of 63% [6]. Laparoscopic resection, especially for gastric GISTs, is feasible, safe, and less invasive than traditional open surgery, with comparable oncological outcomes. Recent clinical outcomes have highlighted the effectiveness of alternative minimally invasive procedures, including submucosal tunneling endoscopic resection, endoscopic full-thickness resection, and laparoscopic endoscopic cooperative surgery [5, 6].

5.4.2 Tyrosine Kinase Inhibitors (TKIs)

The introduction of Imatinib, the first-line TKI, has transformed GIST management [5, 6]. Imatinib has increased 1-year RFS by 15% after surgery and extended overall survival in cases where tumors are metastatic or unresectable [6]. Adjuvant therapy, additional therapy given after primary treatment, with Imatinib for three years is recommended for patients with high-risk GISTs who have undergone complete tumor resection (R0 and R1), enhancing both overall survival and RFS. In cases where Imatinib treatment fails, second-line TKI Sunitinib and third-line multikinase inhibitor Regorafenib can be employed for advanced GISTs [5]. Achieving a permanent cure with TKI alone is challenging. Therefore, early diagnosis and timely surgical resection remain the most promising approaches for a complete cure [5].

5.5 Imaging of GISTs

Several imaging modalities are employed in the identification and evaluation of GISTs, including abdominal ultrasound, computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography (PET) [3]. These modalities serve multiple purposes, including tissue sampling, assessing the tumor's local extent, predicting the stage, and other diagnostic objectives [8].

CT is the primary imaging modality for GISTs, used for initial diagnosis, surgical planning, postsurgical surveillance, and monitoring therapy response. CT provides morphological details that help predict high-grade GISTs and poor prognosis [8]. Although MRI is not typically the first choice for diagnosing GISTs, it offers additional information compared to CT. MRI provides morphological imaging similar to CT but also includes qualitative parameters like the apparent diffusion coefficient (ADC) and perfusion parameters useful in assessing malignancy and treatment response [8]. An example of a GIST tumor at the duodenal-jejunal junction imaged with both CT and MRI is shown in Figure 1 [15].

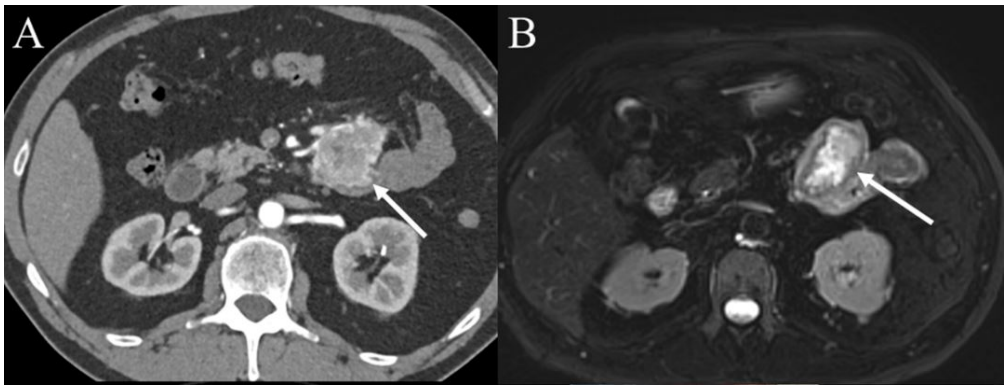


Figure 1: 66 year old man with GIST of the duodenal-jejunal junction. A shows a CT image in the axial plane during arterial phase; B shows a T2-weighted MR image in the axial plane. [15]

Other imaging techniques, such as Fluorodeoxyglucose-PET (FDG-PET), are used for GIST visualization and aid in distinguishing GISTs from non-GISTs, stratifying histopathological risk, evaluating initial disease, and monitoring therapy response. Combining FDG-PET with CT enhances accuracy in response prediction, especially for FDG-avid tumors [8]. Endoscopic examination is another method used for GIST diagnosis [8]. However, endoscopy alone does not provide sufficient information for a differential diagnosis, making high-resolution tomographic imaging through Endoscopic Ultrasound (EUS) necessary. EUS determines the lesion's origin within the gastrointestinal wall, its nature, and size. GISTs typically appear as hypoechoic

solid masses on EUS. High-risk features on EUS include size, irregular borders, heterogeneous echo patterns, and growth during follow-up. Predicting malignancy risk for GISTs smaller than 5 cm requires tissue sampling via EUS-FNA or biopsy [5].

5.6 Histology

Microscopically, GISTs exhibit diverse morphological characteristics and three primary histological subtypes have been identified. These subtypes comprise the spindle cell type (accounting for 70% of cases), the epithelioid type (observed in 20-25% of cases), and a mixed type. Notably, GISTs display considerable variability, ranging from sparsely cellular to highly cellular structures, often characterized by elevated mitotic rates [6]. Figure 2 [3] illustrates the distinctive features of these subtypes. Spindle cell GISTs are characterized by elongated, slender, or fusiform-shaped cells, whereas epithelioid cell GISTs consist of round or polygonal cells. Mixed-type GISTs comprise both spindle and epithelioid components [8].

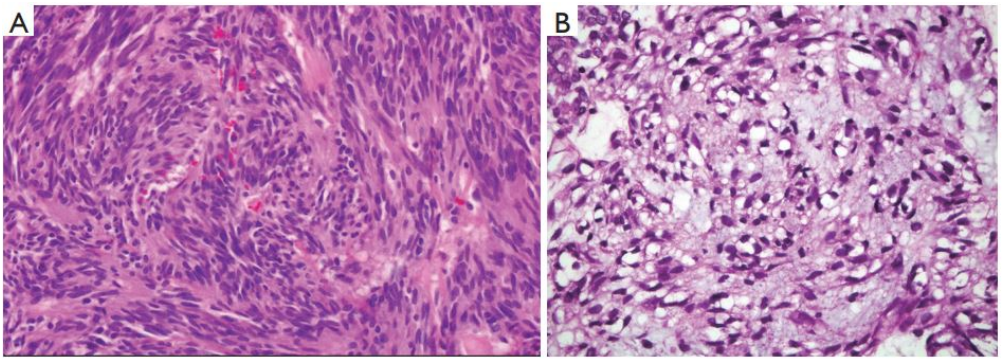


Figure 2: Microscopic photo with HE stain. A shows 40× objective magnification image of a GIST spindle cell type; B shows 100× objective magnification image of a GIST epithelioid type. [3]

5.7 KIT

The KIT receptor, a member of the type 3 tyrosine kinase receptor family, plays a pivotal role in cellular functions such as proliferation, apoptosis, chemotaxis, and adhesion [16]. Activation of KIT occurs upon binding of its ligand, stem cell factor, to its extracellular domain, leading to downstream signaling cascades. This receptor is crucial for the development and maintenance of various cells, including ICCs. In 1998, it was discovered that gain-of-function mutations in KIT serve as key oncogenic drivers in the majority of GISTs [2]. These mutations, present in 80% of cases, result in constitutive, ligand-independent activation of the KIT receptor and its pathways. Consequently, this leads to increased cell proliferation and inhibition of apoptosis. It is essential to emphasize that the detection of KIT by immunohistochemistry does not correlate with the presence of underlying KIT mutations. To determine the presence of mutations, mutational analysis of known genes such as KIT is necessary. This analysis is critical in selecting appropriate therapy and confirming the diagnosis, especially in cases of suspected GISTs that do not stain positive for CD117 or DOG1, another key diagnostic marker [16].

The type of mutation in KIT influences the risk grade of GISTs. The identification of these activating mutations in KIT has provided a therapeutic target for the treatment of GISTs [6]. Therapies aimed at inhibiting KIT have been developed, such as Imatinib, which occupies the ATP binding pocket of KIT, preventing phosphorylation of substrates and downstream signaling. Consequently, this inhibition hampers cell proliferation and promotes apoptosis [6].

Mutations in KIT also influence the response rates to Imatinib. Clinical studies of Imatinib for GISTs treatment have consistently demonstrated varied outcomes based on genotypically defined GIST subsets. Response rates to Imatinib differ for different KIT mutations, and the likelihood of primary resistance to Imatinib is also influenced by KIT mutations [17]. In patients in which the tumors continue to grow within the first six months, primary resistance occurs which is most seen in patients with exon 9 mutant or wild-type tumors [17]. The molecular basis of resistance has been under active investigation. Studies comparing

biopsies from patients with primary and secondary resistance revealed that secondary resistance often involves new mutations in the KIT gene, particularly in the ATP-binding pocket or activation loop. These mutations confer resistance to Imatinib in laboratory experiments. Other causes of secondary resistance include KIT gene amplification and downregulation of KIT expression, leading to a KIT-independent phenotype. In cases of primary resistance, some respond to higher Imatinib doses, especially those with a KIT exon 9 mutation, while the mechanisms behind most primary resistance cases remain unknown [17].

5.7.1 Mutations in KIT and PDGFRA

Genetic mutations in GISTs primarily occur in KIT and PDGFRA (Platelet-Derived Growth Factor Receptor Alpha). PDGFRA is another potential genetic aberration in GISTs [18]. Both KIT and PDGFRA are receptor tyrosine kinases, integral proteins on cell surfaces that play pivotal roles in cellular signaling and regulation [18].

KIT and PDGFRA genes are situated on chromosome 4q12 and encode closely related transmembrane glycoproteins. These proteins share a specific molecular structure comprising an extracellular (EC) domain with five Ig-like loops responsible for ligand binding and dimerization. Additionally, they possess a cytoplasmic domain including a juxtamembrane (JM) region and a split tyrosine kinase (TK) domain, encompassing TK1 and TK2 [18]. Activation of KIT and PDGFRA occurs upon binding with their respective ligands, stem cell factor, and platelet-derived growth factors (PDGFs), leading to the regulation of essential cellular functions [18].

Mutations in KIT and PDGFRA predominantly affect the exons responsible for the functional domains of these Tyrosine Kinase Receptors (TKRs). These mutations can be classified by type, altered domain, and treatment implications. Possible mutation types include deletions, point mutations, internal tandem duplications, insertions, and complex mutations. Altered domains are categorized into dimerization or JM domains (regulatory mutations) and TK1 and TK2 domains (enzymatic mutations). Mutations detected before targeted treatment initiation are referred to as primary mutations, whereas those identified during targeted treatment, leading to acquired resistance to TKIs, are termed secondary mutations. Figure 3 [18] illustrates the distribution and frequency of mutations in KIT and PDGFRA [18].

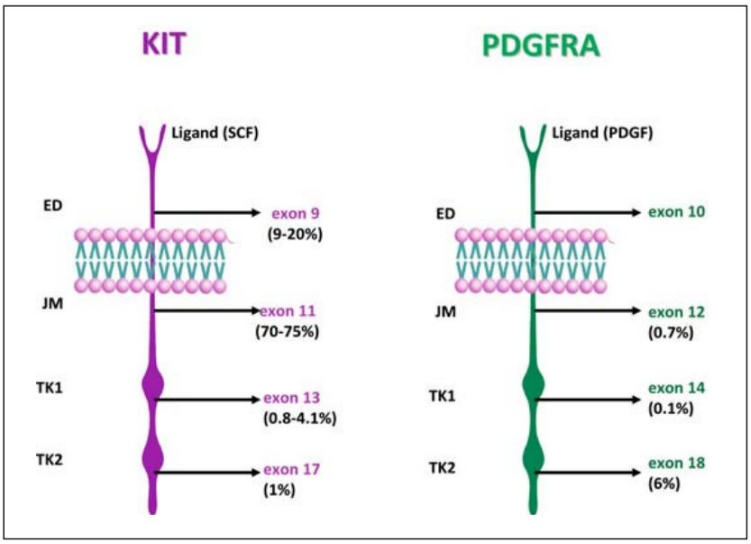


Figure 3: Distribution of mutations in KIT and PDGFRA in GISTs, with frequency of mutation indicated in brackets. [18]

Mutations in KIT predominantly impact exons 9, 11, 13, and 17, as illustrated in Figure 3. The juxtamembrane domain, encoded by exon 11, is the most commonly affected region, with approximately two-thirds of GISTs displaying alterations in this exon. These alterations include deletions, insertions, substitutions, or combinations thereof. Exon 9, responsible for the extracellular domain, is mutated in 9-20% of GIST cases. Mutations in the kinase I domain (exon 13) or kinase II domain (exon 17) occur less frequently [17].

Primary PDGFRA mutations are predominantly found in exons 12 and 18, with less common occurrences in exon 14 (Figure 3). Approximately 6% of GISTs exhibit PDGFRA mutations in exon 18, corresponding to exon 17 of KIT, which encodes the second kinase domain. The most prevalent mutation in this exon is the D842V substitution, identified in up to 75% of all PDGFRA-mutated tumors. Exon 12 mutations (Juxtamembrane domain) are the second most common, occurring in around 1–2% of GISTs. Mutations in exon 14 are exceptionally rare, accounting for less than 0.1% of cases [17].

5.8 Other expression markers and mutations

Currently, in pathology the expression of several other markers and mutations in other genes are examined that may also give an indication of the prognosis of GISTs. Besides CD117, there are other proteins that are expressed in GISTs. This is especially important when GISTs stain negative for GISTs which is approximately 5% of the GIST tumors [19].

5.8.1 DOG1 expression

DOG1, also known as Ano1/TMEM16A, is elevated in GIST cells, aiding in diagnosis and distinguishing them from similar tumors. Overexpression of DOG1 can alter the normal functioning of the cell membrane. Proteins like DOG1 often have specific roles in regulating cellular processes, so when their levels are abnormal, it can affect cell signaling and communication. While DOG1 is closely linked to GIST diagnosis, its widespread expression across tumor types may limit its prognostic value [19–21]. There is however some potential in using a combination of both CD117 and DOG1, since a study by Lopes et al. in 2010 [19] has shown that this can define the diagnosis of GIST in more than 99% of the cases.

5.8.2 SDHB expression and SDH mutation

SDHB is a part of the Succinate Dehydrogenase (SDH) enzyme complex, which plays a role in the mitochondrial electron transport chain and the Krebs cycle (citric acid cycle). In some GIST cases, especially those with a distinct morphology known as wild-type GISTs (not detectible KIT or PDGFRA mutations), loss of SDHB expression can occur [22]. This loss indicates dysfunctional SDH, often due to mutations in SDH genes, leading to a more aggressive clinical course and are associated with a higher risk of metastasis and recurrence [22]. This type of mutations are associated with a hereditary cancer syndrome known as Paraganglioma Syndrome (PGL) or Carney-Stratakis Syndrome. SDH deficiency results in the accumulation of HIF-1α, triggering pathways promoting tumor growth and inhibiting apoptosis [22].

5.8.3 BRAF mutation

BRAF mutations are relatively rare compared to other genetic mutations like those found in KIT, PDGFRA, or SDH genes. In SDH-deficient wild type (WT) GISTs, around 15% of cases have mutations in BRAF or RAS genes [23]. The BRAF gene encodes for a serine/threonine-protein kinase. Its function is to control proliferation and differentiation through the Ras-Raf-MAPK pathway [24]. BRAF mutations can lead to the production of a mutated BRAF protein that is abnormally activated, causing uncontrolled cell growth and division. The specific consequences of a BRAF mutation in GISTs can vary, but generally, it contributes to the development and progression of the tumor [24].

5.8.4 NTRK mutation

Neurotrophic Tyrosine Receptor Kinase (NTRK) genes, including NTRK1, NTRK2, and NTRK3, encode proteins that are involved in normal neural development and function, namely tropomyosin receptor kinase (TRK) A, B and C respectively [25]. These genes produce abnormal receptors that, when activated by specific growth factors called neurotrophins, trigger signaling pathways regulating cell growth, differentiation, and survival. Mutations (like for example gene fusions) involving NTRK genes can lead to constitutive activation of these receptors, causing abnormal cellular behaviors and potentially contributing to cancer development [25].

The discovery of NTRK gene fusions in various cancers, including GISTs, has led to the development of targeted therapies known as TRK inhibitors. Drugs like larotrectinib and entrectinib have been designed to specifically inhibit the activity of abnormal TRK proteins. These targeted therapies have shown remarkable

efficacy in patients with NTRK fusion-positive cancers, often leading to significant tumor shrinkage and improved outcomes [25].

5.9 Genetic classification of GISTs

In summary, there are several genetic and molecular markers of GISTs that can help make a differential diagnosis. Twenty years after the discovery of KIT mutations in GISTs, the diagnosis of GISTs significantly improved through the use of immunohistochemical (IHC) panels (CD117/DOG1) and molecular analysis (KIT/PDGFRA). These methods have become the gold standard for GIST diagnosis, enabling precise identification and guiding clinicians toward personalized treatments. The entire genetic classification system is shown in Figure 4 [26].

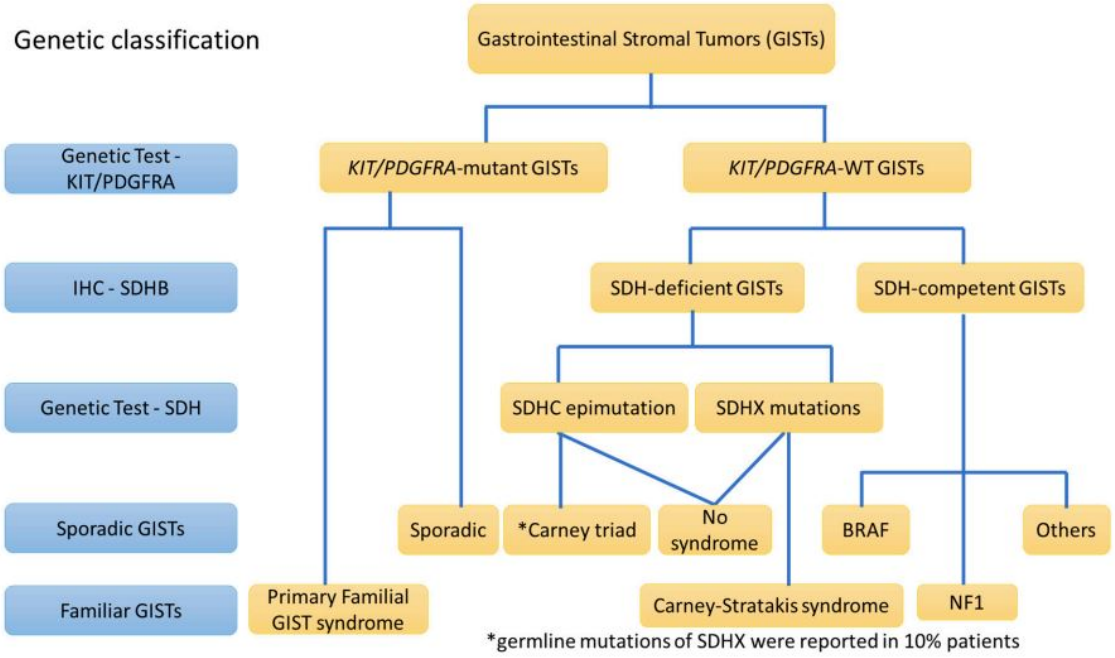


Figure 4: Genetic classification of GIST. GISTs are categorized genetically into two main groups: those with mutations in KIT or PDGFRA genes, and those without (WT KIT/PDGFRA GISTs). Within the latter group, WT KIT/PDGFRA GISTs are further classified based on succinate dehydrogenase (SDH) expression into SDH-deficient and SDH-competent GISTs. While most mutations are random (sporadic), some are inherited (germline mutations), which are associated with syndromic or familial forms of GISTs. [26]

6 Technical background

Imaging can be used in risk stratification of GISTs. In both radiomics and deep learning algorithms, imaging features are used. In this section, radiomics and deep learning are explained more in depth.

6.1 Radiomics

In radiomics, the use of quantitative imaging features is combined with machine learning [10]. Studies have shown that there is much potential for radiomics in several areas, like lung, liver, brain and sarcomas [27]. The imaging modality that is used also varies, like MRI, CT and PET. Radiomics has been used to predict several clinical outcomes, like survival, therapy response and genetic mutations [10]. An overview of the common workflow for radiomics is shown in Figure 5 [27]. The radiomics workflow typically starts with image acquisition, collecting standardized medical images. Image segmentation follows, identifying regions of interest (ROIs). In feature extraction, quantitative features like shape, intensity, and texture are derived from ROIs. Feature selection then picks the most relevant features for model building, where predictive models are created using statistical or machine learning methods [27].

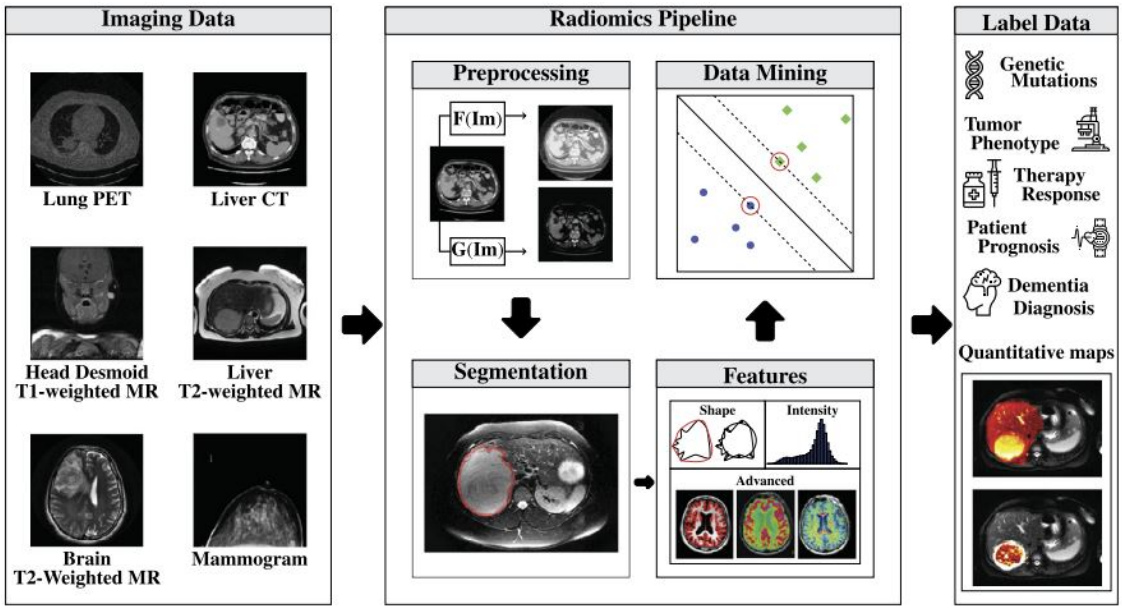


Figure 5: A common radiomics workflow: an overview. This workflow typically starts with images acquisition, followed by image segmentation. Afterwards, features are extracted and selected, which are then used in the model. [27]

A feature can be described as any piece of information which can be relevant for solving the task at hand. Imaging features can be subdivided into three different groups [27]:

- *Morphological features*: Morphological features are features used to describe the shape of an object.
- *First order features*: First order features in image analysis refer to statistical measures that describe the distribution of gray levels or intensities. These features include common observations made by clinicians such as mean and maximum intensity, but also more complex statistics like intensity skewness.
- *Higher order features*: Higher order features are employed to identify distinct patterns within images, like stripes.

6.1.1 Radiomics in GIST diagnosis and prognosis

Some research has been conducted in the field of radiomics and GIST diagnosis and risk stratification. Studies [11] have demonstrated a heterogeneous methodology regarding types of radiomics features, and analysis techniques. Since an accurate diagnosis and risk stratifaction of GISTs is however crucial for determination

of the appropriate treatment, there is much need for a reliable model [11].

There are several advantages to radiomics models for GISTs. First of all, radiomics models allow for quantitative evaluation of tumor characteristics [28]. They provide objective measurements that can aid in diagnosis and treatment planning. Secondly, radiomics models are non-invasive and rely on imaging techniques like MRI and CT which are routinely used in clinical practice [11]. Furthermore, radiomics has the potential to identify specific imaging biomarkers that can predict prognosis, treatment response, and guide personalized treatment strategies for GIST patients [11].

On the other hand, radiomics methods for GISTs also have several disadvantages [11]. There is a lack of standardization in radiomics methodology, including imaging acquisition, feature extraction, and radiomics software. This makes it challenging to compare results between studies and apply radiomics models in different populations. In addition, most of the current published studies on radiomics of GISTs are retrospective and performed in single centers [11]. Furthermore, while some studies have included external validation cohorts, the validation of radiomics models in independent cohorts is still limited [11].

In summary, radiomics methods for GISTs offer the advantages of quantitative assessment, non-invasiveness, and potential for personalized medicine. However, the lack of standardization, limited validation, and reliance on specific imaging modalities are some of the disadvantages that need to be addressed for the wider application of radiomics in clinical practice [11].

6.2 Machine and deep learning

Machine learning involves training computers to learn from data and improve their performance on specific tasks over time using various algorithms and techniques [29]. Artificial neural networks (ANNs) are vital components of machine learning algorithms. They are inspired by biological systems and consist of interconnected neurons. ANNs adapt their structure to perform different machine learning tasks. ANNs are organized into layers: an input layer receives data, hidden layers learn non-linear mappings, and an output layer produces results. Hyperparameters such as the number of layers and neurons are set manually [29, 30].

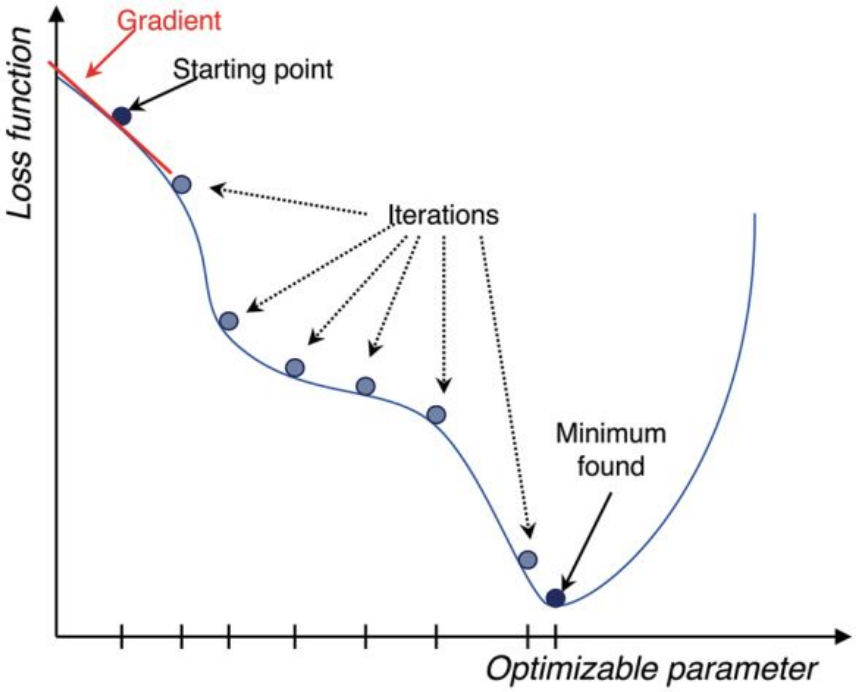


Figure 6: The gradient descent method. [31]

Deep neural networks (DNNs) are a type of ANN with multiple hidden layers. They excel at automatically extracting hierarchical features from data. This makes them ideal for processing complex, high-dimensional

data because their layered structure allows them to capture intricate patterns and relationships [29]. Predicting outcomes using neural networks involves sequential computation of node activations in each layer. Neural networks are refined through parameter adjustments, including weights and biases. Parameters are typically optimized using gradient descent (see Figure 6 [31]), a method that iteratively adjusts parameters to minimize the difference between predicted and actual outcomes, as measured by a loss function [31]. Other optimization methods, such as evolutionary algorithms and genetic algorithms, can also be used. In addition to weights and biases, other parameters such as dropout rates, regularization parameters, and batch normalization parameters may also be optimized. After multiple iterations over each sample in the training dataset, the parameters converge toward values that optimize the model's accuracy [31].

6.2.1 Convolutional neural network

The most popular approach in deep learning for imaging is to use a Convolutional Neural Network (CNN), which is a subset of DNNs. CNNs use convolutional layers to scan input data, like images. The convolution operation involves sliding a small filter (also known as a kernel) over the input image, producing a feature map that highlights the presence of certain features or patterns [32].

After each convolution layer, the feature maps are typically passed through activation functions to introduce non-linearity [32]. This allows the CNN to learn more complex relationships between the features. The output of the convolutional layers is then flattened and fed into fully connected layers, which perform the final classification or regression task. These fully connected layers are similar to traditional neural networks and can learn to map the extracted features to the desired output. This structure is shown in Figure 7 [33].

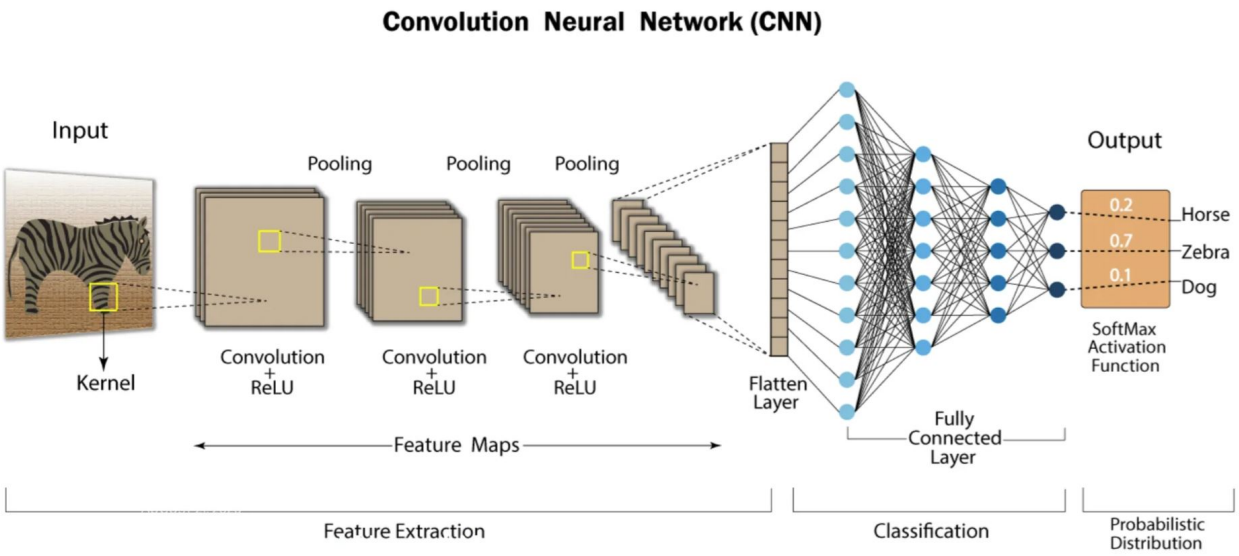


Figure 7: The structure of a convolutional neural network (CNN). Convolutional Layers apply filters to input images to detect features, Pooling Layers reduce the spatial dimensions and computational load, ReLU Layers (activation function) introduce non-linearity, and Fully Connected Layers integrate the features to make final predictions. [33]

Efforts to enhance CNN performance led to the development of deeper and more complex networks. However, deeper networks became harder to train and often resulted in decreased performance. Refinements in network architecture were necessary to address these challenges [34].

CNN's, or deep learning in general has some advantages compared to radiomics. Deep learning has a distinct ability to automatically learn and optimize feature representations directly from the raw image data, often leading to superior performance in capturing complex patterns improving accuracy and performance of the model [11]. However, they also have drawbacks. Training is time-intensive, taking hours to days, even with specialized hardware, contrasting with faster training times for machine learning algorithms. Constrained memory capacity, particularly for hardware, hampers widespread adoption of 3D CNN [27].

7 Methods and materials

In this research, we have developed four models for predicting survival and RFS in patients using data from the Dutch GIST Consortium. The first models two are based solely on clinical data and the final two models utilize imaging data.

7.1 Data and patient characteristics

The data used in this research has been collected in the Dutch GIST Consortium (DGC). The DGC [35] is comprised of GIST expertise centers in the Netherlands, with the goal of centralizing patient care and guaranteeing high-quality treatment. A 'Standard of care' protocol has been developed to guide diagnostic and treatment procedures for GIST [35]. Through their secure GIST Registry, they collect anonymous patient data, facilitating research inquiries. Clinical data, including patient information, pathology, surgery, medical history, treatment, and measurement data, are available for all patients in the GIST consortium. This clinical data is collected from all participating clinical centers: Erasmus Medical Center (EMC), Antonie van Leeuwenhoek (AVL), University Medical Center Groningen (UMCG), Leiden University Medical Center (LUMC), and Radboud University Medical Center (RUMC). In this research, both imaging and clinical data collected in the DSG are used.

A flowchart showing the inclusion and exclusion of patients is given in Figure 8. The upper part shows the inclusion of patients for the patients in the clinical dataset. The lower green part shows the inclusion of patients for the imaging dataset. The clinical dataset is used in the models based on clinical data, while the imaging dataset is used in the models based on imaging data.

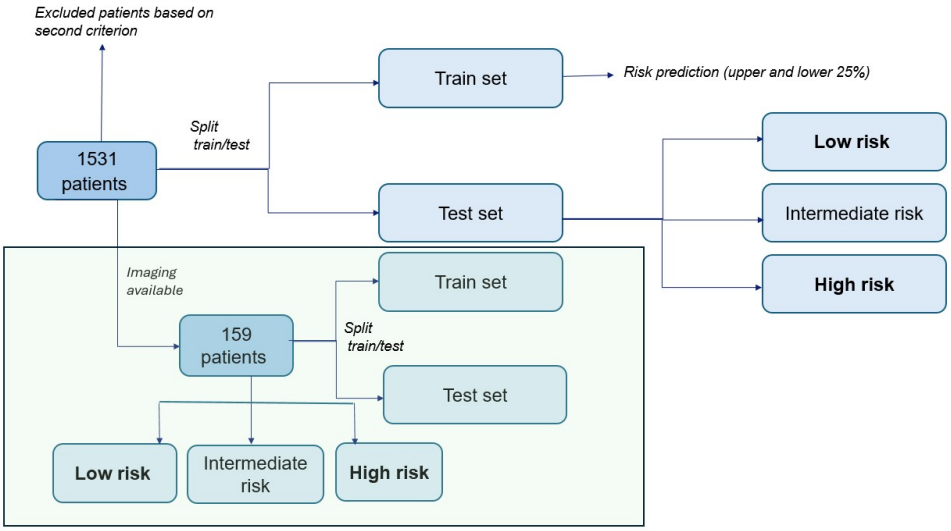


Figure 8: Flowchart showing the inclusion and exclusion of patients. The upper part shows in the inclusion of the clinical dataset, while the green part shows the inclusion of patients for the imaging dataset.

The following inclusion criteria were applied to patients in the GIST consortium:

1. Patients who underwent tumor resection.
2. Time from tumor resection or systemic therapy to the event (death or recurrence) or most recent update date is less than 25 years and available. If the time is more than 25 years or not available, these patients are not included in the dataset.

A total of 1531 patients were included in the clinical dataset. These 1531 were selected based on the first criterion. The patients were split into a 80-20% train-test split. A total of 51 patients were excluded when predicting survival and 32 when predicting RFS based on the second inclusion criterion. Patients were divided into risk groups to look into the ability of the model to stratify risk. This division in groups was made using the top and bottom 25 percentile of risk scores from the training set as cutoff points for the low- and

high-risk group. The risk predictions used were the risk predictions of the model of which its performance is analysed. These cutoff values were subsequently applied to the test set, delineating patients into distinct risk categories (high- and low-risk for the top and bottom 25 percentile of risk scores respectively). The patients inbetween the low- and high-risk prediction values were considered to have an intermediate risk score. The division into three risk groups was chosen to ensure that no random cut-off value was used to divide the patients into a low- and high-risk group. The intermediate-risk category, while not directly analyzed further, provides context by indicating patients whose risk falls between clearly defined high- and low-risk thresholds.

The following additional inclusion criterion was applied for the imaging dataset: Availability of contrast-enhanced CT (CECT) of the GIST, conducted within 6 months before either tumor resection or systemic therapy, with clear tumor visibility. If systemic therapy was administered before tumor resection (neoadjuvant), the scan conducted within 6 months prior to systemic therapy was used. If systemic therapy was administered after surgery, the scan conducted within 6 months prior to surgery was selected. If multiple scan dates were available, the scan closest to the event was chosen. A total of 159 patients were included in dataset including imaging, since for only these patients, a CT baseline scan was available.

Finally, the imaging dataset that included 159 patients was divided into the low-, high-, or intermediate-risk groups. This division into risk groups was made based on the outcome observed 5-years post imaging and was purely made to simplify the problem, in contrast to the division into risk groups in the clinical dataset in which the division was made to look into the model's ability to stratify risk. Patients were categorized into the high-risk group if recurrence or cancer-related death occurred within 5 years after the imaging date. Patients were classified into the low-risk group if they had been included for at least 5 years without experiencing recurrence or cancer-related death. The remaining patients were assigned to the intermediate-risk group. The intermediate-risk group, although identified, is not further analyzed to maintain focus on the clear differentiation between low- and high-risk categories. This 5-year criterion was chosen to ensure that a sufficient number of patients are identified as high-risk, while still focusing on a relatively early period that indicates aggressive disease and poorer prognosis.

7.1.1 Clinical dataset

Table 2: Demographics of included patients in the clinical dataset

		Survival		**RFS**	
		Train	*Test*	*Train*	*Test*
Number of patients		1184	296	1200	299
Sex					
	Male	616	146	624	147
	Female	567	150	575	152
	Unknown	1	0	1	0
Age		63.5 ± 12.6	62.1 ± 12.3	62.4 ± 12.7	61.9 ± 12.4
Tumor location					
	Colon	14	5	14	5
	Duodenal	67	15	68	16
	Esophagus	6	3	6	3
	Gastric	732	197	737	199
	Other	21	6	22	6
	Rectum	64	3	66	3
	Small bowel	280	67	287	67
Duration to event		3490 ± 2264	3701 ± 2191	3128 ± 2212	3355 ± 2192
Event occurred		175	44	274	66
Censored		1009	252	926	233

The patient characteristics of patients included in the clinical dataset are shown in Table 2. These patients are from all expertise centers in the DGC (EMC (n=349), AVL (n=464), UMCG (n=268), LUMC (n=234), RUMC (n=216)). The data that has been extracted from the clinical data collected in DGC is shown in

Table 3.

The features used with the clinical dataset were selected based on their availability for all patients. Features with data present for over half of the patients and deemed predictive of patient survival based on previous literature research were included. However, some patients had missing information for certain features. Missing data were addressed using multiple imputation techniques [36], which generate realistic values based on the distributions and relationships among existing variables in the dataset. For numerical data, imputation was directly applied, while for categorical values, the data were first encoded before imputation and decoded afterward. To integrate categorical features, one-hot encoding was employed to convert these features into numerical values.

Table 3: Features included in the feature sets. The first feature set only contains clinical variables while the extensive feature set also contains features derived from pathology.

Minimal feature set	**Extensive feature set**
Sex	Cell type
Age	CD117
Tumor location	DOG1
Tumor status at diagnosis	KIT
Primary tumor size	KIT mutation location
Largest lesion	BRAF
Number of lesions	Mitotic count

Two feature sets were constructed based on the features utilized for prediction. The first feature set (minimal feature set) comprised solely clinical variables such as age and sex. The second feature set (extensive feature set) also incorporated variables derived from pathology data. Two feature sets were identified to look whether the additional variables derived from pathology data, like mutation status, would result in a better prognosis prediction. The features included are detailed in Table 3.

The prognosis prediction was determined based on the occurrence of an event, which could be death, recurrence, or censorship. The time to event was calculated by subtracting the baseline date from the event date. The baseline date was defined as follows: the earlier of the systemic therapy start date or the surgery date was selected. The recurrence date was determined based on measurements results on imaging, specifically the date on which the imaging measurements indicated recurrence. Based on these two possible outcomes, there were two possible outcome predictions, one with the time to death and one with time to recurrence.

7.1.2 Imaging dataset

The patient characteristics of the patients included in the imaging dataset are outlined in Table 4. The patients were split into a 80-20% train-test split. These patients are only from the the EMC (n=61) and AVL (n=98). As patients come from two medical centers and different scanners were used, scans were conducted in multiple planes and reconstructed using more than one reconstruction kernel. Additionally, the chosen imaging protocol may vary across scans. The scan has been selected that was conducted a maximum of 6 months before either tumor resection or systemic therapy. The reconstruction kernel was chosen based on tumor visibility and image quality. When several protocols or reconstruction kernels were available where tumor visibility was acceptable, the image with the highest resolution was chosen. It is assumed that the selected imaging protocols and kernels were chosen based on their ability to visualize tumors effectively, and it is expected that variations among these protocols do not significantly impact the results when used interchangeably.

Similar to the previously described model, the prognosis prediction was determined based on the occurrence of an event, such as death, recurrence, or censorship. The time to event was again calculated by subtracting the baseline date from the event date of which the baseline date is now the date of imaging.

The imaging dataset was divided into the low-, high-, or intermediate-risk groups. These patients are again only from the the EMC (n=61) and AVL (n=98) and classified as high- and low-risk based on the outcome

Table 4: Demographics of included patients in the imaging dataset

		Survival		**RFS**	
		Train	*Test*	*Train*	*Test*
Number of patients		127	32	127	32
Sex					
	Male	58	14	58	14
	Female	69	18	69	18
Age		63.2 ± 12.4	59.2 ± 14.5	63.2 ± 12.4	59.2 ± 14.5
Tumor location					
	Colon	0	0	0	0
	Duodenal	9	4	9	4
	Esophagus	3	0	3	0
	Gastric	88	20	88	20
	Other	5	0	5	0
	Rectum	5	2	5	2
	Small bowel	17	6	17	6
Duration to event		1935 ± 1003	2092 ± 1186	1800 ± 975	1908 ± 1042
Event occurred		19	2	24	4
Censored		108	30	103	28

observed 5-years post imaging as previously described. The patients were split into a 80-20% train-test split. For classification based on cancer-related death, a total of 78 patients were included, with 69 belonging to the low-risk group and 9 to the high-risk group. Similarly, for classification based on cancer-related death or recurrence, a total of 81 patients were included, with 62 in the low-risk group and 19 in the high-risk group.

7.2 Modelling of survival

Survival analysis, also known as time-to-event analysis, is a statistical method used to analyze the time it takes for an event of interest to occur, like death or recurrence. This method is particularly relevant in medical research, where it aids in assessing patient prognosis, treatment efficacy, and disease progression. The event represents the occurrence of a specific event, often referred to as the "death event," while time denotes the point of the initial observation, also termed the "birth event." The duration between the first observation and the event of interest, in this case death or recurrence is termed "time to event" [37].

During the observation period in survival analysis, not all subjects experience the event of interest. This leads to censorship, where the exact occurrence time of the event is unknown for certain subjects. Censoring occurs either because the event has not yet happened or because the study concludes before the event's occurrence. Properly accounting for censored data in survival analysis is crucial to prevent biased estimations and accurately assess outcomes. The most common form of censorship in survival analysis is right censoring, where the death event has not occurred by the end of the study. Conversely, left censoring occurs when the birth event is not observable. Here, only right censoring occured.

7.2.1 Survival function

The survival function shows the probability that an individual has 'survived' beyond time t. The survival function (S(t)) is denoted as:

$$S(t) = Pr(T > t) \tag{1}$$

Here, T is the random lifetime variable drawn from the studied population. The survival function ranges between 0 and 1 and is a non-increasing function of time t.

7.2.2 Hazard function

The hazard function, which is derived from the survival function shows the probability of the death event occurring at time t, given that the subject did experience the death event up to time t. The function shows the instantaneous potential per unit time for the occurrence of the event. The hazard function (h(t)) is

denoted as:

$$h(t) = lim(\delta \to 0)\frac{Pr(t \leq T < t + \delta | T \geq t)}{\delta} \tag{2}$$

7.2.3 Methods for survival analysis

Survival data analysis employs three main methodological categories:

- *Parametric models*: Parametric methods in survival analysis assume a specific distribution for the survival times, such as exponential, Weibull, or lognormal. An example of a parametric model is the linear regression model.
- *Non-parametric models*: Non-parametric methods in survival analysis make no assumptions about the underlying distribution of survival times. The Kaplan-Meier estimator is a common non-parametric method for estimating survival curves.
- *Semi-parametric models*: Semi-parametric methods combine aspects of both parametric and non-parametric methods to offer a balance between flexibility and structure. The Cox Proportional-Hazards model (CPH) is a widely used semi-parametric method.

7.2.4 Kaplan Meier estimator

The Kaplan-Meier estimator, a widely used non-parametric method, estimates survival probabilities by dividing observed event times into intervals. It offers valuable insights into survival patterns but lacks the ability to incorporate covariates into survival estimations. Within each interval, the probability of surviving until the end of that interval is computed using the following equation:

$$S(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} \tag{3}$$

Here, n_i is a number of individuals who are at risk at time point t_i and d_i is a number of subjects that experienced the event at time t_i.

7.2.5 Cox proportional hazard (CPH) model

The CPH model integrates time, censorship, and covariates to assess survival distributions. It assumes proportional hazards over time and is expressed as:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k) \tag{4}$$

Here, the H(t,X) represents the hazard function, $h_0(t)$ is the baseline hazard function, giving the hazard when all covariates are zero. Finally, the β are the coefficients of the different covariates in the model.

7.2.6 Deep learning model based on clinical characteristics

The integration of machine learning, particularly deep learning, into survival analysis has shown promising results. Deep learning models, including Cox-based and discrete-time models, have demonstrated efficacy in capturing nonlinear relationships and handling high-dimensional data [38].

While traditional Cox regression assumes linear relationships between factors and event risk, more complex survival models are necessary for personalized treatment recommendations [39]. To achieve this, several deep learning modifications and extensions of the Cox regression model have been proposed [38]. DeepSurv, an extension of the feed-forward neural network proposed by Faraggi-Simon [40], has shown potential in capturing nonlinear log-risk functions in survival data [39]. The model optimizes the Cox partial likelihood and incorporates regularization to prevent overfitting. The basic structure of the network is shown in Figure 9 [39]. The model that was used to predict survival and RFS in GIST patients is based on this DeepSurv network architecture.

The input of the network is the baseline data x (feature sets). The hidden layers of the network consist of a fully-connected layer, which is followed by a dropout layer. The output is a single node which estimates the

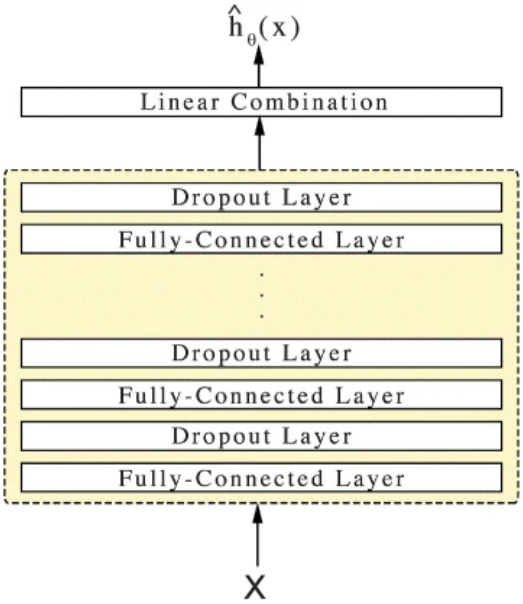


Figure 9: Diagram of the basic structure of DeepSurv. Baseline data x is taken as an input and is processed through hidden layers with weights θ, and outputs the predicted log-risk function. [39]

log-risk function in the model (see equation 4). The network is trained and the gradient descent method is used to find the weights of the network. The optimization of the classical Cox regression runs by optimizing the Cox partial likelihood. The likelihood is defined by the following expression [39]:

$$\text{Partial Likelihood} = \prod_{i=1}^{n} \frac{\exp(\beta^\top z_i)}{\sum_{j \in R(t_i)} \exp(\beta^\top z_j)} \tag{5}$$

where:

n is the number of observations,
z_i is the vector of covariates for observation i,
β is the vector of coefficients,
$R(t_i)$ is the set of individuals who are at risk at time t_i.

The loss function used in the model was the negative log partial likelihood from the equation above but with additional regulation. The regulation parameter is λ.

$$\text{Regularized Negative Log Partial Likelihood} = -\sum_{i=1}^{n} \left[\beta^\top z_i - \log \left(\sum_{j \in R(t_i)} \exp(\beta^\top z_j) \right) \right] + \frac{\lambda}{2} \|\beta\|_2^2 \tag{6}$$

The model performance was measured with the concordance index (C-index). The C-index represents the fraction of comparable pairs where the predictions and outcomes align, indicating concordance. Two samples are comparable under the following conditions: either both samples experienced an event at different times or one sample with a shorter observed survival time experience an event. A pair is incomparable if both samples experienced events simultaneously. The C-index is calculated by the following equation [41], where the indices i and j refer to pairs of observations in the sample:

$$C = \frac{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) \cdot I(\eta_j > \eta_i) \cdot \Delta_j}{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) \cdot \Delta_j} \tag{7}$$

where:

$\tilde{T}_i$ and $\tilde{T}_j$ represent the estimated survival times,
η_i and η_j represent the predicted risk scores,
Δ_j is an indicator variable,
I is the indicator function.

High- and low-risk score To further evaluate the survival prediction, a high- and low-risk group have been defined, based on the risk prediction of the model. This will give more insight on the model's ability to make a risk stratification.

7.2.7 Deep learning model based on imaging

Several extensions have been made on DeepSurv including networks focused on the applicability to high-dimensional data like images [38].

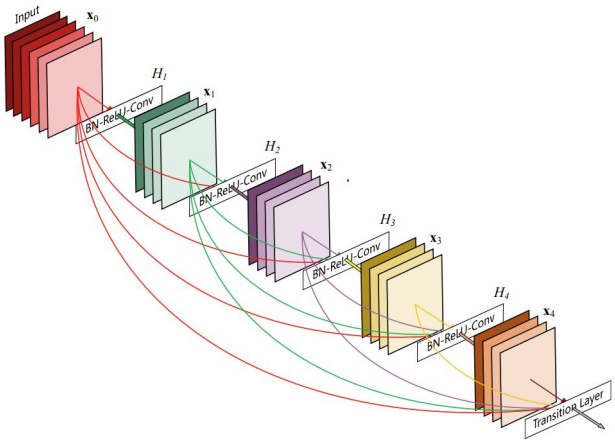


Figure 10: Dense block of a DenseNet with a growth rate of k = 4. Each layer within the DenseBlock receives feature maps from all previous layers and passes its own features to all subsequent layers. [42]

In this model, a DenseNet 121 architecture was used. DenseNet, which is short for Densely Connected Convolutional Networks, is type of CNN of which its key feature is its dense connectivity pattern. In this pattern, each layer receives input from all preceding layers. The dense connectivity helps to promote feature reuse. Since each layer receives feature maps from all preceding layers, the network can be more compact. The growth rate k is the additional number of channels for each layer. The architecture of a building block of a DenseNet, called a dense block is shown in Figure 10 [42]. The dense blocks are connected by transition layers, which perform pooling and convolution operations [42].

7.2.8 Classification model

To simplify the problem, the survival prediction model has been rebuilt into a classification model. This classification model used the DenseNet 121 architecture as in the deep learning model based on imaging, but predicted whether the patients belong to either the low- or high-risk group. The loss function to train the network was the Cross-Entropy Loss. This loss function measures the performance of a classification model whose output is a probability value between 0 and 1. It measures how "wrong" the model's predictions are by comparing the predicted probabilities to the true distribution of labels. The equation for Cross-Entropy Loss for a single training example is given by:

$$\text{Cross-Entropy Loss} = -\sum_i y_i \log(p_i) \tag{8}$$

The outcome of the model is assessed with the accuracy with the following equation:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \tag{9}$$

In addition to the accuracy, the sensitivity and specificity were determined. When evaluating accuracy, sensitivity and specificity, a cutoff of 0.5 was applied for determining positive and negative predictions. Sensitivity identifies the model's ability to detect high-risk cases, while specificity denotes its capability to recognize low-risk cases. Additionally, the AUC (area under the ROC curve) was calculated. The ROC curve illustrates a model's performance across various classification thresholds by plotting true positive rates against false positive rates. AUC quantifies the overall accuracy of the model, ranging from 0 (random guessing) to 1 (perfect predictions).

7.2.9 Cross-validation

The performance of a model is most accurately assessed using cross-validation techniques. One commonly used method is k-fold cross validationn [43]. In k-fold cross validation, the dataset is divided into k subsets or folds. The model is trained on k minus 1 folds and validated on the remaining fold. This process is repeated k times, with each fold being used as the validation set once. The overall performance is then averaged across all folds, providing a more reliable estimate of the model's generalization ability.

To optimize hyperparameters while also performing cross-validation, nested k-fold cross validation can be used. In nested cross validation, there is double loop, an outer loop and an inner loop. The outer loop is used the assess the quality of the model and the inner loop is used for hyperparameter optimization as shown in Figure 11 [44]. This procedure ensures that there will not be a biased evaluation of the performance [43].

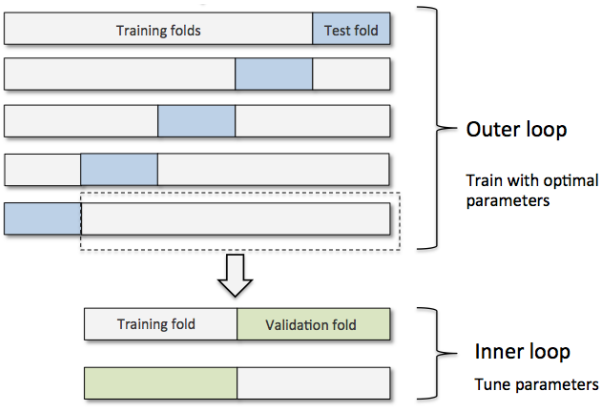


Figure 11: Nested cross validation. There is an outer and inner loop, where the outer loop is used to assess quality of the model, while the inner loop is used to tune parameters of the network. [44]

Hyperparamaters were optimized using Optuna with a Tree-structured Parzen Estimator (TPE) algorithm. TPE is an Bayesian optimization algorithm that uses the probability distribution to represent the hyperparamater search space, adjusting the distribution according to the performance of the explored hyperparameters [45]. The hyperparameters were optimized based on the average validation C-index over the final 10 epochs. This approach has been chosen to ensure the stability of the network, as certain combinations of hyperparameters led to models exhibiting unpredictable behavior. Unstable models can result in varying performance or diverging learning curves, undermining the reliability of the model and complicating interpretation.

Furthermore, for unbalanced dataset it is useful to use stratified k-fold cross validation. This method involves dividing the dataset into k equal-sized folds while ensuring that each fold maintains the same class distribution as the original dataset [46]. By stratifying the data based on class labels, the aim is to mitigate the effects of class imbalances and ensure the generalization ability of the model.

7.2.10 Interpretability

Feature exploratory analysis To identify which features independently influence survival and RFS, the CPH model was utilized. Features deemed significant were those that showed a clear impact on survival probabilities.

Shapley analysis To gain deeper insights into the factors influencing the outcomes of the deep learning model, Shapley analysis can be used. The Shapley value, initially conceptualized in cooperative game theory, offers a systematic approach to attribute contributions of individual features to model predictions [47]. The calculation of Shapley values involves considering all possible permutations of feature combinations and evaluating the difference in predictions when a specific feature is included or excluded. Mathematically, the Shapley value ϕ_i for a feature i in a cooperative game with N features is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

where:

S represents a set of features excluding i,
$v(S)$ denotes the model's prediction when the features in set S are present,
$v(S \cup \{i\})$ denotes the model's prediction when the feature i is added to the set S,
$|S|$ denotes the cardinality of set S,
N represents the set of all features.

In simpler terms, Shapley values quantify the average marginal contribution of a feature to the prediction compared to all possible feature combinations. It provides a way to interpret the relative impact of inputs on the model's output by assessing the change in predictions as each feature is included, while considering all possible permutations.

Grad-CAM To look into the interpretability of the imaging-based model, Grad-CAM, which is short for Gradient-weighted Class Activation Mapping can be used. Grad-CAM is a technique used for visualization and understanding the decisions made by the CNN. It operates by generating a heatmap that highlights pixels with the highest gradients, specifically for a given class prediction [48].

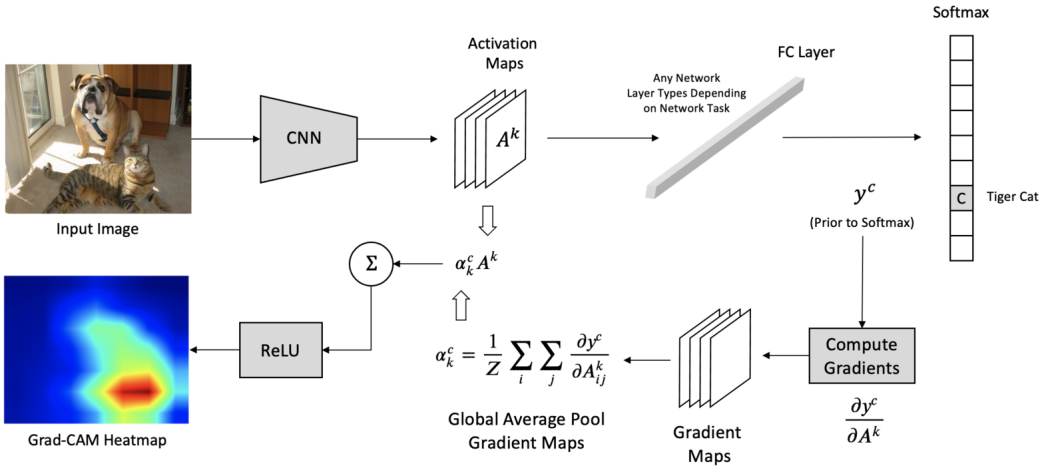


Figure 12: Grad-CAM Visualization Process: This Figure shows the steps involved in generating a Grad-CAM heatmap for an input image. The input image is processed through a CNN to produce activation maps. Gradients of the target class score are computed with respect to these activation maps. These gradients are then averaged to obtain importance weights, which are used to create a heatmap highlighting regions of the image that are important for the prediction. The final heatmap is overlaid on the original image, providing a visual explanation of the model's decision. [48]

When a CNN predicts a class, its final convolutional layer generates feature maps representing learned image features. Grad-CAM computes gradients of scores for the target class with respect to these feature maps, indicating how changes in features affect class scores [48]. These gradients are spatially averaged to assign importance weights to each feature map channel. Multiplying these weights with feature maps highlights regions influencing the CNN's decision. The resulting weighted feature maps are summed to produce a heatmap, visually depicting regions crucial for the CNN's decision, as illustrated in Figure 12 [48].

7.3 Experimental set-up

7.3.1 Models based on clinical characteristics

The patients from the clinical dataset have been used in these experiments. The deep learning model based on clinical characteristics consisted of three hidden layers, which consisted of fully-connected layers with nonlinear activation functions (SeLU) followed by dropout. The loss function used to train the network was the average negative log partial likelihood with regularization. A learning rate of 0.0001 and learning rate decay of 0.005 were used with a inverse time decay as learning rate scheduler. Both values were chosen based on manual experimentation, where the learning curves were monitored to ensure stable convergence without overshooting. The model was trained for 800 epochs, as it was observed that the loss stopped decreasing significantly beyond this point. This ensured that the training was efficient and prevented unnecessary computation without further improvement in model performance. The performance of the model was evaluated with the C-index.

Table 5: Hyperparamaters used for model evaluation. Set 1 and 2 relate to the minimal and extensive feature sets respectively that are used to make a prediction.

	Set 1 (Survival)	**Set 1 (RFS)**	**Set 2 (Survival)**	**Set 2 (RFS)**
# Nodes layer 2	99	95	100	100
# Nodes layer 3	55	71	67	67
Dropout	0.225	0.184	0.274	0.274
L2 reg	13.152	12.796	14.598	14.598

Other hyperparameters of the network include the number of nodes in each layer, dropout probability and L2 regularization coefficient. The hyperparameters were optimized and the model was evaluated using nested 5-fold cross validation (5 outer folds, 3 inner folds). The window in which the hyperparameters were optimized was 1 to 100 for the number of nodes, 0 to 0.5 for dropout and 0 to 20 for the L2 regularisation factor. The parameters to be optimized and the ranges chosen were based on the parameters and ranges used in DeepSurv for different datasets [39]. The hyperparameters used for all evaluation metrics of this model are shown in Table 5. For comparison, a CPH model was also evaluated with 5-fold cross validation with the same folds. In addition, the models were evaluated using the original train-test splits (Table 2) to ensure similar of performance of this split compared to the cross-validation experiment.

High- and low-risk score To compare the survival distributions of the two groups (high- and low-risk), a log-rank test was performed. P-values below the 0.05 threshold were considered a significant difference between low- and high-risk. For this evaluation, the original train-test splits were used (Table 2). For comparison, the CPH model was evaluated similarly, using the same train-test splits.

Interpretability To identify which features independently influence the outcome of the CPH model, significance testing with a Wald test was performed. Features with p-values below the 0.05 threshold were considered to significantly influence survival probabilities. For this evaluation, the original train-test splits were used (Table 2).

To identify which features the deep learning model relied on for its predictions, Shapley analysis was conducted. For this evaluation, the original train-test splits were utilized (Table 2). The model was initially trained on the training set, and then the test set was used to compute the mean Shapley values for each feature.

7.3.2 Models based on imaging

The patients from the imaging dataset have been used in these experiments. Some pre-processing has been performed on the images to ensure compatibility with the model. First, the data was resampled to ensure consistent spacing within the images to a spacing of (1, 1, 3). This specific spacing was chosen to ensure the images do not exceed GPU memory, while still maintaining as much resolution as possible. Secondly, the intensity of the images was scaled to normalize the intensity values of medical images to the standard

range [0, 1], which can improve the performance and stability of the model [49, 50]. The intensity range of [-125, 225] of the original image was scaled to this standard range to get rid of background and focus on soft tissue. Finally, zero padding was performed to the size of (32, 32, 32), to ensure the minimum required size of a DenseNet 121 is achieved. This padding was only performed, when the size of the image was smaller than (32, 32, 32), which was rarely the case.

The loss function to train the network was the average negative log partial likelihood with regularization. Like in the deep learning model based on clinical characteristics, a learning rate of 0.0001 and learning rate decay of 0.005 were used with a inverse time decay as learning rate scheduler. The model performance was evaluated with the C-index. The model was trained for 100 epochs, since it was observed that the loss stopped decreasing significantly beyond this point. The model was evaluated using the original train-test split (Table 4).

Classification model The classification model predicted whether the patients belonged to either the low- or high-risk group. The loss function to train the network was the CrossEntropyLoss. A learning rate of 0.00001 has been used. Both values were again chosen based on manual experimentation, where the learning curves were monitored to ensure stable convergence without overshooting. The model was again trained for 100 epochs, after which the loss stopped decreasing significantly, indicating that training was sufficient. The model performance was evaluated with the accuracy, sensitivity, specificity and AUC. The model was evaluated using stratified 5-fold cross validation, because of the class imbalance in the dataset.

Interpretability To visualize the decision-making process of the model, Grad-CAM was used. The Grad-CAM was targeted on the final convolutional layer of the DenseNet 121. This corresponds to the final convolutional layer of the 16th denselayer in the 4th denseblock. The model was trained on the train dataset and heatmaps were generated for three random images of the dataset. The heatmaps are shown of the middle part of the image so at depth/2 of the images.

8 Results

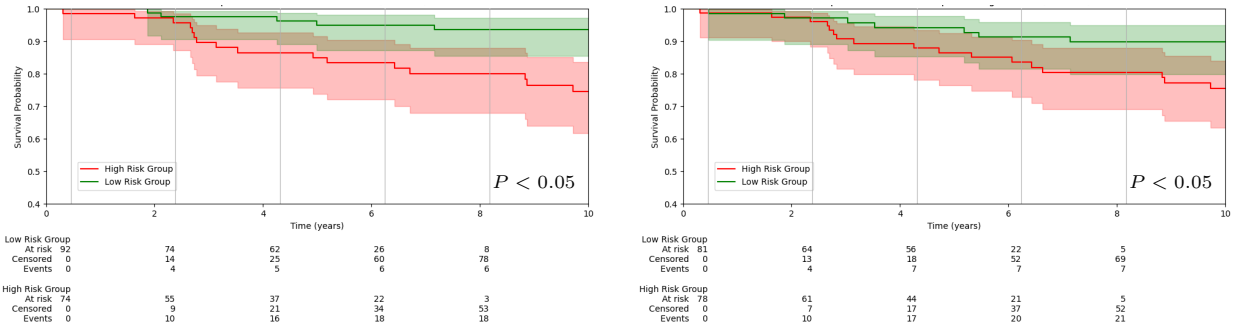
8.1 Models based on clinical characteristics

The C-indices for the different feature sets and predictions are shown in Table 6. Using only clinical features (minimal feature set) the deep learning model achieves a C-index of 0.63 (95% CI: 0.61 - 0.66). This is similar to the performance of the CPH model (0.64 (95% CI: 0.60 - 0.69)). Next, the deep learning model (0.60 (95% CI: 0.58 - 0.62)) also achieves comparable results to CPH model (0.63 (95% CI: 0.61 - 0.64)) for RFS prediction. Finally, for both methods and outcomes, additional features of the extensive feature set do not improve the results (Table 6). The deep learning model also achieves similar results to the CPH model when using the original train-test split (Table 6). The results of the C-index therefore show that the performance of the deep learning model is similar to the performance of the CPH model.

Table 6: C-index of CPH and the deep learning model. The C-indices are given for both the minimal and extensive feature and for both the prediction of survival and RFS. The results are shown for 5-fold cross validation and the original train-test split.

		Cross validation (95% CI)	**Train-test split**
CPH model	*Minimal feature set (survival)*	0.64 (0.60 - 0.69)	0.64
	Minimal feature set (RFS)	0.63 (0.61 - 0.64)	0.65
	Extensive feature set (survival)	0.64 (0.60 - 0.70)	0.64
	Extensive feature set (RFS)	0.63 (0.60 - 0.66)	0.66
Deep learning model	*Minimal feature set (survival)*	0.63 (0.61 - 0.66)	0.61
	Minimal feature set (RFS)	0.60 (0.58 - 0.62)	0.60
	Extensive feature set (survival)	0.64 (0.61 - 0.68)	0.64
	Extensive feature set (RFS)	0.62 (0.58 - 0.65)	0.63

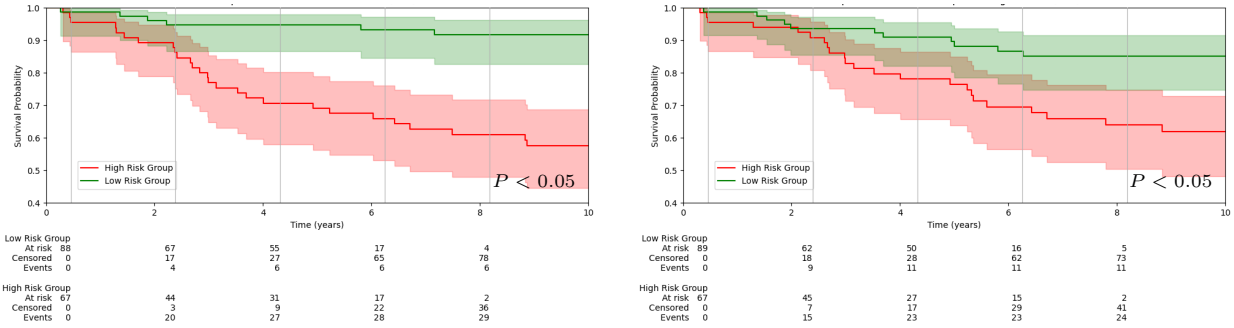
8.1.1 Low- and high-risk group



(a) CPH model *(b) Deep learning model*

Figure 13: Kaplan–Meier plots showing high- and low-risk groups for survival prediction for the extensive feature set. The high-risk (red line) and low-risk (green line) patients from each model were divided based on the upper and lower 25% of the risk predictions of the train set. The P-value for the log-rank test comparing the low- and high-risk group is shown in the Figure. The number of patients at risk at different points in time are shown for both groups below the Figure.

For the extensive feature set, both the CPH model (P: <0.05) as the deep learning model (P: <0.05) show significantly different survival probabilities between the low- and high-risk groups for survival prediction (Figure 13). Both the CPH (P: <0.05) as the deep learning model (P: <0.05) also show significantly different survival probabilities for the low- and high-risk group for RFS prediction (Figure 14). For the minimal feature set, the CPH model shows significantly different survival probabilities between the low- and high-risk group (P <0.05) when predicting survival (Figure S1a). The deep learning model does not show a significant difference in survival probabilities (P: 0.10) (Figure S1b). Furthermore, for RFS prediction, the CPH model



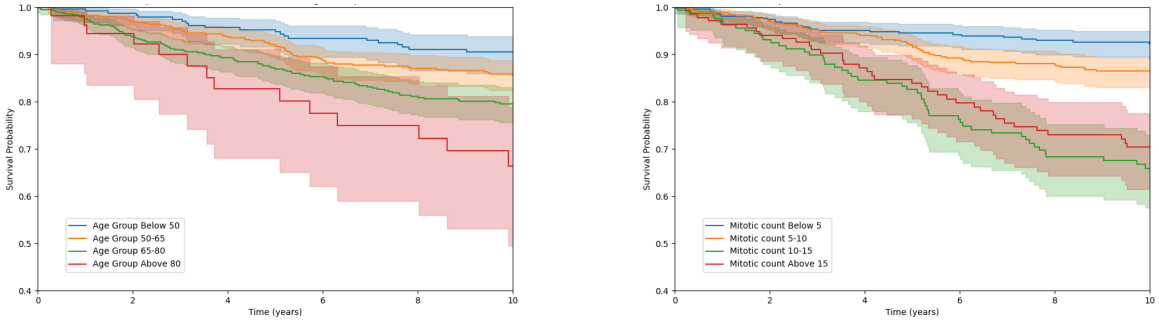
(a) CPH model *(b) Deep learning model*

Figure 14: Kaplan–Meier plots showing high- and low-risk groups for RFS prediction for the extensive feature set. The high-risk (red line) and low-risk (green line) patients from each model were divided based on the upper and lower 25% of the risk predictions of the train set. The P-value for the log-rank test comparing the low- and high-risk group is shown in the Figure. The number of patients at risk at different points in time are shown for both groups below the Figure.

again shows a significant difference (P: <0.05), while the deep learning model does not (P: 0.20) (Figure S2).

For the extensive feature set when predicting survival, more patients are included in the low-risk group in the CPH model (n_{low}=92) than in the deep learning model (n_{low}=81) (Figure 13). Conversely, the deep learning model includes more patients in the high-risk group (n_{high}=78) compared to the CPH model (n_{high}=74). For RFS prediction, the patient distribution is similar between the CPH model (n_{low}=88, n_{high}=67) and the deep learning model (n_{low}=89, n_{high}=67) (Figure 14). For the minimal feature set predicting survival, the CPH model includes more patients in both the high- and low-risk groups (n_{low}=78, n_{high}=72) compared to the deep learning model (n_{low}=68, n_{high}=64) (Figure S1). For RFS prediction, the CPH model includes more patients in the low-risk group (n_{low}=89) than the deep learning model (n_{low}=64), but slightly fewer in the high-risk group (n_{high}=66) compared to the deep learning model (n_{high}=69) (Figure S2).

8.1.2 Interpretability

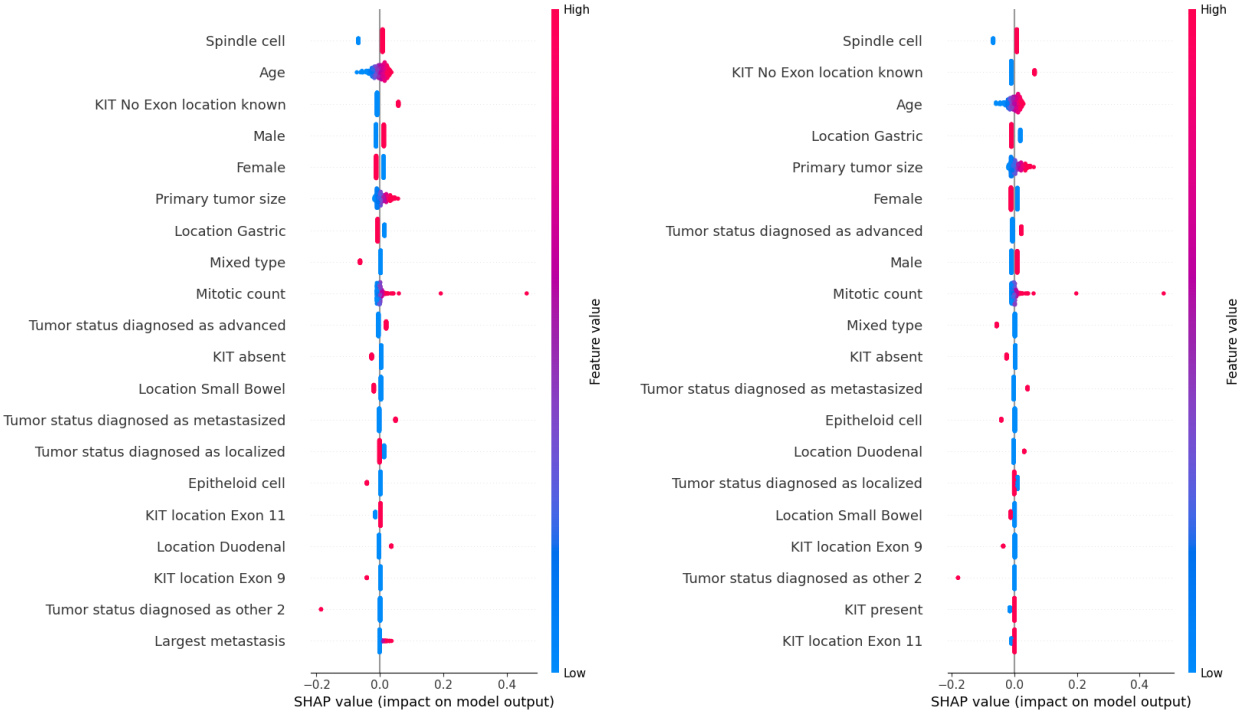


(a) Influence of different age *(b) Influence of different mitotic count*

Figure 15: Kaplan-Meier curves showing the influence of feature change of the features that significantly (P <0.05) influence the survival based on the CPH model

Several factors are important in the CPH model for survival and RFS prediction. In the minimal feature set, age significantly (P: <0.05) influences survival and RFS (Figure 15a, Figure S3a), where an increased age results in a lower survival probability. Moreover, age is more predictive for survival than RFS. Similarly, an increase in size of the primary tumor also significantly (P: <0.05) influences a lower probability for RFS (Figure S3b). In the extensive feature set, an increase in mitotic count also lowers the survival probability and RFS significantly up to a mitotic count of 10 (P: <0.05) (Figure 15b, Figure S3c). A mitotic count >15 results in a higher survival probability compared to a mitotic count of 10-15 for both predicting survival

(Figure 15b) as RFS (Figure S3c).



(a) Survival prediction *(b) RFS prediction*

Figure 16: Shapley analysis of the deep learning model for the extensive feature set. The mean shapley value of the test set is shown. The features are ranked from most important to least important. High feature values are shown in red and low feature values are shown in blue.

The results of the Shapley anlaysis are shown in Figure 16 and S4. For the survival prediction with the extensive feature set (Figure 16a), the cell histology as 'spindle cell' is the most important feature. The age is the 2nd most important important feature, while the mitotic count is only the 9th most important feature in the model outcome. Both an increase in age as an increase in mitotic count have a positive influence on the model output (increase in risk). This is similar for the RFS prediction of the extensive feature set (Figure 16b) for which the cell histology as 'spindle cell' is again the most important feature. Age and mitotic count are the 3rd and 9th most important features. The primary tumor size is the 5th most important feature in the model outcome prediction. An increase in primary tumor size positively influences the model outcome, thus increasing the risk of recurrence. For the extensive feature set, the Shapley analysis shows that the importance of each feature does not exactly align with the importance of the features when interaction is not taken into account as in the CPH model.

For the survival prediction with the minimal feature set (Figure S4a), the tumor status at diagnosis marked as 'localized' is the most important feature for the model output. The age is the 2nd most important feature. An increase in age has again a positive influence on the model output (increase in risk). For the RFS prediction with the minimal feature set (Figure S4b), the tumor status at diagnosis marked as 'localized' is again the most important feature for the model output. The primary tumor size is the 6th most important feature, while the age was the 7th most important feature in predicting the RFS. For the minimal feature set, the Shapley analysis also shows that the most important features influencing model outcome in the deep learning model, again do not align with the features that were most important in the CPH model.

8.2 Models based on imaging

During the training and evaluation of the deep learning model, the C-index was fluctuating around 0.5, while additional training time does not help with increasing the C-index. Despite efforts in model architecture design and training parameter tuning, the model only yields a train C-index of 0.517 and a test C-index of 0.5,

which indicates the model is just guessing. Therefore, the problem has been simplified into a classification problem.

8.2.1 Classification model

The accuracy, AUC, specificity and sensitivity of the classification model are shown in Table 7. The accuracy of the classification model, using a cutoff of 0.5 for class prediction, predicting high- or low-risk with the event being recurrence or death is relatively high (0.77 (95% CI: 0.74 - 0.79)). The accuracy is even higher with the event being only death (0.89 (95% CI: 0.86 - 0.91)). However, when looking at the AUC, it is relatively low for both the event being recurrence and death (0.59 (95% CI: 0.51 - 0.67)) and the event being only death (0.57 (95% CI: 0.46 - 0.68)). When taking the specificity and sensitivity into account, it becomes clear that all cases are assigned to the low-risk group: the specificity is 1.00 and the sensitivity is 0.00 for both death as death or recurrence, again using a cutoff of 0.5 for class prediction.

Table 7: Performance of classification model, showing the accuracy, AUC, specificity, and sensitivity for 5-fold cross-validation.

	Accuracy (95% CI)	**AUC (95% CI)**	**Specificity**	**Sensitivity**
Death/recurrence	0.77 (0.74 - 0.79)	0.59 (0.51 - 0.67)	1.00	0.00
Death	0.89 (0.86 - 0.91)	0.57 (0.46 - 0.68)	1.00	0.00

8.2.2 Interpretability

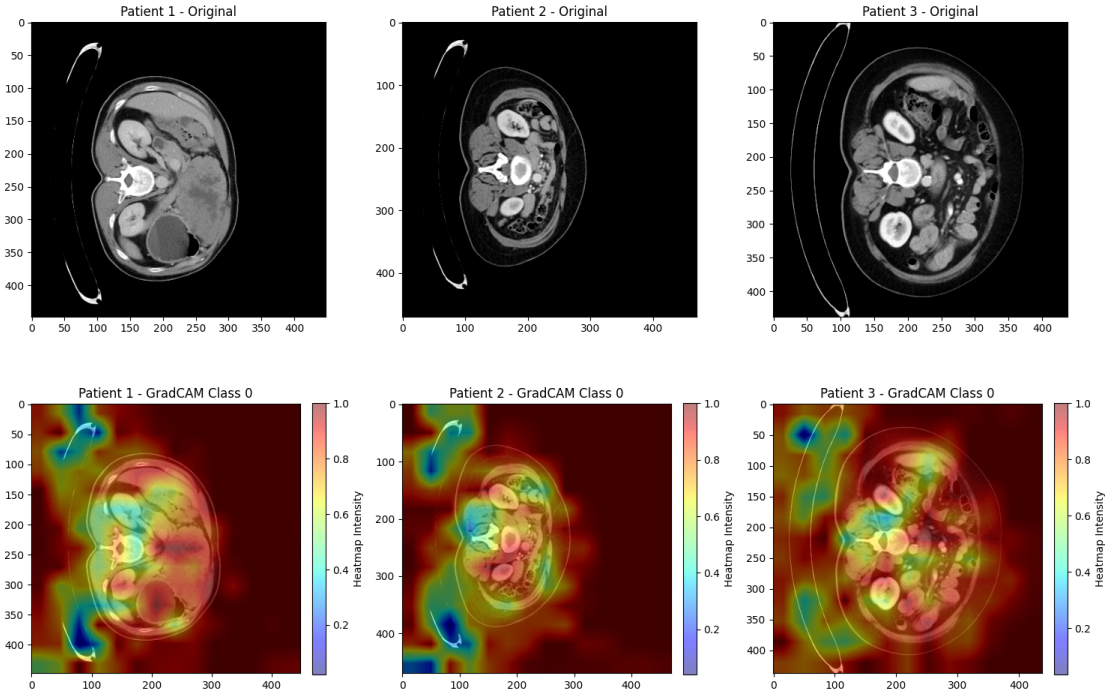


Figure 17: Grad-CAM analysis of three patients. Here the original CT scan images alongside the GradCAM heatmaps for the low-risk class (class 0) are displayed. Each column corresponds to one patient, with the top row showing the original images and the bottom row showing the GradCAM heatmaps overlaid.

The results of the Grad-CAM analysis are presented in Figure 17. For all patients, in the original CT scan, a cross-sectional view of the abdominal region is shown. The heatmap reveals that for all three patients, the model focuses on some parts of the abdomen, but there are especially high activations in the background of the images. For the first patient, the tumor is shown in the image and there is some activation in this area,

which may suggest that the model has learned certain features within this area being indicative of low-risk classification. There are however also activations in the abdomen at places that are not tumor tissue. For the other two patients, the tumor is not shown in the Figure, but there are high activations present in the abdomen.

9 Discussion

9.1 Models based on clinical characteristics

The deep learning and CPH model provide similar performance for predicting survival and RFS based on the C-index. The performance is relatively consistent across different train-test splits in the cross-validation experiments, indicating consistency and robustness in the predictive capabilities of both models. No improvements are found when adding features from the extensive feature set, contrary to expectations given the prognostic value of genetic markers in the additional feature set [7]. However, it is possible that the initial set of features in the minimal feature set already captures most relevant information necessary for survival and RFS prediction. Therefore, despite the prognostic value of genetic markers, their incorporation does not enhance model performance.

The deep learning model performs worse than the CPH model in patient risk stratification. This is especially shown in the distribution of patients into low- and high-risk groups by the different models. Generally, the CPH model includes more patients in the low-risk group, indicating superior classification of patients compared to the deep learning model. Shapley analysis shows that the deep learning model considers not only the features most important in the CPH model's prediction but also additional features such as 'localized' tumor status at diagnosis or 'spindle cell' tumor histology, which are less relevant in the CPH model. These features might interact with other factors, influencing model outcomes. While deep learning models are able to capture complex patterns and interactions in data [29], the effectiveness of these capabilities depends on the quality and relevance of the interaction of the features inputted into the model. These findings show that the use of interaction between features does not increase model performance.

Shapley analysis highlights that the tumor status at diagnosis marked as 'localized' is most important for survival and RFS prediction using the minimal feature set. Although this feature is not identified as important in the CPH model, its effect is intuitive: The tumor status at diagnosis marked as 'localized' has a negative influence on the model output (decrease in risk) while the tumor status at diagnosis not marked as 'localized' has a positive influence on the model output (increase in risk). Shapley analysis shows that the tumor histology being 'spindle cell' is most important for the prediction of survival and RFS using the extensive feature set. A spindle cell histology decreases the risk of death or recurrence in the deep learning model, in contrast to Joensuu et al. in 2008 [7], who described the occurrence of spindle cell histology with a poorer 5-year RFS. This discrepancy might stem from the deep learning model capturing complex interactions with spindle cell histology that this study did not consider. These results underscore the need for further investigation to understand the underlying mechanisms and current interaction between features in the deep learning model. It appears that the model might be overfitting by capturing overly complex interactions, as its performance does not surpass that of the simpler and more interpretable CPH model in patient risk stratification.

Feature exploration with the CPH model reveals that only age and mitotic count are significant for survival prediction, with primary tumor size additionally significant for RFS prediction. Increases in age, mitotic count, and primary tumor size decrease survival and recurrence probabilities, except for an increase of mitotic count from 10-15/50 HPF to >15/50 HPF. The mitotic count of high-risk GIST is according to the NIH classification system >5/50 HPF [3]. The relative impact may diminish beyond this threshold, possibly because counts meeting the high-risk threshold are already surpassed.

While the C-index indicates similar performance for both models, risk classification favors the CPH model over the deep learning model. The C-index measures a model's ability to rank survival times accurately [51]. Thus, despite similar C-index values, differences in stratifying risk groups suggest that focusing solely on the C-index might be insufficient. Consideration of additional metrics such as the Brier score, which assesses the accuracy of predicted survival probabilities, could provide a more comprehensive evaluation.

9.1.1 Conclusion

Overall, this study demonstrates the feasibility of predicting survival and RFS using a deep learning model based on clinical features. While the model's performance closely matches that of the CPH model, the latter shows a slight advantage in stratifying patients into low- or high-risk groups. This indicates the need for

further refinement of the deep learning model.

To improve the deep learning model and adress overfitting, several strategies can be implemented, such as simplifying the model architecture and exploring data augmentation. Simplifying the model architecture by reducing the number of layers or neurons can help to reduce overfitting. For example, the hidden layers could be reduced from three to two layers. In addition, the number of neurons in each layer could be decreased. Currently, the number of neurons are optimized between 1 and 100. Decreasing this upper limit may help in increasing model performance.

Data augmentation is another valuable technique that can increase the variability of the training set. This approach can help the model generalize better to unseen data. Approaches for data augmentation of tabular data mainly involve dealing with class imbalance or aiming to increase dataset size and diversity. Balancing the dataset through techniques like oversampling the minority class or undersampling the majority class will prevent the model from becoming biased towards the majority class, ensuring a fairer learning process [52]. More advanced techniques could also be used like Synthetic Minority Over-sampling Technique (SMOTE), where synthetic samples for the minority class are created by interpolating between existing samples based on their nearest neighbors [53]. Other data augmentation techniques that focus more on dataset size and diversity could be applying transformations to the features, like the logarithm. Previously described methods, like data augmentation and changes to model architecture might also help stabilize the model. In conclusion, further optimization of the model through for example architectural simplification and robust data augmentation techniques holds potential for developing an even better deep learning framework capable of using clinical features to predict survival and RFS effectively.

9.2 Models based on imaging

When predicting survival or RFS with only imaging data, the model performs no better than random guessing. Efforts are made to determine whether this low performance is due to the model itself or the imaging data.

First, to investigate if the amount of patients or the specific patient population is responsible for the low performance, the clinical data-based model has been trained and tested only on patients for whom imaging data is available. In this subset, the model has been unable to learn the survival probability, as indicated by a C-index no higher than 0.5, which indicates random guessing. When the clinical data-based model has been trained on all patients except those for whom imaging was available and tested on those patients, the model has performed similarly to the other splits used in cross-validation. These findings suggest that the dataset size might be the primary issue. This finding is not surprising, as large datasets are crucial for training accurate models, particularly in tasks like predicting patient outcomes such as survival [54]. In survival analysis, typically not all patients have experienced the event during the study period, making it challenging for the model to learn effectively from limited data, potentially requiring more samples to adequately represent the survival event.

Secondly, it has been tested whether changing the batch size to 1, which is necessary for the model based on imaging data due to GPU memory constraints, has been causing the problem. In the model based on clinical characteristics, a batch size of the entire training dataset has been used. This necessary adjustment to 1 may have led to some model instability, and therefore it is possible that the model has not been able to make predictions when only a batch size of 1 is used. When using a batch size of 1 in the model based on clinical characteristics, the C-index has fluctuated around 0.5 even when trained for a long time, indicating that the model has not performed well with such a small batch size. A small batch size can lead to instability in model training due to noisier gradient estimates [55]. This instability arises because small batches provide high variance in gradient estimates, resulting in a less reliable optimization process. Since each gradient update is variable, the model could have struggled to converge to the global minimum, which may have led to poorer overall performance. There are still several adjustments that could have been made to the model, like gradient accumulation to address these issues. However, these previous findings suggest that the complexity of the problem may have exceeded the capacity of the model to learn from the available data, especially since the model based on clinical characteristics has also performed poorly trained on only the subset of patients with imaging data. Therefore, the problem has been simplified, prioritizing the demonstration of the feasibility of using deep learning models with imaging data to predict the risk of GIST patients.

To simplify the problem, the survival prediction model has been changed into a classification problem, in which the patients for whom imaging data is available are divided into low-, high-, and intermediate-risk groups. Although the results show high accuracy for this model, the sensitivity and specificity show that the model assigns all samples to the low-risk group. The AUC is also low, indicating that the model is not good at classifying patients into the right risk group. A possible explanation could be that an imbalanced dataset is used. Way more patients belong to the low-risk group than the high-risk group, most likely introducing bias towards the low-risk group. When classification algorithms are applied to imbalanced data, aiming for overall accuracy often causes the model to favor the majority class (low-risk) and neglect the minority class (high-risk) [56]. This results in poor classification performance for the low-risk group.

To gain more insight into the model's decision-making process, a Grad-CAM analysis has been performed to visualize regions of interest for the model within the CT images. Although the analysis shows some focus on the abdominal region, suggesting that the model is learning specific features associated with low-risk classification, there is also high activation in the background of the image and non-tumor regions in the abdomen. This indicates that model predictions may be artifact-driven and therefore not reliable.

While Grad-CAM can offer valuable insights into the model's decision-making processes, it has some limitations and challenges that need to be taken into account [57]. It has been suggested [57] that some saliency methods, like Grad-CAM, may implicitly be implementing unsupervised image processing techniques like edge detection or denoising, which are not considered explanation methods. To differentiate these methods from model-sensitive explanations, visual inspection is insufficient, and therefore conclusions drawn from this analysis may not be reliable.

9.2.1 Conclusion

Imaging data may contain more relevant information that could be useful to make an even better prediction on survival and RFS. However, the survival prediction model performs no better than random guessing, and the classification model categorizes all patients as low-risk, failing to differentiate between low- and high-risk groups.

To investigate whether imaging data can actually help in predicting survival and RFS, more imaging data of patients should be added. Furthermore, the Grad-CAM analysis shows that the model's predictions are not based on relevant tumor features which also indicates that the model is not performing well. It would be useful to look into the pre-processing steps. Segmenting the images could further improve the performance of the model by making sure the model is making its prediction based on imaging features of the tumor itself. This could also more simply be achieved by using a boundary box.

Additionally, employing class balancing techniques can help address the dataset's imbalance. Ensuring a more equal distribution of classes will mitigate bias and enable the model to learn from both low- and high-risk patient groups effectively. Furthermore, data augmentation techniques can be used to generate synthetic data from existing samples, augmenting the dataset size and diversity. Data augmentation techniques that could be used are for example rotation and flipping of the images [58]. This approach could be useful in training the model on a more comprehensive dataset, improving its robustness and generalization capabilities. Furthermore, additional images of data of the patients of which imaging data is already available could be used. Currently, only one image is used for each patient with best tumor visibility. However, using other images in which the tumor is also visible in acceptable resolution, helps to increase the dataset size. In conclusion, addressing these challenges through dataset expansion, refined pre-processing, class balancing, and data augmentation techniques holds promise for evaluation of the usefulness of image data in predicting survival and RFS.

9.3 Limitations

There are several limitations of this study. First, the interpretability and the ability to stratify risk of the model based on clinical characteristics have only been tested using one single train-test split, while cross validation would have provided a more accurate assessment of the performance. Although the C-indices were relatively similar across different train-test splits, it is still not guaranteed that the performance of

risk stratification and interpretability were also generalizable across the different train-test splits. This is especially important, since the C-index measures a model's ability to rank survival times accurately [51] and not the model's risk stratification abilities. In addition, the imaging-based deep learning survival model was also only evaluated using one specific train-test split. Therefore it is possible, that an unlucky train-test split was used causing poor performance of the model. However, for the classification model based on the same imaging data, cross validation was applied and this model gave a low AUC across all folds. This is an indication that the poor performance of the imaging-based deep learning survival model was not only caused by an unlucky train-test split.

Furthermore, the imaging data used in this study were only obtained from two clinical centers (EMC and AVL), which may limit the generalizability of the findings to the entire population. The variability in imaging protocols, equipment, and patient demographics across different centers could influence the model's performance and its applicability to diverse clinical settings.

Finally, applicable to both the imaging and clinical models, the retrospective nature of the study introduces biases. Retrospective studies rely only on existing data, which may not capture all relevant variables or reflect current clinical practices. This limitation can affect the model's accuracy and its ability to predict future outcomes accurately.

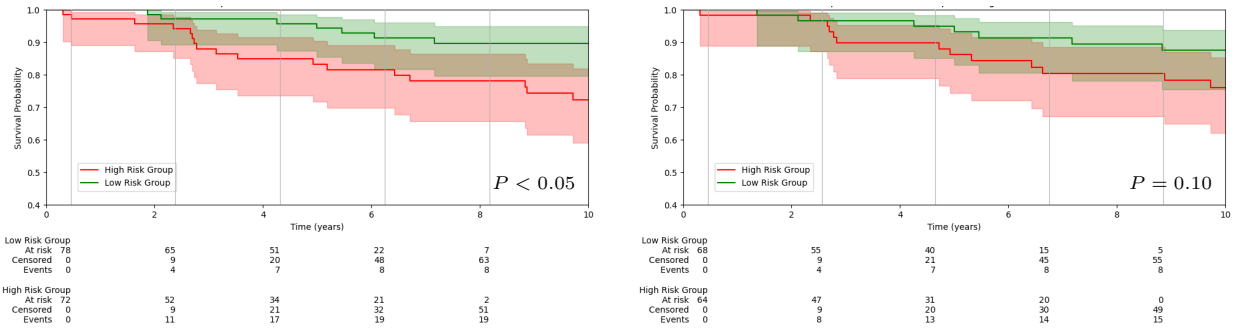
9.4 Future works

The final goal is to make a model that incorporates both imaging as clinical data to make a prediction on the survival or RFS. Both imaging data as non-imaging data could hold important information that might help to make a correct prognosis prediction [59]. Imaging data, can provide detailed insights into tumor characteristics, such as shape, texture, and heterogeneity [8], which might not be fully captured by clinical data alone. Combining both clinical and imaging data in deep learning methods could therefore help to make an even better prognosis prediction. Deep learning architectures designed in this study may be useful to incorporate into this final model. There are however still some challenges in the development of this model. Besides improvement of the model based on clinical and imaging features individually, there are several strategies that can be used for combining both clinical and imaging data.

One approach is to use a multi-modal deep learning architecture that can process and integrate both types of data simultaneously. For instance, separate branches of the network could be designed to handle imaging data (e.g. using convolutional neural networks) and clinical data (e.g. using fully connected layers), with a fusion layer that combines the learned features before making a final prediction [59]. A different strategy involves the use of attention mechanisms to weigh the importance of different data types dynamically [60]. This can help the model to focus on the most relevant features from both datasets.

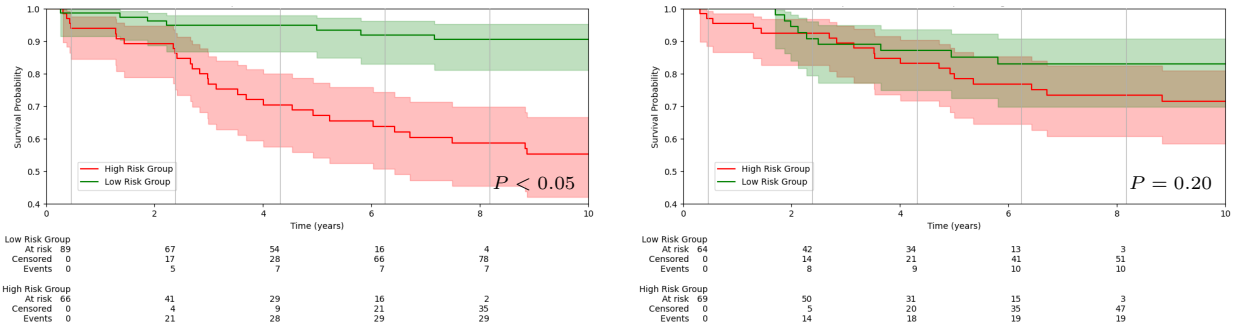
While the model is still in early development, the final goal is clinical translation. Collaborations with domain experts, such as radiologists and oncologists, will be essential to validate the model's findings and ensure its clinical relevance. By addressing these challenges, we can move closer to developing a comprehensive and accurate predictive model that leverages the full spectrum of available data, ultimately improving patient outcomes and advancing personalized medicine.

S1 Supplementary material



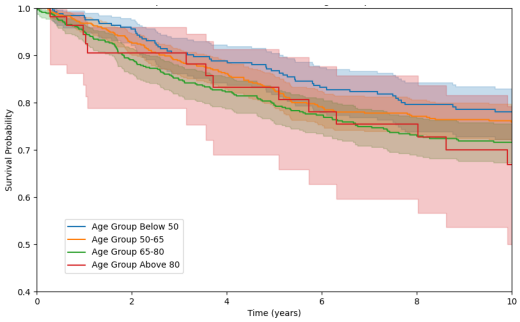
(a) CPH model *(b) Deep learning model*

Figure S1: Kaplan–Meier plots showing high- and low-risk groups for survival prediction for the minimal feature set. The high-risk (red line) and low-risk (green line) patients from each model were divided based on the upper and lower 25% of the risk predictions of the train set. The P-value for the log-rank test comparing the low- and high-risk group is shown in the Figure. The number of patients at risk at different points in time are shown for both groups below the Figure.

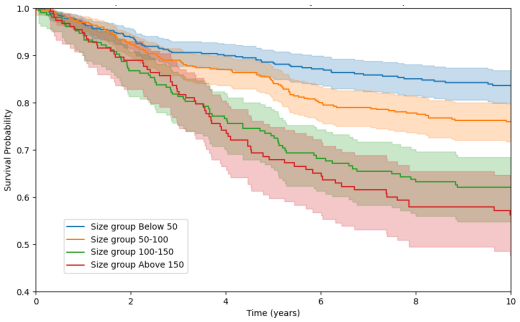


(a) CPH model *(b) Deep learning model*

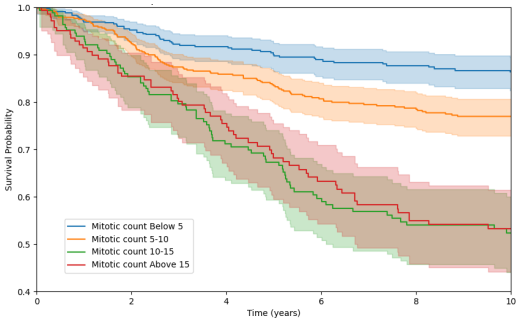
Figure S2: Kaplan–Meier plots showing high- and low-risk groups for RFS prediction for the minimal feature set. The high-risk (red line) and low-risk (green line) patients from each model were divided based on the upper and lower 25% of the risk predictions of the train set. The P-value for the log-rank test comparing the low- and high-risk group is shown in the Figure. The number of patients at risk at different points in time are shown for both groups below the Figure.



(a) The influence of different age

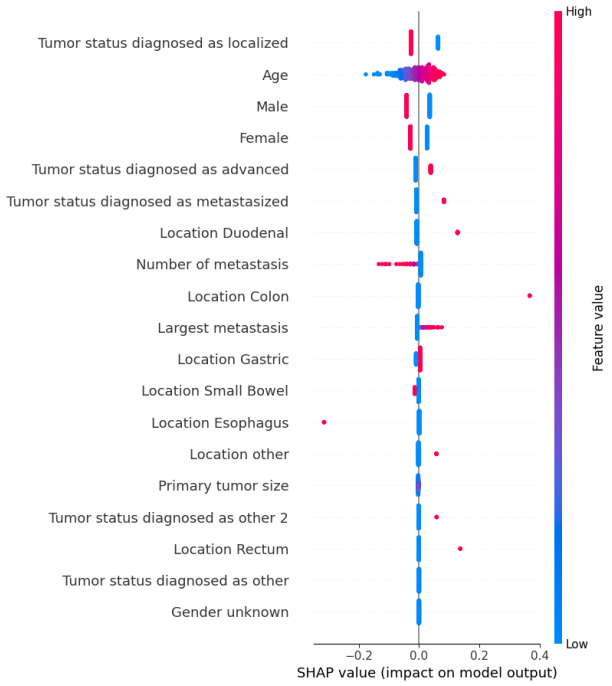


(b) The influence of different primary tumor sizes

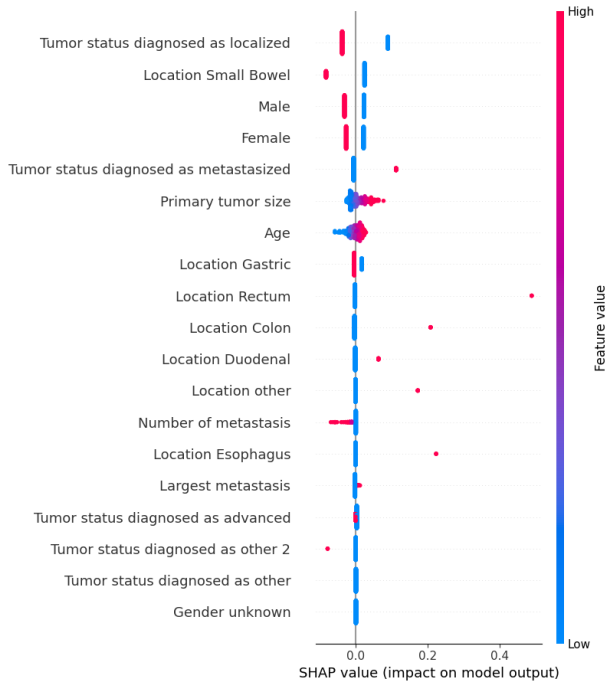


(c) The influence of different mitotic count

Figure S3: Kaplan-Meier curves showing the influence of feature change of the features that significantly (P <0.05) influence the RFS based on the CPH model



(a) Survival prediction



(b) RFS prediction

Figure S4: Shapley analysis of the deep learning model for the minimal feature set. The mean shapley value of the test set is shown. The features are ranked from most important to least important. High feature values are shown in red and low feature values are shown in blue.

References

[1] H. R. Unalp, H. Derici, E. Kamer, A. D. Bozdag, E. Tarcan, and M. A. Onal, "Gastrointestinal stromal tumours: outcomes of surgical management and analysis of prognostic variables," *Can. J. Surg.*, vol. 52, pp. 31–38, Feb. 2009.

[2] C. L. Corless, J. A. Fletcher, and M. C. Heinrich, "Biology of gastrointestinal stromal tumors," *Journal of Clinical Oncology*, vol. 22, pp. 3813–3825, Sept. 2004.

[3] T. M. Parab, M. J. DeRogatis, A. M. Boaz, S. A. Grasso, P. S. Issack, D. A. Duarte, O. Urayeneza, S. Vahdat, J.-H. Qiao, and G. S. Hinika, "Gastrointestinal stromal tumors: a comprehensive review," *Journal of Gastrointestinal Oncology*, vol. 10, pp. 144–154, Feb. 2018.

[4] E. Wu, S. Y. Son, V. Gariwala, and C. O'Neill, "Gastric gastrointestinal stromal tumor (gist) with co-occurrence of pancreatic neuroendocrine tumor," *Radiology Case Reports*, vol. 16, p. 1391–1394, June 2021.

[5] K. Akahoshi, M. Oya, T. Koga, and Y. Shiratsuchi, "Current clinical management of gastrointestinal stromal tumor," *World Journal of Gastroenterology*, vol. 24, pp. 2806–2817, July 2018.

[6] X. Zhao and C. Yue, "Gastrointestinal stromal tumors," *Journal of Gastrointestinal Oncology*, vol. 3, pp. 189–208, Apr. 2012.

[7] H. Joensuu, "Risk stratification of patients diagnosed with gastrointestinal stromal tumor," *Human Pathology*, vol. 39, pp. 1411–1419, Oct. 2008.

[8] A. Inoue, S. Ota, M. Yamasaki, B. Batsaikhan, A. Furukawa, and Y. Watanabe, "Gastrointestinal stromal tumors: a comprehensive radiological review," *Japanese Journal of Radiology*, vol. 40, pp. 1105–1120, July 2022.

[9] C. Zhou, X. Duan, X. Zhang, H. Hu, D. Wang, and J. Shen, "Predictive features of CT for risk stratifications in patients with primary gastrointestinal stromal tumour," *European Radiology*, vol. 26, pp. 3086–3093, Dec. 2015.

[10] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. de Jong, J. van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, pp. 749–762, Oct. 2017.

[11] R. Cannella, L. L. Grutta, M. Midiri, and T. V. Bartolotta, "New advances in radiomics of gastrointestinal stromal tumors," *World Journal of Gastroenterology*, vol. 26, pp. 4729–4738, Aug. 2020.

[12] B. Kang, X. Yuan, H. Wang, S. Qin, X. Song, X. Yu, S. Zhang, C. Sun, Q. Zhou, Y. Wei, F. Shi, S. Yang, and X. Wang, "Preoperative CT-based deep learning model for predicting risk stratification in patients with gastrointestinal stromal tumors," *Frontiers in Oncology*, vol. 11, Sept. 2021.

[13] K. Søreide, O. M. Sandvik, J. A. Søreide, V. Giljaca, A. Jureckova, and V. R. Bulusu, "Global epidemiology of gastrointestinal stromal tumours (GIST): A systematic review of population-based cohort studies," *Cancer Epidemiology*, vol. 40, pp. 39–46, Feb. 2016.

[14] J. J. Cuaron, K. A. Goodman, N. Lee, and A. J. Wu, "External beam radiation therapy for locally advanced and metastatic gastrointestinal stromal tumors," *Radiation Oncology*, vol. 8, Nov. 2013.

[15] M. Barat, A. Pellat, A. Dohan, C. Hoeffel, R. Coriat, and P. Soyer, "CT and MRI of gastrointestinal stromal tumors: New trends and perspectives," *Canadian Association of Radiologists Journal*, June 2023.

[16] M. C. Heinrich, C. L. Corless, G. D. Demetri, C. D. Blanke, M. von Mehren, H. Joensuu, L. S. McGreevey, C.-J. Chen, A. D. Van den Abbeele, B. J. Druker, B. Kiese, B. Eisenberg, P. J. Roberts, S. Singer, C. D. Fletcher, S. Silberman, S. Dimitrijevic, and J. A. Fletcher, "Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor," *Journal of Clinical Oncology*, vol. 21, p. 4342–4349, Dec. 2003.

[17] C. L. Corless and M. C. Heinrich, "Molecular pathobiology of gastrointestinal stromal sarcomas," *Annual Review of Pathology: Mechanisms of Disease*, vol. 3, pp. 557–586, Feb. 2008.

[18] J. Martín-Broto, L. Rubio, R. Alemany, and J. A. López-Guerrero, "Clinical implications of KIT and PDGFRA genotyping in GIST," *Clinical and Translational Oncology*, vol. 12, pp. 670–676, Oct. 2010.

[19] L. F. Lopes, R. B. West, L. M. Bacchi, M. van de Rijn, and C. E. Bacchi, "DOG1 for the diagnosis of gastrointestinal stromal tumor (GIST): Comparison between 2 different antibodies," *Applied Immunohistochemistry &amp Molecular Morphology*, vol. 18, pp. 333–337, July 2010.

[20] K. Jansen, N. Farahi, F. Büscheck, M. Lennartz, A. M. Luebke, E. Burandt, A. Menz, M. Kluth, C. Hube-Magg, A. Hinsch, D. Höflmayer, S. Weidemann, C. Fraune, K. Möller, P. Lebok, G. Sauter, R. Simon, R. Uhlig, W. Wilczak, F. Jacobsen, S. Minner, R. Krech, T. Clauditz, C. Bernreuther, D. Dum, T. Krech, A. Marx, and S. Steurer, "DOG1 expression is common in human tumors: A tissue microarray study on more than 15, 000 tissue samples," *Pathology - Research and Practice*, vol. 228, p. 153663, Dec. 2021.

[21] W. Swalchick, R. Shamekh, and M. M. Bui, "Is DOG1 immunoreactivity specific to gastrointestinal stromal tumor?," *Cancer Control*, vol. 22, pp. 498–504, Oct. 2015.

[22] Y.-M. Wang, "Succinate dehydrogenase-deficient gastrointestinal stromal tumors," *World Journal of Gastroenterology*, vol. 21, no. 8, p. 2303, 2015.

[23] K. Jašek, B. Váňová, M. Grendár, A. Štanclová, P. Szépe, A. Hornáková, V. Holubeková, L. Plank, and Z. Lasabová, "BRAF mutations in KIT/PDGFRA positive gastrointestinal stromal tumours (GISTs): Is their frequency underestimated?," *Pathology - Research and Practice*, vol. 216, p. 153171, Nov. 2020.

[24] I. Hostein, N. Faur, C. Primois, F. Boury, J. Denard, J.-F. Emile, P.-P. Bringuier, J.-Y. Scoazec, and J.-M. Coindre, "iBRAF/imutation status in gastrointestinal stromal tumors," *American Journal of Clinical Pathology*, vol. 133, pp. 141–148, Jan. 2010.

[25] Z. Cao, J. Li, L. Sun, Z. Xu, Y. Ke, B. Shao, Y. Guo, and Y. Sun, "GISTs with NTRK gene fusions: A clinicopathological, immunophenotypic, and molecular study," *Cancers*, vol. 15, p. 105, Dec. 2022.

[26] C.-E. Wu, C.-Y. Tzen, S.-Y. Wang, and C.-N. Yeh, "Clinical diagnosis of gastrointestinal stromal tumor (GIST): From the molecular genetic point of view," *Cancers*, vol. 11, p. 679, May 2019.

[27] M. P. Starmans, S. R. van der Voort, J. M. C. Tovar, J. F. Veenland, S. Klein, and W. J. Niessen, "Radiomics," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 429–456, Elsevier, 2020.

[28] M. P. A. Starmans, M. J. M. Timbergen, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, R. S. Dwarkasing, F. E. J. A. Willemssen, W. J. Niessen, C. Verhoef, S. Sleijfer, J. J. Visser, and S. Klein, "Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach," *Journal of Digital Imaging*, vol. 35, pp. 127–136, Jan. 2022.

[29] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685–695, Apr. 2021.

[30] F. Bre, J. M. Gimenez, and V. D. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings*, vol. 158, pp. 1429–1441, Jan. 2018.

[31] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang, "Deep learning: A primer for radiologists," *RadioGraphics*, vol. 37, pp. 2113–2131, Nov. 2017.

[32] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, p. 611–629, June 2018.

[33] N. Shahriar, "What is Convolutional Neural Network—CNN (Deep Learning) — nafizshahriar.medium.com." `https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5`. [Accessed 08-11-2023].

[34] P. M. Cheng, E. Montagnon, R. Yamashita, I. Pan, A. Cadrin-Chênevert, F. P. Romero, G. Chartrand, S. Kadoury, and A. Tang, "Deep learning: An update for radiologists," *RadioGraphics*, vol. 41, pp. 1427–1445, Sept. 2021.

[35] "dutchsarcomagroup.nl." `https://www.dutchsarcomagroup.nl/over-sarcomen/dutch-gist-consortium-dgc/`. [Accessed 02-04-2024].

[36] P. Li, E. A. Stuart, and D. B. Allison, "Multiple imputation: A flexible tool for handling missing data," *JAMA*, vol. 314, p. 1966, Nov. 2015.

[37] P. Schober and T. R. Vetter, "Survival analysis and interpretation of time-to-event data: The tortoise and the hare," *Anesthesia amp; Analgesia*, vol. 127, p. 792–798, Sept. 2018.

[38] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, "Deep learning for survival analysis: a review," *Artificial Intelligence Review*, vol. 57, Feb. 2024.

[39] 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, Feb. 2018.

[40] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in Medicine*, vol. 14, p. 73–82, Jan. 1995.

[41] M. Schmid, M. N. Wright, and A. Ziegler, "On the use of harrell's c for clinical risk prediction via random survival forests," *Expert Systems with Applications*, vol. 63, p. 450–459, Nov. 2016.

[42] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, July 2017.

[43] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," *Expert Systems with Applications*, vol. 182, p. 115222, Nov. 2021.

[44] "[Tutorial] - Cross Validation & Nested CV — kaggle.com." `https://www.kaggle.com/code/jacoporepossi/tutorial-cross-validation-nested-cv`. [Accessed 23-05-2024].

[45] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis," *Healthcare Analytics*, vol. 4, p. 100218, Dec. 2023.

[46] S. Szeghalmy and A. Fazekas, "A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning," *Sensors*, vol. 23, p. 2333, Feb. 2023.

[47] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, International Joint Conferences on Artificial Intelligence Organization, Aug. 2017.

[48] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017.

[49] S. Masoudi, S. A. Harmon, S. Mehralivand, S. M. Walker, H. Raviprakash, U. Bagci, P. L. Choyke, and B. Turkbey, "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research," *Journal of Medical Imaging*, vol. 8, Jan. 2021.

[50] L. B. de Amorim, G. D. Cavalcanti, and R. M. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109924, Jan. 2023.

[51] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, p. 1105–1117, Jan. 2011.

[52] C. Xie, R. Du, J. W. Ho, H. H. Pang, K. W. Chiu, E. Y. Lee, and V. Vardhanabhuti, "Effect of machine learning re-sampling techniques for imbalanced datasets in 18f-fdg pet-based radiomics model on prognostication performance in cohorts of head and neck cancer patients," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 47, p. 2826–2835, Apr. 2020.

[53] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, p. 122778, June 2024.

[54] Z. Obermeyer and E. J. Emanuel, "Predicting the future — big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, p. 1216–1219, Sept. 2016.

[55] Y. Zhang, H. Qu, C. Chen, and D. Metaxas, "Taming the noisy gradient: Train deep neural networks with small batch sizes," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019, International Joint Conferences on Artificial Intelligence Organization, Aug. 2019.

[56] Q. Li, C. Zhao, X. He, K. Chen, and R. Wang, "The impact of partial balance of imbalanced dataset on classification performance," *Electronics*, vol. 11, p. 1322, Apr. 2022.

[57] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Red Hook, NY, USA), p. 9525–9536, Curran Associates Inc., 2018.

[58] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, Dec. 2022.

[59] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, Oct. 2020.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.