PIDEEL: METABOLIC PATHWAY-INFORMED DEEP LEARNING MODEL FOR SURVIVAL ANALYSIS AND PATHOLOGICAL CLASSIFICATION OF GLIOMAS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Gün Kaynar

June 2024

PiDeeL: metabolic pathway-informed deep learning model for survival analysis and pathological classification of gliomas
By Gün Kaynar
June 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek(Advisor)

Can Alkan

Gülden Olgun

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

PIDEEL: METABOLIC PATHWAY-INFORMED DEEP LEARNING MODEL FOR SURVIVAL ANALYSIS AND PATHOLOGICAL CLASSIFICATION OF GLIOMAS

Gün Kaynar
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
June 2024

The real-time assessment of tumor characteristics during surgery is crucial, offering the potential to create a feedback mechanism for surgeons. Such feedback enables surgeons to make more informed decisions regarding the extent of tumor resection and whether to adopt a more aggressive or conservative approach. Although metabolomics-based tumor pathology prediction methods exist, their predictive performance is often constrained by the limited size of available datasets. Additionally, the feedback about the tumor tissue could be enhanced in terms of content and accuracy. This thesis introduces PiDeeL, a metabolic pathway-informed deep learning model designed to perform survival analysis and pathology assessment based on metabolite concentrations. We demonstrate that integrating pathway information into the model architecture significantly reduces parameter complexity while enhancing survival analysis and pathological classification performance. Our results indicate that PiDeeL improves tumor pathology prediction performance over state-of-the-art methods, achieving a 3.38% improvement in the Area Under the ROC Curve and a 4.06% improvement in the Area Under the Precision-Recall Curve. Similarly, regarding the time-dependent concordance index (c-index), PiDeeL exhibits superior survival analysis performance, with a 4.3% improvement compared to current leading methods. Furthermore, importance analyses conducted on input metabolite features and pathway-specific neurons of PiDeeL provide valuable insights into tumor metabolism. Applying this model in the surgical setting will assist surgeons in dynamically adjusting their surgical plans, ultimately leading to more accurate prognosis estimates tailored to the specifics of the surgical procedure.

Keywords: Nuclear magnetic resonance spectroscopy, Metabolomics, Deep learning, Intraoperative feedback, Survival analysis, Pathways.

ÖZET

PIDEEL: GLİOMALARIN SAĞKALIM ANALİZİ VE PATOLOJİK SINIFLANDIRMASI IÇİN METABOLİK YOLAK BİLGİLİ DERİN ÖĞRENME MODELİ

Gün Kaynar
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Haziran 2024

Cerrahi operasyon sırasında tümör özelliklerinin anlık olarak değerlendirilmesi önemlidir ve cerraha işlem sırasında geri bildirim mekanizması kurma potansiyeline sahiptir. Bu tür geri bildirimlerin mevcut olması durumunda, cerrahlar tümörün rezeksiyonu konusunda daha esnek veya daha sıkı olmaya karar verebilirler. Metabolomik tabanlı tümör patolojisi tahmini yapmak için yöntemler olsa da, model karmaşıklığı ve tahmin performansı küçük veri seti boyutlarıyla sınırlıdır. Ayrıca tümör dokusuna sağlanan geri bildirimlerin aktardığı bilgiler hem içerik hem de doğruluk açısından geliştirilebilir. Bu çalışmada, metabolit konsantrasyonlarına dayalı olarak hayatta kalma analizi ve patoloji değerlendirmesi gerçekleştirmek için metabolik yolak bilgili bir derin öğrenme modeli (PiDeeL) öneriyoruz. Yolak bilgisinin model mimarisine dahil edilmesinin parametre karmaşıklığını önemli ölçüde azalttığını ve daha iyi hayatta kalma analizi ve patolojik sınıflandırma performansı sağladığını gösteriyoruz. Bu tasarım kararlarıyla, PiDeeL'in, AUROC açısından tümör patolojisi tahmin performansını %3,38 ve AUPR açısından %4,06 oranında geliştirdiğini gösteriyoruz. Benzer şekilde, zamana bağlı uyumluluk indeksi (c-endeksi) açısından PiDeeL'in, en son teknolojiyle karşılaştırıldığında daha iyi sağkalım analizi performansı (%4,3 oranında iyileşme) elde ettiğini gösteriyoruz. Ayrıca, PiDeeL'in yolaklara özgü nöronlarının yanı sıra girdi metabolit özellikleri üzerinde gerçekleştirilen önem analizlerinin tümör metabolizmasına ilişkin bilgiler sağladığını da gösteriyoruz. Bu modelin ameliyathanede kullanılmasının, cerrahların ameliyat planını anında ayarlamalarına yardımcı olacağını ve o anki cerrahi prosedürlere uygun daha iyi prognoz tahminleri elde edilmesini sağlayacağını öngörüyoruz.

Anahtar sözcükler: Nükleer Manyetik Rezonans Spektroskopisi, Metabolomik, Derin öğrenme, Cerrahi geri bildirim, Sağkalım analizi, Yolaklar.

Acknowledgement

Bana karşı sonsuz güveni ve anlaşıyı, her soruna bulduğu çözümleri ve karizmasıyla bana araştırmacı olmayı öğreten danışmanım Ercüment Çiçek'e,

Derin bilgisiyle projelerimde nerede takılsam hemen ofisine gidip soruma yanıt alarak çıktığım ikinci araştırma danışmanım Can Alkan'a,

Tez komitemin değerli üyesi Gülden Olgun'a

Her konuda desteğiyle şimdilerde en sevdiğim iş olan algoritmaları bana öğreten hocam Uğur Doğrusöz'e,

Neşesi ve esprileriyle her konuştuğumuzda yeni bir problem ve çözümünü öğreten hocam Cevdet Aykanat'a,

Hiçbir yerde bulunamayacak kadar dolu bir Deep Learning dersini Bilkent'te veren hocam Hamdi Dibeklioğlu'na,

Her başım sıkıştığında koştuğum üstat Serkan Demirci'ye,

İlk makalemi yazarken birkaç defa beraber sabahladığımız dostum Doruk Çakmakçı'ya,

Beraber yediğimiz yemekleri unutmak mümkün olmayan değerli dostlarım Kutay Taşcı ve Arçin Ergüzen'e

Ofise gelmemin en büyük sebepleri değerli dostlarım Sina Barazandeh ve Alper Yılmaz'a,

Ev arkadaşlıkları ile Bilkent'i güzelleştiren sevgili dostlarım Aydamir Mirzayev ve Pedram Tootoonchian'a,

Lojman partilerimiz ve kutu oyunu gecelerimizle Ankara'nın soğuğunu dindiren dostlarım Mousa Azari, Güneş Tok, Umut Taylan ve Melis Karaca'ya,

Akademik her konuda bana yol gösteren dayım Mehmet Koyutürk'e,

Güvenlerini ve sevgilerini hep hissettiğim, önce Boğaziçi'ne, ardından Bilkent'e gitmemde ve Carnegie Mellon'dan kabul almamda her türlü katkısı olan annem Dilek Kaynar, babam Alp Kaynar, kardeşim Tan Kaynar ve ananem Gül Koyutürk'e,

İnceliği ve hoş düşüncesi ile İstanbul'u, Ankara'yı ve Pittsburgh'u güzelleştiren sevgilim Cansın'a,

Sonsuz teşekkürler.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Gliomas have uncertain mortality rates due to variations in suspected brain cell types and differences in cancer progression dynamics [1]. The 2021 World Health Organization (WHO) classification of central nervous system tumors [2] categorizes gliomas based on cell types (e.g., astrocytes, oligodendrocytes) and four pathological grades (I-IV), considering histopathologic, molecular, and genetic characteristics. These grades reflect increasing malignancy and decreasing prognostic optimism. Given the prevalence of gliomas among adult brain tumors [3], developing a surgical management strategy that supports personalized clinical decision-making through intraoperative or preoperative feedback is essential.

To maximize the resection of tumor tissue, various imaging [4], spectroscopy, mass spectrometry [5, 6, 7, 8, 9, 10, 11, 12, 13], and optical spectrometry techniques [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26] have been proposed. Among these, ^{1}H High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HRMAS NMR) spectroscopy measures metabolic activity in unprocessed tissue samples [27] within 20 minutes, producing a time domain signal (Free Induction Decay - FID) composed of decaying exponential functions. The FID signal is preprocessed to obtain the HRMAS NMR spectrum in the frequency domain.

Post-preprocessing, the HRMAS NMR spectrum is suitable for analyzing residual tumor tissue in the excision cavity, providing feedback to surgeons. This feedback loop requires a pathologist knowledgeable about tumor metabolism and an NMR technician to identify key metabolites. However, this method is limited by expert availability during surgery and the manual quantification process, which is slow and prone to issues with overlapping metabolite peaks [28].

Due to the high-dimensional and complex nature of preprocessed HRMAS NMR spectra, data-driven computational modeling (e.g., machine learning) can complement the manual feedback loop. Various supervised machine-learning models of glioma tissue samples have been explored for automated pathological classification [29, 30]. The dataset used in these studies [31] includes HRMAS NMR spectra from 568 glioma and healthy control samples, which, despite being the largest of its kind, is relatively small, limiting model complexity and performance.

Improving deep learning models in small data regimes involves making informed architectural choices to simplify models. Incorporating prior biological information can reduce neuron connectivity, mitigating overfitting and enhancing interpretability. Though not applied in metabolomics, biologically informed deep learning architectures have been successful in other domains, such as metastasis prediction in prostate cancer [32], phenotype prediction from genetic variants [33], survival prediction in glioblastoma [34], and multi-modal data integration [35].

Pathological classification alone may not fully determine cancer progression. In glioblastoma multiforme (GBM), a grade IV astrocytic glioma, prognosis is typically poor due to its malignancy [1]. However, prolonged survival has been observed in IDH wildtype GBM [36, 37, 38], suggesting the need for survival risk estimates in surgical feedback. Recent deep learning models for survival prediction include DeepSurv [39], DeepHit [40], PC-Hazard [41], and Surv-Net [42].

We propose a deep neural network architecture (PiDeeL) to predict glioma patient survival and pathological classification using HRMAS NMR signals to provide surgeons with feedback (Figure 1.1). While automated pathological classification methods exist, this is the first study to combine these two sources of information for surgical feedback. We incorporate prior biological information from metabolic networks into the neural network architecture to enhance prediction interpretability, which is crucial for surgeons. PiDeeL outperforms state-of-the-art models in pathological classification and achieves a c-index of 68.7% for survival prediction. We also identify important metabolites and pathways, discussing their relevance in gliomas.

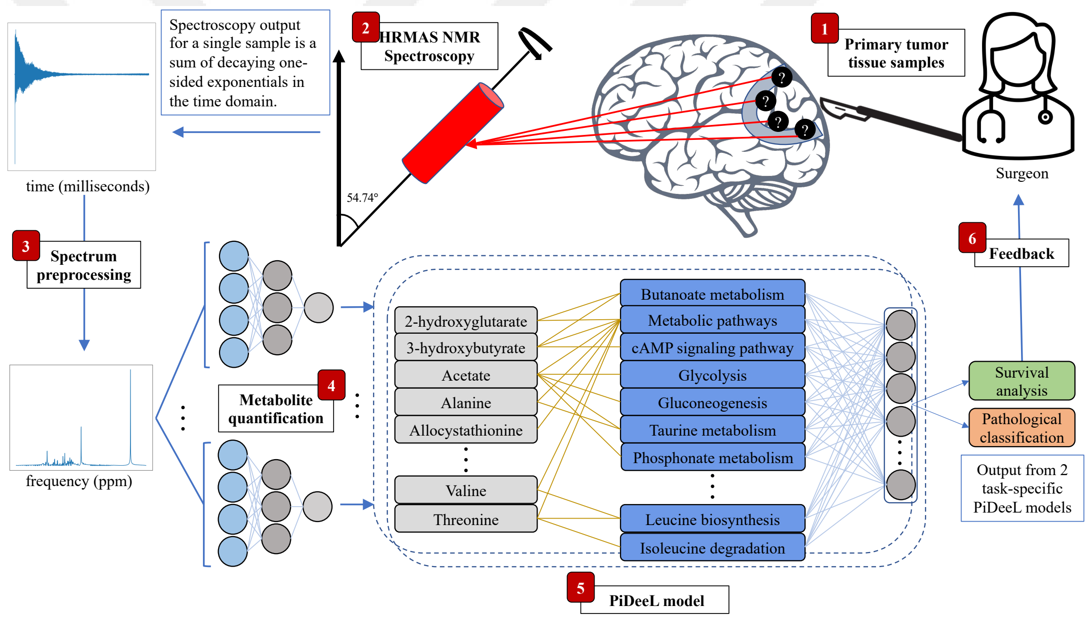

Figure 1.1: PiDeeL assisted intraoperative feedback loop. (1) The surgeon removes primary tumor tissue from the patient's brain and prepares specimens from the resected tissue. (2) The prepared samples are sent for HRMAS NMR spectroscopy analysis. (3) Time domain signals from the spectroscopy of each sample are preprocessed to reveal metabolite peaks in the corresponding frequency domain spectrum. (4) Thirty-seven metabolites are quantified from the preprocessed spectra in milliseconds using metabolite-specific two-layer fully-connected neural networks. (5) PiDeeL converts metabolite concentrations into survival analysis and pathological classification predictions, also in milliseconds. (6) Based on this feedback, the surgeon decides whether to stop or continue the surgery, guided by PiDeeL's predictions.

Chapter 2

Methods

2.1 Dataset

The dataset used in this thesis is a subset of the patient cohort available at `https://zenodo.org/record/5781769`. It includes 568 HRMAS NMR spectra from 458 glioma samples (80.7% aggressive and 19.3% benign) and 110 control tissue specimens, all collected during surgeries on 513 patients at University Hospitals of Strasbourg.

2.1.1 Tissue Sample Collection

All tissue specimens were collected either through a pneumatic system linking the neurosurgery operating theater to the NMR room (Hautepierre Hospital, University Hospitals of Strasbourg) or from samples stored at the tumor biobanks in Strasbourg and Colmar, with minimal ischemic delays (average time 2 ± 1 min). The collected samples were suitable for proper HRMAS NMR analysis and had a low necrosis-to-tumor ratio. After HRMAS NMR analysis, inserts for half the content of each sample were cut. The percentage of tumor cells in the sample was calculated relative to the total surface area as well as frozen hematoxylin and eosin-stained sections. Refer to Supplementary Table C.1 for the histopathological classification of the tumor tissue samples.

2.1.2 Ethics Statement

The tissue specimens used in this thesis were gathered from samples stored in the tumor biobanks of Strasbourg and Colmar, or through a pneumatic system. The collection, storage, and compilation of the dataset were approved by the Ethics Committee (no. 2003-100), Ethics Committee (no. 2013-37, 12.11.2013), and the Ethics Committee of Strasbourg (Comité de protection des personnes, Est IV). Written consent was obtained from all patients who participated in this study.

2.1.3 HRMAS NMR Spectrum Acquisition

A Bruker Avance III 500 spectrometer, equipped with a 4-mm triple-resonance gradient HRMAS probe (1H, 13C, and 31P) operating at a proton frequency of 500.13 MHz, was used to acquire spectra from biopsy samples. The samples were prepared by combining $30\mu L$ KelF inserts with 15- to 18-mg tissue samples at 20°C. To set the lock frequency of the spectrometer, $10\mu L$ of D_2O was added to the insert. Spectra were acquired using a Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence with a $285\mu s$ inter-pulse delay and a 10-minute acquisition time. The CPMG pulse train lasted 93 ms due to 328 loops. During acquisition, the temperature was maintained at 4°C to minimize tissue degradation effects, resulting in a one-dimensional (1D) proton (1H) spectrum for each specimen. Additionally, a simulated Lorentzian-shaped ERETIC [43] signal, calibrated to 69.6 nmoles of protons, was digitally added to the spectrum at 10 ppm. This digitally added ERETIC signal also accounted for pulse width differences between the standard acetate sample and the biopsy sample, making the method insensitive to the salt effect [44]. The calibrated ERETIC signal enabled the quantification of all identified metabolites in the spectrum.

2.1.4 Preprocessing

To prepare the dataset for survival analysis, we included samples from primary tumor tissue specimens of glioma origin that had recorded dates of metabolomics-guided surgery and either the date of death or last follow-up. We excluded specimens resected from healthy control patients and excision cavities of glioma patients. The resulting dataset consists of 384 HRMAS NMR spectra, with one sample per glioma patient, comprising 299 (77.9%) aggressive and 85 (22.1%) benign tumors, with subtype and grade information detailed in Supplementary Table C.1. Of the patient cohort, 70.1% (269 patients) were reported deceased before January 2021. The remaining 115 patients were considered censored, as their survival status is unknown. The distribution of patient age and time-to-event for deceased and censored patients is shown in panels A, B, and C of Supplementary Figure B.1, respectively. Additionally, the ERETIC-CPMG HRMAS NMR FID signals for these 384 glioma patients, whose spectra were acquired as described in Section 2.1.3, are available at `https://zenodo.org/record/7228791`.

2.2 Automated Metabolite Quantification

The automated metabolite quantification pipeline described in [30] was employed to quantify the following 37 metabolites: 2-hydroxyglutarate, 3-hydroxybutyrate, Acetate, Alanine, Allocystathionine, Arginine, Ascorbate, Aspartate, Betaine, Choline, Creatine, Ethanolamine, GABA, Glutamate, Glutamine, Glutathione (GSH), Glycerophosphocholine, Glycine, hypo-Taurine, Isoleucine, Lactate, Leucine, Lysine, Methionine, Myo-inositol, NAL, NAA, o-Acetylcholine, Ornithine, Phosphocholine, Phosphocreatine, Proline, scyllo-Inositol, Serine, Taurine, Threonine, and Valine.

To prepare our HRMAS NMR spectra for this pipeline, we first removed the initial 70-time points in the FID signal corresponding to the spectroscopy machine's (Bruker, in our case) digital filter. The clipped time domain signal was then transformed to the frequency domain using a Discrete Fourier Transform without windowing and was phase-corrected after applying a 10 Hz line broadening. The magnitude of the resulting signal (ranging from -2 ppm to 12 ppm) was normalized by tissue sample mass and scaled using the constant acetate reference-based scaling factor provided in the cited paper. Finally, the scaled magnitude signal was binned into $1,401$ bins of 0.01 ppm width by summing the intensity of the frequencies mapped to each bin. The resulting signal served as input to the publicly available trained 2-layer fully-connected neural network-based quantification models for each metabolite of interest, available at `https://github.com/gunkaynar/targeted_brain_tumor_margin_assessment`.

2.3 Problem Formulation

In this thesis, we address pathological classification and survival analysis tasks utilizing the complete output of the metabolite quantification pipeline detailed in Section 2.2, along with tumor malignancy labels, patient-specific survival event indicators (i.e., censored or deceased), and time-to-event variables.

For a given sample i, we denote the feature vector, tumor malignancy label, survival event indicator, and time-to-event as $\widehat{M}^{(i)}$, $Y_m^{(i)}$, $Y_e^{(i)}$, and $Y_d^{(i)}$, respectively. The feature vector, $\widehat{M}^{(i)} \in \mathbb{R}^{37}$, is a 37-dimensional vector representing the full output of the automated metabolite quantification pipeline. Formally, let the predicted concentration of a specific metabolite j for a sample set of size N be $\widehat{M}_j{}^{(1:N)} \in \mathbb{R}^{N\times 1}$. The complete output of the automated metabolite quantification pipeline for 37 metabolites and a sample set of size N is denoted as the concatenation of all predicted quantification vectors across metabolites:

$$\widehat{M}^{(1:N)} = [\widehat{M}_1^{(1:N)}, \widehat{M}_2^{(1:N)}, \cdots, \widehat{M}_{37}^{(1:N)}] \in \mathbb{R}^{N\times 37} \tag{2.1}$$

The survival event indicator, $Y_e^{(i)} \in \{0, 1\}$, is a binary variable indicating whether the sample is from a deceased (i.e., 1) or a censored patient (i.e., 0). Time to survival event, $Y_d^{(i)} \in [0, \infty]$, is a non-negative continuous variable defined as the time (in days) between the metabolomics-guided tumor removal surgery and the date of death or last follow-up for deceased and censored patients, respectively. The tumor malignancy label, $Y_m^{(i)} \in \{0, 1\}$, indicates whether the sample is from a tumor with aggressive (i.e., 1) or benign (i.e., 0) pathology, similar to the formulations presented in [29, 30].

For our survival analysis task, we use the hazard function denoted as $\lambda(t, \widehat{M}^{(i)})$ for a given sample i, where t is the time-to-event (i.e., $Y_d^{(i)}$). Assuming proportional hazards from Cox's model [45], the hazard function factorizes [39, 46] into a time-dependent baseline hazard function, $h_0(t)$, and a sample-dependent risk function, $h(\widehat{M}^{(i)})$, as given in Eq 2.2.

$$\lambda(t, \widehat{M}^{(i)}) = h_0(t) \cdot e^{h(\widehat{M}^{(i)})} \tag{2.2}$$

We train survival analysis models based on this hazard function formulation on a dataset of N samples to learn a mapping function $f : \mathbb{R}^{37} \rightarrow [-\infty, \infty]$ such that $f(\widehat{M}^{(1:N)}) = h(\widehat{M}^{(1:N)})$. The trained models aim to predict risk scores for glioma patients based on their metabolic profile. For these models, we use $Y_e^{(i)}$ and $Y_d^{(i)}$ as labels.

In contrast, we train pathological classification models on a dataset of N samples to learn a mapping function $g : \mathbb{R}^{37} \rightarrow \{0, 1\}$ such that $g(\widehat{M}^{(1:N)}) = Y_m^{(1:N)}$, which is a binary classification problem similar to [29, 30].

2.4 Baseline Models

This section details the models benchmarked against PiDeeL for survival analysis and pathological classification tasks as formulated in Section 2.3.

2.4.1 Survival Analysis

For the survival analysis task, we use the following baseline models: Cox Proportional Hazards (Cox-PH), Component-wise Gradient Boosting (CWGB), Random

Survival Forest (RSF), DeepHit [40], PC-Hazard [41], and DeepSurv [39].

The Cox-PH model is a linear model that assumes proportional hazards, where the instantaneous hazard function is factored into a product of a baseline hazard function and a sample-dependent risk function, $h(\widehat{M}^{(i)})$, as explained in Section 2.3. Based on the formulation by [45], the risk function estimate for sample i, $\widehat{h}_\beta(\widehat{M}^{(i)})$, is provided by the linear model in Eq 2.3:

$$\widehat{h}_\beta(\widehat{M}^{(i)}) = \beta^t \widehat{M}^{(i)} \tag{2.3}$$

where $\widehat{M}^{(i)}$ is the feature vector of metabolite concentrations for sample i, and β is the coefficient vector. The model is fitted by maximizing the Cox partial likelihood [39, 45, 46], given in Eq 2.4:

$$L(\beta) = \prod_{i:Y_e^{(i)}=1} \frac{exp(\widehat{h}_\beta(\widehat{M}^{(i)}))}{\sum_{j:Y_d^{(j)}>Y_d^{(i)}} exp(\widehat{h}_\beta(\widehat{M}^{(j)}))} \tag{2.4}$$

The gradient boosting approach iteratively improves the prediction performance of an ensemble of base learners (e.g., linear models). In each round, new base learners are trained to minimize the gradient of prediction error of the current ensemble model, and the best-performing candidate is added to the ensemble. We focused on the component-wise variant of gradient boosting, which uses component-wise least squares as the base learner [47]. Under the proportional hazards assumption, the instantaneous hazard function is factored as given in Eq 2.2.

RSF [48] is a nonlinear ensemble method that constructs a predefined number of survival trees using random subsets of features and bootstraps of training data to estimate the cumulative hazard function, $\Lambda(t)$, defined in Eq 2.5. Internal nodes of survival trees are split to maximize the log-rank test [49, 50] statistic between resulting child nodes.

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{2.5}$$

Neural network-based approaches to survival analysis were introduced by [46]. More recently, deeper variants such as DeepSurv [39], DeepHit [40], and PC-Hazard [41] have been proposed.

DeepHit is a multi-layer perceptron that models discrete time and competing risks. Since we only have one event type, we use DeepHit-Single (hereafter DeepHit) for our task. The model discretizes event times to create a uniform grid of time intervals within the duration threshold in the sample set. The model assumes time is discrete from T_0 to T_{final}. Let $\widehat{h}(\widehat{M}) = \widehat{h}_0(\widehat{M}) + \cdots + \widehat{h}_{final}(\widehat{M})$ denote the estimated hazard (probability) mass function output by the model. The estimated survival function is shown in Eq 2.6.

$$\widehat{S}(T_j|\widehat{M}) = 1 - \sum_{k=1}^{j} \widehat{h}_k(\widehat{M}) \tag{2.6}$$

The model has m output neurons with linear activation, where m is a tunable hyperparameter, and uses two loss functions: discrete negative log-likelihood loss ($loss_L$) and ranking loss ($loss_{rank}$), as shown in Eq 2.7 and Eq 2.8, respectively.

$$loss_L = -\sum_{i=1}^{N} [Y_e^{(i)} \log(\widehat{h}_{Y_d^{(i)}}(\widehat{M}^{(i)})) + (1 - Y_e^{(i)}) \log(\widehat{S}(T_i|\widehat{M}^{(i)}))] \tag{2.7}$$

$$loss_{rank} = \sum_{j:Y_d^{(j)} > Y_d^{(i)}} Y_e^{(i)} \exp(\frac{\widehat{S}(T_i|\widehat{M}^{(i)}) - \widehat{S}(T_i|\widehat{M}^{(j)})}{\sigma}) \tag{2.8}$$

where σ is a tunable hyperparameter. The loss function of DeepHit is given in Eq 2.9, where α combines the two losses.

$$loss = \alpha loss_L + (1 - \alpha) loss_{rank} \tag{2.9}$$

PC-Hazard [41] is a multi-layer perceptron that parametrizes the continuous-time likelihood contribution function. The model considers hazard as piecewise constant. Time is partitioned from T_0 to T_{final}, and the hazard function $h(t)$ is a step function with non-negative constants $(\eta_1, \cdots, \eta_{final})$, as given in Eq 2.10.

$$h(t) = \eta_{\kappa(t)} \tag{2.10}$$

$\rho(t)$ is the interval fraction of $\kappa(t)$ at time t and is formulated in Eq 2.11.

$$\rho(t) = \frac{t - T_{\kappa(t)-1}}{\Delta_{T_{\kappa(t)}}} \tag{2.11}$$

The model has m output neurons with linear activation and uses a mean negative log-likelihood loss, as shown in Eq 2.12, where $\widehat{\eta}$ is the estimated vector of constants.

$$loss = -\frac{1}{N}\sum_{i=1}^{N}(Y_e^{(i)} \log \widehat{\eta}_{\kappa(t_i)} \widehat{M}^{(i)} - \widehat{\eta}_{\kappa(t_i)} \widehat{M}^{(i)} \rho(t) - \sum_{j=1}^{\kappa(t_i)-1} \widehat{\eta}_j \widehat{M}^{(i)}) \tag{2.12}$$

DeepSurv is a multi-layer perceptron that models a sample-specific risk function, $h(\widehat{M}^{(i)})$, as formulated in Section 2.3. The model has a single output neuron with linear activation and uses the negative log of Cox partial likelihood as the loss function, shown in Eq 2.13.

$$\log(L(\beta)) = -\sum_{i:Y_e^{(i)}=1} (\widehat{h}_\beta(\widehat{M}^{(i)}) - \log \sum_{j:Y_d^{(j)}>Y_d^{(i)}} e^{(\widehat{h}_\beta(\widehat{M}^{(j)}))}) \tag{2.13}$$

We utilize 2, 3, and 4-layer DeepHit, PC-Hazard, and DeepSurv models as baseline models. We use a single neuron (multiple for DeepHit and PC-Hazard) with linear activation to produce the predicted risk scores in the final layer. For our dataset of size N, the trained 3-layer DeepHit, PC-Hazard, and DeepSurv models are summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{2.14}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{2.15}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{2.16}$$

where X_{hid} denotes the output of the first hidden layer, $\widehat{h}^{(i)}$ signifies the predicted risk score, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ indicates the rectified linear unit. For the 2 and 4-layer DeepHit, PC-Hazard, and DeepSurv models, the hidden layer specified in Eq 2.15 is either omitted or replicated. For further details on the baseline neural networks, please refer to Section 2.4.3.

Additionally, we examine the DeepSurv model that utilizes the full spectrum as input. Details on this baseline model are provided in Section 2.4.4, while Section 2.4.2 covers the baseline models employed for the pathological classification task.

2.4.2 Pathological Classification

We consider random forest (RF) from [30] and fully-connected neural networks as baseline models for the pathological classification task.

For the pathological classification task, we use neural network-based models as baseline models. We train 2, 3, and 4-layer fully-connected neural networks with class-weighted binary cross-entropy loss. The trained 3-layer fully-connected networks on our dataset of size N can be described as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{2.17}$$
$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{2.18}$$
$$\widehat{Y_m}^{(1:N)} = Sigmoid(FC^{(1)}(X_{hid}^{(1:N)})) \tag{2.19}$$

where X_{hid} represents the output of the first hidden layer, $\widehat{Y_m}^{(1:N)}$ denotes the predicted tumor pathology, $FC^{(\cdot)}$ signifies a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ indicates the rectified linear unit. We also consider 2 and 4-layer fully-connected networks where the hidden layer specified in Eq 2.18 is either omitted or replicated, respectively.

2.4.3 PiDeeL and other models (DeepHit, PC-Hazard, and DeepSurv)

We evaluate 2, 3, and 4-layer DeepHit, PC-Hazard, and DeepSurv models as baseline models. For DeepSurv, we employ a single neuron in the final layer with linear activation to output the predicted risk scores of samples. For DeepHit and PC-Hazard models, the number of neurons in the final layer is a tunable parameter, referred to as *num_durations*. We empirically select 10 as *num_durations* from the set $\{5, 10, 20, 50, 100\}$.

The trained 3-layer DeepHit and PC-Hazard models on our dataset of size N can be summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{2.20}$$
$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{2.21}$$
$$\widehat{h}^{(1:N)} = FC^{(10)}(X_{hid}^{(1:N)}) \tag{2.22}$$

The trained 3-layer DeepSurv model on our dataset of size N can be summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{2.23}$$
$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{2.24}$$
$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{2.25}$$

where X_{hid} denotes the output of the first hidden layer, $\widehat{h}^{(i)}$ signifies the predicted risk score, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ indicates the rectified linear unit. In the case of 2 and 4-layer DeepHit, PC-Hazard, and DeepSurv models, the hidden layer specified in Eq 2.21 and 2.24 is either omitted or replicated.

2.4.4 Full spectrum model

We consider a fully-connected MLP model that utilizes the entire spectrum as input. It is important to note that this model excludes the automated metabolite quantification workflow outlined in Section 2.2. We train and evaluate 2, 3, and 4-layer versions of this model using the same methodology applied to PiDeeL. The architecture of this model is identical to that of the baseline fully-connected models, with the exception of the input vector shape, which is 16,314 instead of 37.

2.5 Pathway-Informed Deep Learning Model (PiDeeL)

In this section, we introduce PiDeeL, a pathway-informed multi-layer perceptron model designed to integrate biomarker metabolite concentrations at the metabolic pathway level. We utilize the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to parse the metabolite-to-pathway mapping and construct a 37×138 matrix, referred to as PI_matrix, where the entry in the i^{th} row and j^{th} column is 1 if the metabolite i is present in pathway j, and 0 otherwise.

Analogous to the baseline fully-connected neural network models used for survival analysis and pathological classification tasks, we consider 2, 3, and 4-layer PiDeeLs.

2.5.1 Model Architecture

Given a dataset of N samples, the architecture of the 3-layer PiDeeL for the survival analysis can be summarized as follows:

$$PL^{(138)} = \left[\ FC^{(138)} weights\ \right] \left[PI\_matrix \right] \tag{2.26}$$

$$PA^{(1:N)} = ReLU(PL^{(138)}(\widehat{M}^{(1:N)}) + b^{138}) \tag{2.27}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(PA^{(1:N)})) \tag{2.28}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{2.29}$$

where $PL^{(138)}$ and b^{138} denote the weights and biases of the pathway-informed layer, $PA^{(1:N)}$ represents the activations of the first hidden layer (hereafter referred to as pathway-activation), X_{hid} denotes the output of the hidden layers, $\widehat{h}^{(i)}$ signifies the predicted risk score, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ indicates the rectified linear unit. We incorporate the metabolic pathway information into PiDeeL by defining the first layer of the model using PI_matrix. This architectural choice results in a sparse first layer in PiDeeL, as connections in this layer are only allowed between input metabolites and their corresponding pathways. To train PiDeeL for the survival analysis task, we utilize the negative log of Cox Partial Likelihood as the loss function, akin to the DeepSurv models described in Section 2.4.1. We also train PiDeeL with DeepHit and PC-Hazard losses. Our findings indicate that the DeepSurv loss function yields the best performance for PiDeeL. For further details on these experiments, please refer to Section 2.5.4 and Supplementary Figure B.8.

The architecture of PiDeeL for survival analysis and pathological classification tasks varies only in the output layer (i.e., Eq 2.29). To adapt PiDeeL for pathological classification, we employ the sigmoid function as the output activation function by substituting Eq 2.29 with Eq 2.30. Additionally, we use class-weighted binary cross-entropy as the loss function.

$$\widehat{Y_m}^{(1:N)} = Sigmoid(FC^{(1)}(X_{hid}^{(1:N)})) \tag{2.30}$$

2.5.2 Interpretability and Importance Analyses

We perform SHapley Additive exPlanations (SHAP) analysis on our feature vector ($\widehat{M}^{(1:N)}$) and pathway activations ($PA^{(1:N)}$) to assess the importance of each metabolite and pathway on PiDeeL's predictions. We adopt the same 5-fold cross-validation setup (repeated 3 times) and hyperparameter selection described in Section 3.1 of this thesis. During iteration i, SHAP values are calculated on the test fold (fold $(i+1) \mod 5$) using the trained PiDeeL from training fold i. The hyperparameters are consistent across all trained models (5-fold repeated 3 times). In each iteration (i.e., 15 times), we concatenate the SHAP values calculated in that iteration with the previous SHAP values to form a 37×15 matrix. We then average each metabolite's SHAP value vector and sort the mean SHAP values of the 37 metabolites in increasing order. For the calculated SHAP values for each metabolite in the survival analysis and pathological classification tasks, please refer to Supplementary Figure B.4. The three metabolites with the most significant impact on PiDeeL's survival analysis predictions are discussed in Section 3.4 of this thesis.

To assess the interpretability of the pathway profiles, we use the same 5-fold cross-validation setup. This time, we examine the output of the first hidden layer (i.e., pathway-activation vector $PA^{(1:N)}$). We measure the impact of each pathway by calculating the SHAP values of vector $PA^{(1:N)}$ on PiDeeL's predictions. We concatenate and sort the SHAP values of $PA^{(1:N)}$ calculated for each fold. The three most significant pathways are discussed in Section 3.4 of this thesis.

2.5.3 Ablation Studies for Pathway-Informed Architecture

To ascertain that the enhancements achieved by PiDeeL over the baseline models are attributable to the pathway-informed architecture, we conduct a series of ablation tests as detailed below:

- **Test a)** We train fully-connected models with 138 neurons in the first hidden layer to verify if the number of neurons in the first layer (138) of PiDeeL contributes to the performance.
- **Test b)** We train PiDeeLs using a randomly connected PI_matrix instead of the PI_matrix constructed from the KEGG database to determine if the pathway-informed connections offer an advantage over an equivalent number of randomly-connected edges.
- **Test c)** We train PiDeeLs using a shuffled PI_matrix, which is akin to a randomly connected PI_matrix but maintains the in-degrees of pathways.
- **Test d)** We train the DeepSurv model with varying dropout rates $(0.5, 0.6, 0.7, 0.8, 0.9)$ in the first hidden layer to evaluate if the pathway-informed layer provides a benefit over this commonly used regularization technique.

- **Test e)** We train and test DeepSurv and PiDeeL on sample sets of sizes 50, 100, and 200 to assess the models' adaptability to different dataset sizes.

Here, we provide the specifics of the ablation tests performed to determine the contribution of the pathway-informed architecture to the performance of the PiDeeL.

In test a, we validate the impact of the number of neurons in the first layer on performance. Notably, PiDeeL's first hidden layer has 138 neurons, whereas the trained fully-connected networks (baseline models) have 64. We train 2, 3, and 4-layer DeepSurv models using the same 5-fold cross-validation setup, repeating it three times. The model architecture can be summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(138)}(\widehat{M}^{(1:N)})) \tag{2.31}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{2.32}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{2.33}$$

We train 2, 3, and 4-layer PiDeeL with the same 5-fold cross-validation setup, and repeat it 30 times (i.e., 150 iterations).

Recall that $PL^{(138)}$ refers to the weights of the pathway-informed layer of PiDeeL. To construct $PL^{(138)}$, we use Eq 2.34, where we multiply the $FC^{(138)}$ weights by the PI_matrix built from the KEGG database for pathways. In test b, we use a randomly connected PI_matrix to verify the contribution of the PI_matrix to PiDeeL's performance. The number of connections in the first hidden layer (468) is preserved in the random PI_matrix. We train 2, 3, and 4-layer PiDeeLs with the same 5-fold cross-validation setup, repeating it 30 times (i.e., 150 iterations).

$$PL^{(138)} = \left[\ FC^{(138)}weights\ \right]\left[PI\_matrix\right] \tag{2.34}$$

In test c, we construct the PI_matrix similarly to test b. However, we shuffle each metabolite-to-pathway connection vector instead of randomly connecting them. The PI_matrix is a 37×138 matrix; for metabolite j, $PI\_matrix.j$ yields a vector of shape 1×138. We shuffle these 37 vectors and train 2, 3, and 4-layer PiDeeLs using the same 5-fold cross-validation setup, repeating it 30 times (i.e., 150 iterations).

In test d, we verify the effect of the first hidden layer's sparsity on PiDeeL's performance. In PiDeeL, the total connections between the input layer ($\widehat{M}^{(1:N)}$) and the first hidden layer $PL^{(1:N)}$ are 468, whereas a fully-connected network with 37 input features and 138 neurons in the first hidden layer has 5106 connections. We train 2, 3, and 4-layer DeepSurv models with dropout in the first hidden layer using the same 5-fold cross-validation setup, repeating it three times, with different dropout rates from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$.

Finally, in test e, we train and test DeepSurv and PiDeeL on sample sets of sizes 50, 100, and 200. We randomly undersample our dataset (n=384) to sizes 200, 100, and 50. We then train and test DeepSurv and PiDeeL on these sample sets 10 times with a 5-fold cross-validation, repeating it three times. This results in a total of 150 iterations.

2.5.4 PiDeeL with different losses

We train PiDeeL using losses formulated by the DeepHit, PC-Hazard, and DeepSurv models in Eq 2.9, 2.12, and 2.13 of this thesis. Given our dataset of N samples, the architecture of the trained 4-layer PiDeeL for survival analysis can be summarized as follows:

$$PL^{(138)} = \left[\ FC^{(138)}weights\ \right]\left[PI\_matrix\right] \tag{2.35}$$

$$PA^{(1:N)} = ReLU(PL^{(138)}(\widehat{M}^{(1:N)}) + b^{138}) \tag{2.36}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(PA^{(1:N)})) \tag{2.37}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(PA^{(1:N)})) \tag{2.38}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{2.39}$$

We train 4-layer PiDeeLs using the same 5-fold cross-validation setup, repeating it 30 times (i.e., 150 iterations). In Supplementary Figure B.8, we present the performance comparison of PiDeeLs using DeepHit, PC-Hazard (cyan boxplots), and DeepSurv (navy blue boxplot) losses. PiDeeL trained with the DeepSurv loss outperformed models trained with other losses in the median c-index metric. Specifically, the use of DeepSurv loss improved the median c-index achieved by DeepHit and PC-Hazard losses by 9.75% and 4.04%, respectively. Additionally, the variance decreased by 22.1% and 39.2% for DeepHit and PC-Hazard losses, respectively. Therefore, we proceed with using the DeepSurv loss, as formulated in Eq 2.13 of this thesis, in our model.

2.5.5 Multitask learning

We explore a multi-task strategy for PiDeeL for survival analysis and pathological classification tasks. The multi-task PiDeeL learns a function $s : \mathbb{R}^{37} \rightarrow [-\infty, \infty] \times \{0, 1\}$. Note that we use the same feature vector (metabolite profiles) as input. The multi-task PiDeeL architecture can be summarized as follows:

$$PL^{(138)} = \left[\; FC^{(138)} weights \; \right] \left[PI\_matrix \right] \tag{2.40}$$

$$PA^{(1:N)} = ReLU(PL^{(138)}(\widehat{M}^{(1:N)}) + b^{138}) \tag{2.41}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(PA^{(1:N)})) \tag{2.42}$$

$$\widehat{h}^{(1:N)}, \widehat{Y_m}^{(1:N)} = FC^{(2)}(X_{hid}^{(1:N)}) \tag{2.43}$$

We train 2, 3, and 4-layer multi-task PiDeeLs using the negative log of Cox Partial Likelihood and class-weighted binary cross-entropy losses. Hyperparameter configurations for multi-task PiDeeL were selected based on optimizing a linear combination of survival and pathological classification surrogates. Specifically, a grid search based on validation performance determined a 0.125-fold weight for Cox-PH loss and a 0.875-fold weight for binary cross-entropy loss.

Refer to Supplementary Figure B.5 for a visualization of the convergence of the loss functions during training for a 3-layer PiDeeL using a random seed. The left plot shows the absolute values of the Cox-PH and binary cross-entropy losses, while the right plot shows the scaled values of Cox-PH and binary cross-entropy losses.

Chapter 3

Results

3.1 Experimental Setup

In this section, we elaborate on the experimental setup employed to evaluate PiDeeL as well as the baseline models outlined in Section 2.4 within the context of survival analysis. The performance of the survival analysis models is assessed using the time-dependent concordance index (c-index) as described in [51]. The traditional concordance index [52] is computed using individually predicted patient risk scores, whereas the time-dependent version calculates individual risk scores using the predicted risk function and the baseline hazard function derived from the sample set [51, 53].

For each task and model type, we utilize a 5-fold cross-validation procedure, repeating it three times with different random seeds (i.e., 15 iterations). Prior to each iteration, we shuffle and partition our dataset into training, validation, and test sets, ensuring no overlap of samples or patients across folds. The results presented here pertain to the models' predictive capabilities on the test split. During iteration i of the cross-validation process, the Cox-PH, CWGB, and RSF models were tested on fold i, validated on fold $(i+1) \mod 5$, and trained on the remaining three folds. We re-train the selected configuration on the combined training and validation sets to enhance the baseline models. However, for all deep learning models, re-training steps are omitted due to the presence of early-stopping mechanisms.

3.1.1 Hyperparameter Space

In this section, we provide the lists of hyperparameter spaces for the baseline models for both survival analysis and pathological classification tasks.

For the CoxPH model, we employ a ridge-regularized variant to prevent overfitting. Hyperparameter optimization is conducted on the validation split, with the search space restricted to: (1) regularization penalty coefficient $\in \{0, 0.1, \ldots, 0.9, 1\}$; (2) the heuristic for handling event ties, set to either *efron* [54] or *breslow* [55]; and (3) the maximum number of iterations $\in \{100, 200, \ldots, 500\}$.

For the CWGB model, hyperparameter selection on the validation split involves the following search space: the number of boosting stages $\in \{20, 50, 100, 150\}$; the loss function to be optimized $\in \{\text{"Cox-PH"}, \text{"Squared"}\}$; the percentage of base learners to be dropped $\in \{0.0, 0.1, 0.2\}$; and the learning rate $\in \{0.2, 0.1, 0.05\}$.

For the RSF model, similar to the Cox-PH and CWGB models, hyperparameter selection on the validation split uses the following search space: the number of trees in the forest $\in \{20, 50, 100, 150, 300, 400\}$; the maximum depth limit for each tree $\in \{\text{no limit}, 10, 15, 25, 30\}$; the minimum number of samples required to split an internal node $\in \{2, 5, 10, 15\}$; the minimum number of samples required in leaf nodes $\in \{2, 5, 10, 15\}$; and the maximum number of features considered for the best split $\in \{\sqrt{\text{number of metabolites} = 37}, \log_2(37)\}$.

The DeepSurv model's hyperparameter search space is restricted to: (1) batch size for training $\in \{16, 32, 64, 128, \text{full-batch}\}$; and (2) the activation function between layers $\in \{\text{ReLU}, \text{Sigmoid}\}$.

For the pathological classification task, the RF model serves as the baseline. We apply grid search for hyperparameter selection and choose the setting with the best performance (AUC-ROC) on the validation set. The hyperparameter search space for the RF model includes: the number of trees in the forest $\in \{50, 150, 300, 400\}$; the maximum depth limit for each tree $\in \{10, 15, 25, 30\}$; the minimum number of samples required to split an internal node $\in \{5, 10, 15\}$; the minimum number of samples required in leaf nodes $\in \{2, 10, 20\}$; and the criterion for measuring the quality of a split: *Gini-index* and *entropy*.

The other baseline model for the pathological classification task is the fully-connected neural network. The hyperparameter search space is similar to that of DeepSurv: (1) batch size for training $\in \{16, 32, 64, 128, \text{full-batch}\}$; and (2) the activation function between layers $\in \{\text{ReLU}, \text{Sigmoid}\}$.

3.1.2 Hyperparameter Selection

Regardless of the task, we use grid search on the validation split to determine the best-performing hyperparameter values for each model and iteration. Training these models involves a two-step process. During iteration i of the cross-validation setup, models are tested on fold i, validated on fold $(i+1) \mod 5$, and trained on the remaining three folds. After selecting the best-performing hyperparameters based on the c-index from the validation set, we re-train the selected configuration on the combined training and validation splits to obtain stronger baseline models.

Hyperparameters of deep learning-based models (DeepHit, PC-Hazard, and DeepSurv) or proposed (PiDeeL) for survival analysis were tuned based on the loss calculated on validation split predictions. As described in Section 2.4.1 of this thesis, the losses (Eq 2.9, 2.12, and 2.13) adopted for deep learning-based survival analysis are differentiable surrogates for the time-dependent c-index metric [51]. Similarly, hyperparameters of deep learning models for the pathological classification task were tuned based on binary cross-entropy loss, a differentiable proxy for AUC-ROC and AUC-PR, calculated on the validation split. For all deep learning models discussed above, cross-validation folds were distributed to training, validation, and test splits similar to the Cox-PH, CWGB, and RSF models. However, re-training with the union of training and test splits was not performed due to early-stopping dynamics.

To define the hyperparameter space, we sampled five equidistant learning rates from the $[10^{-3}, 10^{-4}]$ interval, conducted experiments, and selected 10^{-4} as the learning rate based on the highest observed performance.

The hyperparameter search space for PiDeeL for both tasks was as follows: (1) batch size for training $\in \{16, 32, 64, 128, \text{full-batch}\}$; and (2) the activation function between layers $\in \{\text{ReLU}, \text{Sigmoid}\}$. After comprehensive experimentation, we identified the optimal settings for our model, selecting batch sizes of 64 and 32 for survival analysis and pathological classification tasks, respectively. Additionally, the ReLU activation function yielded the highest empirical performance, and we proceeded with this choice.

Moreover, we trained models for up to $5,000$ epochs (selected from $\{1000, 2000, \ldots, 6000\}$) and applied early-stopping when the validation loss increased for 150 consecutive epochs (selected from 50, 100, 150, 200). We used the Adam optimizer [56] with a weight decay of 0.002, based on the highest observed performance.

The time discretization parameter *num_durations* for DeepHit and m for PC-Hazard were empirically selected as 10 from $\{5, 10, 20, 50, 100\}$.

3.2 Survival Analysis

Refer to Figure 3.1 for the survival prediction performances of (i) Cox-PH, CWGB, and RSF models (gray boxplots); (ii) DeepHit, PC-Hazard, and DeepSurv models (cyan boxplots); and PiDeeL (navy blue boxplot), with respect to the c-index. Note that Figure 3.1 displays the performance of 4-layer neural network models (DeepHit, PC-Hazard, DeepSurv, and PiDeeL). See Section 3.2.2

and Supplementary Figures B.9, B.10, and B.11 for the performance comparison of 2, 3, and 4-layered DeepHit, PC-Hazard, and DeepSurv models against PiDeeL.

We observe that PiDeeL surpasses other methods, achieving a median c-index of 68.7%. RSF proves to be a particularly robust baseline with a median c-index of 67.8%. Among the other neural network-based models, DeepSurv attains the highest median c-index at 64.3%.

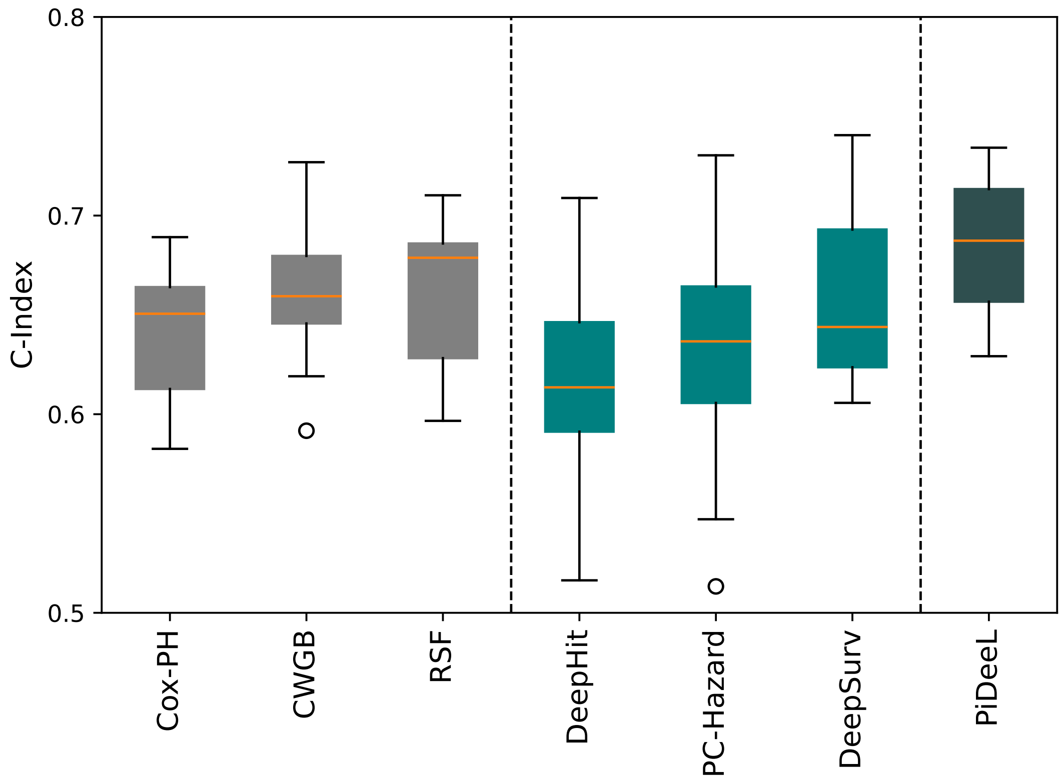

Figure 3.1: Comparison of baseline models and PiDeeL on survival prediction in terms of c-index. This evaluation is conducted using a 5-fold cross-validation repeated 3 times, resulting in a total of 15 iterations.

3.2.1 Full spectrum model

To demonstrate the advantages of the automated metabolite quantification pipeline, we compare PiDeeL with a version of DeepSurv that utilizes the full

HRMAS NMR spectrum instead of quantified metabolites. Supplementary Figure B.6 illustrates the performance comparison of these models with PiDeeL in terms of the c-index. PiDeeL outperforms this baseline, achieving higher median c-index values by 7.9%, 8.3%, and 5.0% for 2, 3, and 4-layered models, respectively. Additionally, PiDeeL reduces the variance by an average of 21.0%. It should be noted that using PiDeeL directly with the full NMR spectrum is infeasible due to the complexity of integrating pathway information into the model.

3.2.2 PiDeeL and other models (DeepHit, PC-Hazard, and DeepSurv)

Refer to Supplementary Figures B.9, B.10, and B.11 for layer-wise comparisons between PiDeeL and DeepHit, PC-Hazard, and DeepSurv. We observe a slight performance increase as the depth of PiDeeL increases from 2 to 4 layers (up to a 0.4% improvement in median c-index). Conversely, the performance of PC-Hazard and DeepSurv models decreases with increasing model depth (up to a 4.4% and 1.5% decrease in median c-index for PC-Hazard and DeepSurv, respectively). When comparing neural network-based models (DeepHit, PC-Hazard, and DeepSurv) and PiDeeL with the same number of layers, the performance gap increases with the number of layers. This highlights the benefit of incorporating biological information into modeling choices. For 2, 3, and 4-layered networks, PiDeeL improves upon DeepHit by 15.8%, 13.1%, and 12.0% in median c-index, respectively. Similarly, PiDeeL outperforms PC-Hazard by 2.6%, 4.0%, and 7.9% and DeepSurv by 2.5%, 3.4%, and 4.3% in median c-index for the same model depths. These results underscore the robustness of sparse architectures in small data regimes for survival analysis tasks.

3.3 Pathological Classification

We evaluate the performance of pathological classification models using the Area Under the ROC Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR) metrics.

Refer to Supplementary Figure B.2 for a visual depiction of pathological classification performance achieved by RF (gray boxplot), fully-connected networks (cyan boxplots), and PiDeeL (navy blue boxplots) with 2, 3, or 4 layers. Similar to the survival analysis results presented in Section 3.2, PiDeeL consistently outperforms the fully-connected baseline, irrespective of architecture depth or the performance metrics used. For AUC-ROC, PiDeeL shows performance improvements of 1.7%, 0.7%, and 2.7% compared to the fully-connected baseline with 2, 3, and 4 layers, respectively. A similar trend is observed for the AUC-PR metric, with improvements of 0.6%, 0.4%, and 0.8% for the same model depths. However, PiDeeL does not show a significant quantitative improvement over the RF model for the pathological classification task. While the performance of fully-connected networks deteriorates with increasing layers, PiDeeL shows a slight increase in median AUC-ROC with model depth.

3.4 Interpretability and Importance Analyses

Combining PiDeeL with downstream feature importance analysis methods, such as SHAP [57], allows for joint analysis of metabolite and pathway importance for model output. In addition to the quantitative analyses, we perform feature importance analyses on the 4-layer PiDeeL to identify metabolites and pathways implicated in gliomas, their progression, and associated patient survival. We calculate SHAP values for input features (i.e., in silico predicted metabolite concentrations - $\widehat{M}^{(1:N)}$) and pathway-informed layer activations ($PA^{(1:N)}$).

Refer to Supplementary Figure B.4 for PiDeeL's metabolite importance plots for survival prediction and tumor pathological classification tasks. We highlight the top 3 metabolites with the highest mean SHAP values for survival analysis: *glutamate*, *glutamine*, and *alanine*. We also identify the top 3 pathways with the highest SHAP values: *Mineral absorption*, *Alanine, aspartate and glutamate metabolism*, and *mTOR signaling pathway*. Figure 3.2 shows the relative importance of metabolites and pathways for PiDeeL's outcomes, with node sizes representing importance and flow thickness indicating metabolite contributions to pathway importance.

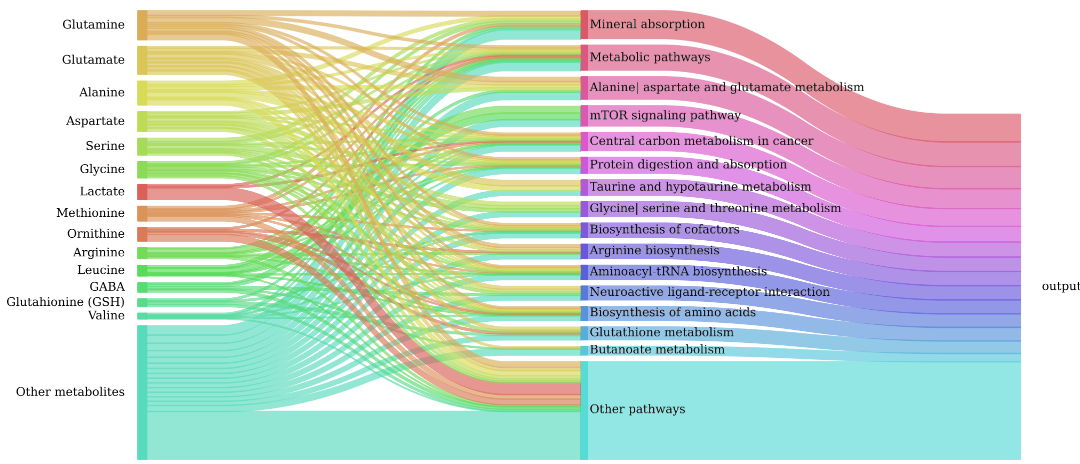

Figure 3.2: Sankey diagram employed to visualize the relative importance of metabolites and pathways for PiDeeL. Nodes on the left represent metabolite level inputs, nodes in the middle correspond to pathway layer activations, and the node on the right denotes the model output. Larger nodes indicate greater importance for the model, while smaller nodes indicate lesser importance. The flows in the diagram illustrate the contributions of metabolites to pathways, where a larger flow signifies a greater contribution to the subsequent layer node. For instance, the contribution of Glutamine to the Mineral absorption pathway is more significant than the contribution of Methionine to the Protein digestion and absorption pathway.

3.4.1 Biological Motivation for Interpretability and Importance Analyses

Through a case-control study, alanine has been reported as a potential biomarker for malignant gliomas [58]. Similarly, glutamate plays a central role in malignant gliomas through multiple mechanisms [59]. Given the poor prognosis of malignant gliomas, the high ranking of alanine and glutamate is expected. Glutamine, the primary precursor of glutamate, is critical for brain function [60], regulating pathways like energy production and redox homeostasis in brain cancer cells [60]. Glutamine has also been identified as a biomarker for glioma progression and treatment response [61], making its high ranking unsurprising.

We provide a biological rationale for the connection between the top three pathways and gliomas. Minerals and their absorption are essential for sustaining life[1]. Calcium has been identified as a dietary component with potential for glioma prevention by influencing apoptosis and DNA repair [62]. Therefore, the importance of the Mineral absorption pathway is biologically justified. Additionally, two of the three metabolites involved in the Alanine, aspartate, and glutamate metabolism pathway are among the top three important metabolites, as discussed previously. Finally, mTOR plays a crucial role in integrating signal transduction and metabolic pathways in glioblastoma [63]. Since approximately 46% of the cohort analyzed in this study were diagnosed with glioblastoma, the prominence of the mTOR signaling pathway is justified. This reinforces the predictive capabilities of PiDeeL and highlights its interpretability at both the metabolite and pathway levels.

We also explore the relationship between the top metabolites and pathways identified. Glutamine has been reported to play a crucial role in the mTOR signaling pathway, acting as a key nitrogen donor and activating mTORC1 [64]. mTORC1 regulates cell growth, metabolism, and autophagy, thereby influencing proliferation. Glutamine's activation effect on the mTORC1 complex controls

[1] KEGG PATHWAY: Mineral absorption - Homo sapiens (human). (n.d.). https://www.genome.jp/pathway/hsa04978

protein synthesis, lipid synthesis, and autophagy, which are often altered in cancer metabolism [65]. Glutamate forms complexes with minerals through chelation, binding to minerals like calcium, sodium, and potassium. These complexes enhance the solubility and availability of these minerals, making them more accessible to brain tissue [66]. Alanine facilitates glutamate intake in brain tissue through aminotransferase, which in turn regulates mineral absorption [67]. The alanine, aspartate, and glutamate metabolism pathway is responsible for the synthesis and interconversion of amino acids, playing a crucial role in the synthesis of arginine, asparagine, aspartate, alanine, glutamate, glutamine, and proline[2]. This pathway is also highly interconnected with other metabolic pathways, with glutamate donating amino groups for the synthesis of non-essential amino acids[3] [68]. Glutamate can be converted into alpha-ketoglutarate through the action of the enzyme glutamate dehydrogenase. This reaction generates reducing equivalents (NADH) and participates in the tricarboxylic acid (TCA) cycle, a key pathway for energy production[4]. Additionally, glutamate is a precursor for glutathione, a critical antioxidant molecule that helps protect cells from oxidative damage and maintain redox balance [69].

3.5 Ablation Studies for Pathway-Informed Architecture

In this section, we present the results of ablation studies discussed in Section 2.5.3. Figure 3.3 demonstrates the outcome of the respective ablation tests (excluding dropout).

Figure 3.3 is divided into three panels to show the results for 2, 3, and 4-layer models. In each panel, the boxplots from left to right illustrate the performance

[2] KEGG PATHWAY: Alanine, aspartate and glutamate metabolism - Reference pathway. (n.d.). https://www.genome.jp/pathway/map00250

[3] KEGG PATHWAY: Nitrogen metabolism - Reference pathway. (n.d.). https://www.genome.jp/pathway/map00910

[4] KEGG PATHWAY: Citrate cycle (TCA cycle) - Reference pathway. (n.d.). https://www.genome.jp/pathway/map00020

of three ablation tests (cyan boxplots): DeepSurv model with 138 neurons in the first layer, PiDeeL with a randomly connected PI_matrix, PiDeeL with a randomly shuffled PI_matrix, and PiDeeL (navy blue boxplots).

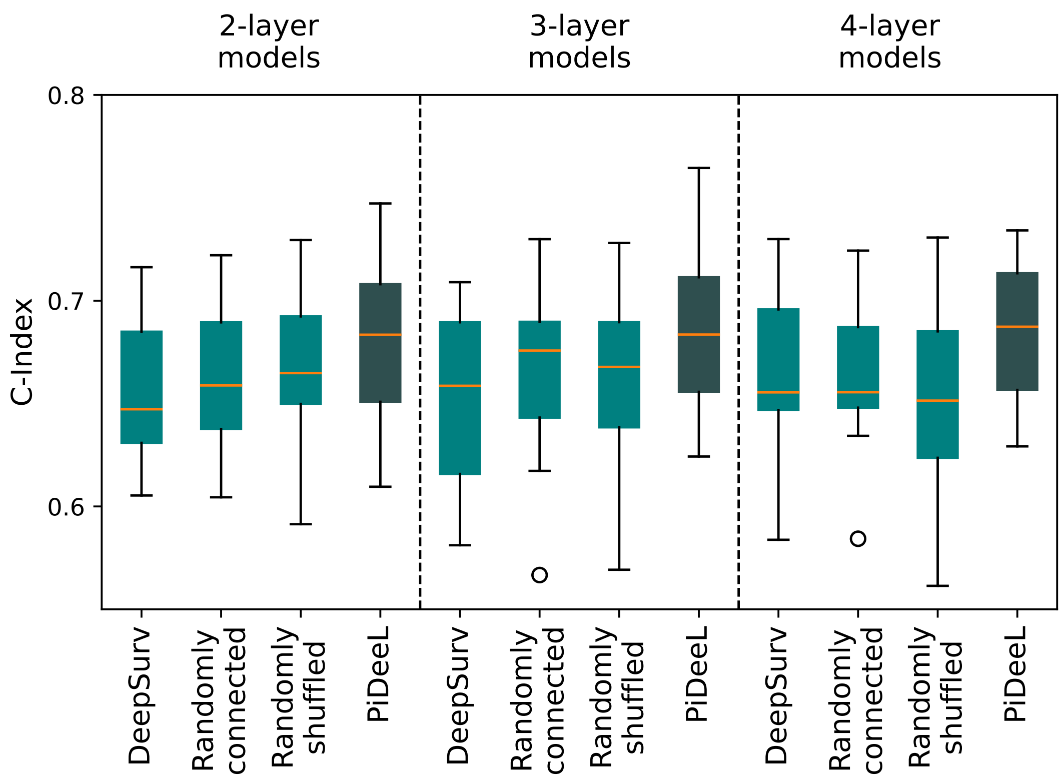

Figure 3.3: Performance comparison of PiDeeL on the survival analysis task against DeepSurv with 138 neurons in the first hidden layer, PiDeeL with a randomly connected PI_matrix, and PiDeeL with a shuffled PI_matrix, with respect to c-index.

We observe that PiDeeL improves the median c-index over DeepSurv with 138 neurons in the first layer by 3.6%, 2.6%, and 3.6% for 2, 3, and 4-layer networks, respectively. Additionally, using PiDeeL over DeepSurv reduces the variance by an average of 29.7%. It is noteworthy that the original DeepSurv model has 64 neurons in the first layer, indicating that the performance improvement over DeepSurv is attributable to pathway information rather than the larger number of neurons in the first layer.

We compare the performance of PiDeeL with the PI_matrix against PiDeeL with a random PI_matrix. In this setup, we randomly connect our metabolite nodes (input vectors) and pathway activation layer neurons. We constructed randomly connected 37×138 matrices, preserving the total number of connections in the original PI_matrix (468). We observe that PiDeeL improves the median c-index over PiDeeL with a random PI_matrix by 3.7%, 1.2%, and 4.9% for 2, 3, and 4-layer networks, respectively.

We also compare the performance of PiDeeL with the PI_matrix against PiDeeL with a shuffled PI_matrix. Instead of randomly connecting the PI_matrix, we shuffle the rows to preserve each metabolite-to-pathway profile. We observe a similar trend, with the PI_matrix improving the performance over the shuffled PI_matrix in terms of median c-index by 2.8%, 2.4%, and 4.9%. Additionally, using the PI_matrix decreases the variance by an average of 4.6%.

Figure 3.4 shows the survival prediction performance comparison of DeepSurv with various dropout rates (d) and PiDeeL. We observe that PiDeeL substantially outperforms DeepSurv with dropout, regardless of the dropout rate. Additionally, DeepSurv's performance deteriorates as the dropout rate increases. The 2-layer PiDeeL achieves a median c-index of 68.3%, whereas the DeepSurv model attains median c-index scores of 58.1%, 58.6%, 51.6%, 51.1%, and 47.0% for dropout rates of 0.5, 0.6, 0.7, 0.8, and 0.9, respectively. A similar trend is observed with increasing layers for both models. This indicates that the sparsity provided by dropout does not enhance the performance of the fully-connected model for survival prediction, whereas PiDeeL's pathway-informed sparsity with the PI_matrix improves performance.

In Supplementary Figure B.7, we present the performance benchmark of PiDeeL and DeepSurv on different sample set sizes. For a sample set of size 200, PiDeeL improves the median c-index over DeepSurv by 5.4%, 7.6%, and 5.7% for 2, 3, and 4-layer networks, respectively. For a sample set of size 100, PiDeeL outperforms DeepSurv by 7.6%, 5.8%, and 11.8% in terms of median c-index. Finally, for a sample set of size 50, the performance gap between PiDeeL and DeepSurv increases significantly. PiDeeL improves the median c-index over DeepSurv by

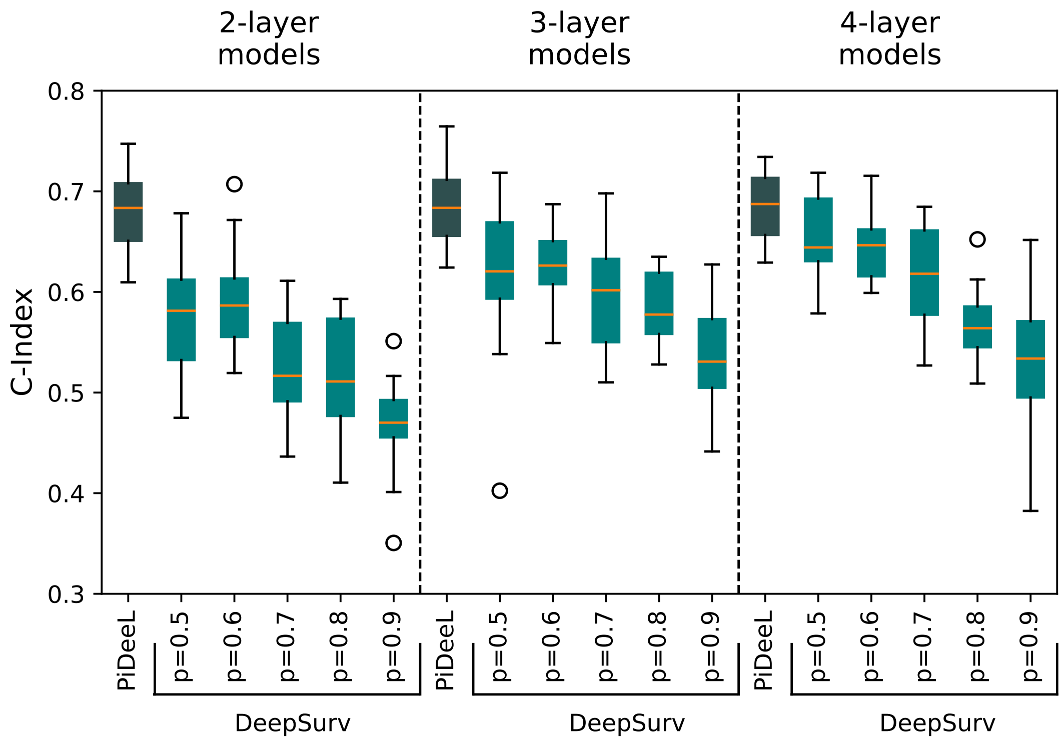

Figure 3.4: Performance comparison of PiDeeL on the survival analysis task against DeepSurv with varying dropout rates with respect to c-index. This evaluation is based on a 5-fold cross-validation repeated 3 times, resulting in a total of 15 iterations.

4.9%, 13.5%, and 18.6% for 2, 3, and 4-layer networks, respectively. These results indicate that the pathway-informed architecture improves performance and mitigates overfitting by reducing model complexity.

3.6 Validation on an independent dataset

We acquired an independent HRMAS NMR glioma dataset from Firdous et al. [58]. The raw FID files underwent preprocessing consistent with the methods applied to our dataset. The outcome of this preprocessing pipeline is a spectrum vector of length 16.314. Subsequently, we normalized the signal intensity relative to the mean intensity of our samples and calibrated the spectrum by shifting it leftward by 1515 ppm to ensure alignment with our data. This dataset encompasses 42 individuals, of which 26 are glioma samples. Initially, we estimated the metabolite quantities for these 26 samples to derive the feature vectors.

The labels in the dataset from Firdous et al. include tumor grade (i.e., I-IV) and tumor subtype (e.g., AST, GBM). Our dataset similarly comprises these two labels, specifically tumor grade and tumor subtype. Given the absence of survival information (i.e., event indicator Y_e and/or time-to-event Y_d) in the Firdous et al. dataset, we simulated $Y_e^{(i)}$ and $Y_d^{(i)}$ for each glioma sample (i), taking into account their grades and subtypes.

For each subtype and grade (e.g., ASTIII, OASTII), we assumed that the survival labels in the Firdous et al. dataset are uniformly distributed within the range of the minimum and maximum time-to-event values in our dataset corresponding to the same subtype and grade. Consequently, for each glioma sample in the Firdous et al. dataset, we randomly assigned a time-to-event label from the appropriate distribution. Similarly, we assumed that the event indicator in the Firdous et al. dataset follows a Bernoulli distribution for each subtype and grade, mirroring the proportion of deceased and censored samples in our dataset. We then randomly drew from the respective distribution to simulate an event indicator label. This simulation was repeated 100 times with 3 different random seeds (i.e., 300 total iterations).

We evaluated the survival prediction performance of various neural network-based models and PiDeeL using this simulated data. The performance benchmark of these models and PiDeeL is presented in Figure 3.5. The median c-indices of baseline models are below 50%, with DeepHit, PC-Hazard, and DeepSurv achieving median c-indices of 47.4%, 47.7%, and 46.9%, respectively. PiDeeL surpasses DeepHit, PC-Hazard, and DeepSurv models, attaining a median c-index of 52.1%. Additionally, PiDeeL reduces variance by an average of 15.2%.

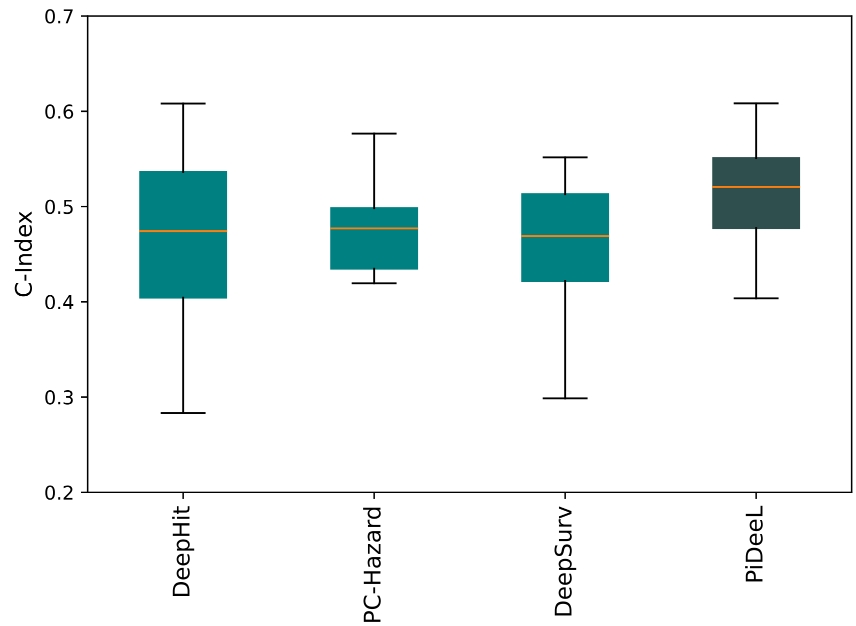

Figure 3.5: Performance comparison of PiDeeL and other models on an independent dataset with respect to c-index. This evaluation is based on 100 iterations with 3 different random seeds, totaling 300 iterations.

3.7 Multitask Learning

Please refer to Supplementary Figure B.3 for the comparative performance analysis of single and multi-task learning approaches in survival analysis and pathological classification tasks using PiDeeL. Across all evaluated model depths, the multi-task modeling strategy demonstrates at least a slight advantage in survival analysis performance. The 4-layer multi-task PiDeeL achieves the highest median c-index of 69.5%. In the pathological classification task, the 2-layer multi-task PiDeeL outperforms all single-task and multi-task models with a median AUC-ROC of 91.9% and a median AUC-PR of 97.5%. However, the exploration of multi-task learning in PiDeeL yielded marginal performance improvements (0.8% in median c-index and 1.8% in median AUC-ROC and 0.4% in median AUC-PR in 4-layer and 2-layer multi-task PiDeeLs); thus, we opted to proceed with the deployment of two task-specific single-task PiDeeLs.

Chapter 4

Discussion

The ^{1}H HRMAS NMR spectrum has proven to be a reliable resource for differentiating malignant from healthy samples obtained from the excision cavity during surgery. This technique is particularly advantageous as it is non-destructive and capable of analyzing small samples of raw tissue specimens [27]. Machine learning techniques that learn from the NMR signal to distinguish between healthy and tumor tissues, as well as benign and aggressive tumors, have been pivotal. However, their performance has been constrained by the small size of training sets [29, 30], which limits the application of complex architectures such as deep neural networks that learn a hierarchical composition of complex features. In this thesis, we successfully employed such a hierarchically complex model to distinguish benign from aggressive tumors with comparable performance. Despite utilizing a deep model, we reduced the number of trainable parameters by integrating prior biological information derived from metabolic networks into the model architecture.

Tumor pathology provides critical information for the surgeon, who might opt to resect more tumor tissue, risking the patient's well-being if the tumor is likely aggressive, or might risk leaving residual tumor tissue if the tumor is likely benign. Nevertheless, pathology alone does not definitively determine prognosis. Therefore, it is also crucial to estimate patient survival, which serves as a proxy for prognosis, to inform the surgeon. To this end, we trained a pathway-informed model to predict patient survival and demonstrated high performance. An illustrative example of the utility of combining survival prediction with pathological classification is as follows: Our model accurately predicted long survival for a patient with aggressive glioma who survived 5,656 days post-surgery. In this scenario, the surgeon could adopt a more conservative approach compared to relying solely on a prediction of aggressive tumor pathology. Conversely, our model correctly predicted low survival for a patient with a benign tumor who survived only 43 days post-surgery.

We also explored a multi-task learning architecture that combines two models to predict both labels simultaneously. Our results indicated that this approach only marginally increased performance for both tasks. Consequently, we decided to proceed with the utilization of two task-specific single-task PiDeeLs.

One limitation of our study is the dependency on the automated metabolite quantification model, which has demonstrated promising performance and generalizability. However, this component comprises individual multi-layer perceptrons for 37 metabolites for which a training dataset exists. We believe our model would benefit from a more comprehensive spectrum of metabolites to predict patient survival, which is a complex outcome to approximate. The human metabolome consists of approximately 3,000 metabolites, and we are currently limited to utilizing only 1% of this information source.

Bibliography

[1] A. Omuro and L. M. DeAngelis, "Glioblastoma and other malignant gliomas: a clinical review," *Jama*, vol. 310, no. 17, pp. 1842–1850, 2013.

[2] A. Osborn, D. Louis, T. Poussaint, L. Linscott, and K. Salzman, "The 2021 world health organization classification of tumors of the central nervous system: What neuroradiologists need to know," *American Journal of Neuroradiology*, vol. 43, no. 7, pp. 928–937, 2022.

[3] Q. T. Ostrom, H. Gittleman, P. Liao, T. Vecchione-Koval, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan, "Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2010–2014," *Neuro-oncology*, vol. 19, no. suppl_5, pp. v1–v88, 2017.

[4] W. Stummer, M. J. van den Bent, and M. Westphal, "Cytoreductive surgery of glioblastoma as the key to successful adjuvant therapies: new arguments in an old discussion," *Acta neurochirurgica*, vol. 153, no. 6, pp. 1211–1218, 2011.

[5] M. V. Brown, J. E. McDunn, P. R. Gunst, E. M. Smith, M. V. Milburn, D. A. Troyer, and K. A. Lawton, "Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies," *Genome medicine*, vol. 4, no. 4, p. 33, 2012.

[6] K.-C. Schafer, J. Balog, T. Szaniszló, D. Szalay, G. Mezey, J. Dénes, L. Bognár, M. Oertel, and Z. Takáts, "Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry," *Analytical Chemistry*, vol. 83, no. 20, pp. 7729–7735, 2011.

[7] L. S. Eberlin, I. Norton, D. Orringer, I. F. Dunn, X. Liu, J. L. Ide, A. K. Jarmusch, K. L. Ligon, F. A. Jolesz, A. J. Golby, *et al.*, "Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors," *Proceedings of the National Academy of Sciences*, vol. 110, no. 5, pp. 1611–1616, 2013.

[8] V. Pirro, C. M. Alfaro, A. K. Jarmusch, E. M. Hattab, A. A. Cohen-Gadol, and R. G. Cooks, "Intraoperative assessment of tumor margins during glioma resection by desorption electrospray ionization-mass spectrometry," *Proceedings of the National Academy of Sciences*, vol. 114, no. 26, pp. 6700–6705, 2017.

[9] S. Santagata, L. S. Eberlin, I. Norton, D. Calligaris, D. R. Feldman, J. L. Ide, X. Liu, J. S. Wiley, M. L. Vestal, S. H. Ramkissoon, *et al.*, "Intraoperative mass spectrometry mapping of an onco-metabolite to guide brain tumor surgery," *Proceedings of the National Academy of Sciences*, vol. 111, no. 30, pp. 11121–11126, 2014.

[10] D. Calligaris, D. R. Feldman, I. Norton, P. K. Brastianos, I. F. Dunn, S. Santagata, and N. Y. Agar, "Molecular typing of meningiomas by desorption electrospray ionization mass spectrometry imaging for surgical decision-making," *International journal of mass spectrometry*, vol. 377, pp. 690–698, 2015.

[11] D. Calligaris, D. R. Feldman, I. Norton, O. Olubiyi, A. N. Changelian, R. Machaidze, M. L. Vestal, E. R. Laws, I. F. Dunn, S. Santagata, *et al.*, "Maldi mass spectrometry imaging analysis of pituitary adenomas for near-real-time tumor delineation," *Proceedings of the National Academy of Sciences*, vol. 112, no. 32, pp. 9978–9983, 2015.

[12] A. K. Jarmusch, V. Pirro, Z. Baird, E. M. Hattab, A. A. Cohen-Gadol, and R. G. Cooks, "Lipid and metabolite profiles of human brain tumors by desorption electrospray ionization-ms," *Proceedings of the National Academy of Sciences*, vol. 113, no. 6, pp. 1486–1491, 2016.

[13] B. Fatou, P. Saudemont, E. Leblanc, D. Vinatier, V. Mesdag, M. Wisztorski, C. Focsa, M. Salzet, M. Ziskind, and I. Fournier, "In vivo real-time mass

spectrometry for guided surgery application," *Scientific reports*, vol. 6, no. 1, pp. 1–14, 2016.

[14] W. Stummer, U. Pichlmeier, T. Meinel, O. D. Wiestler, F. Zanella, H.-J. Reulen, A.-G. S. Group, *et al.*, "Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase iii trial," *The lancet oncology*, vol. 7, no. 5, pp. 392–401, 2006.

[15] A. Tsugu, H. Ishizaka, Y. Mizokami, T. Osada, T. Baba, M. Yoshiyama, J. Nishiyama, and M. Matsumae, "Impact of the combination of 5-aminolevulinic acid–induced fluorescence with intraoperative magnetic resonance imaging–guided surgery for glioma," *World neurosurgery*, vol. 76, no. 1-2, pp. 120–127, 2011.

[16] M. J. Colditz and R. L. Jeffree, "Aminolevulinic acid (ala)–protoporphyrin ix fluorescence guided tumour resection. part 1: Clinical, radiological and pathological studies," *Journal of Clinical Neuroscience*, vol. 19, no. 11, pp. 1471–1474, 2012.

[17] B. Montcel, L. Mahieu-Williame, X. Armoiry, D. Meyronet, and J. Guyotat, "Two-peaked 5-ala-induced ppix fluorescence emission spectrum distinguishes glioblastomas from low grade gliomas and infiltrative component of glioblastomas," *Biomedical optics express*, vol. 4, no. 4, pp. 548–558, 2013.

[18] Y. Li, R. Rey-Dios, D. W. Roberts, P. A. Valdés, and A. A. Cohen-Gadol, "Intraoperative fluorescence-guided resection of high-grade gliomas: a comparison of the present techniques and evolution of future strategies," *World neurosurgery*, vol. 82, no. 1-2, pp. 175–185, 2014.

[19] F.-K. Lu, D. Calligaris, O. I. Olubiyi, I. Norton, W. Yang, S. Santagata, X. S. Xie, A. J. Golby, and N. Y. Agar, "Label-free neurosurgical pathology with stimulated raman imaging," *Cancer research*, vol. 76, no. 12, pp. 3451–3462, 2016.

[20] M. Jermyn, J. Mercier, K. Aubertin, J. Desroches, K. Urmey, J. Karamchandiani, E. Marple, M.-C. Guiot, F. Leblond, and K. Petrecca, "Highly accurate

detection of cancer in situ with intraoperative, label-free, multimodal optical spectroscopy," *Cancer research*, vol. 77, no. 14, pp. 3942–3950, 2017.

[21] M. Jermyn, J. Desroches, J. Mercier, K. St-Arnaud, M.-C. Guiot, F. Leblond, and K. Petrecca, "Raman spectroscopy detects distant invasive brain cancer cells centimeters beyond mri capability in humans," *Biomedical optics express*, vol. 7, no. 12, pp. 5129–5137, 2016.

[22] D. T. M. Chan, H. Y.-P. Sonia, and W. S. Poon, "5-aminolevulinic acid fluorescence guided resection of malignant glioma: Hong kong experience," *Asian journal of surgery*, vol. 41, no. 5, pp. 467–472, 2018.

[23] D. A. Orringer, B. Pandian, Y. S. Niknafs, T. C. Hollon, J. Boyle, S. Lewis, M. Garrard, S. L. Hervey-Jumper, H. J. Garton, C. O. Maher, *et al.*, "Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated raman scattering microscopy," *Nature biomedical engineering*, vol. 1, no. 2, p. 0027, 2017.

[24] F. Poulon, H. Mehidine, M. Juchaux, P. Varlet, B. Devaux, J. Pallud, and D. A. Haidar, "Optical properties, spectral, and lifetime measurements of central nervous system tumors in humans," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[25] T. C. Hollon, S. Lewis, B. Pandian, Y. S. Niknafs, M. R. Garrard, H. Garton, C. O. Maher, K. McFadden, M. Snuderl, A. P. Lieberman, *et al.*, "Rapid intraoperative diagnosis of pediatric brain tumors using stimulated raman histology," *Cancer research*, vol. 78, no. 1, pp. 278–289, 2018.

[26] Z. Xue, L. Kong, C.-c. Pan, Z. Wu, J.-t. Zhang, and L.-w. Zhang, "Fluorescein-guided surgery for pediatric brainstem gliomas: Preliminary study and technical notes," *Journal of Neurological Surgery Part B: Skull Base*, vol. 79, no. S 04, pp. S340–S346, 2018.

[27] M. Gogiashvili, J. Nowacki, R. Hergenröder, J. G. Hengstler, J. Lambert, and K. Edlund, "Hr-mas nmr based quantitative metabolomics in breast cancer," *Metabolites*, vol. 9, no. 2, p. 19, 2019.

[28] E. O. Karakaslar, B. Coskun, H. Outilaft, I. J. Namer, and A. E. Cicek, "Predicting carbon spectrum in heteronuclear single quantum coherence spectroscopy for online feedback during surgery," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 2, pp. 719–725, 2019.

[29] D. Cakmakci, E. O. Karakaslar, E. Ruhland, M.-P. Chenard, F. Proust, M. Piotto, I. J. Namer, and A. E. Cicek, "Machine learning assisted intraoperative assessment of brain tumor margins using hrmas nmr spectroscopy," *PLoS Computational Biology*, vol. 16, no. 11, p. e1008184, 2020.

[30] D. Cakmakci, G. Kaynar, C. Bund, M. Piotto, F. Proust, I. J. Namer, and A. E. Cicek, "Targeted metabolomics analyses for brain tumor margin assessment during surgery," *Bioinformatics*, 2022.

[31] D. Cakmakci, E. O. Karakaslar, E. Ruhland, M.-P. Chenard, F. Proust, M. Piotto, I. J. Namer, and A. E. Cicek, "1h hrmas nmr dataset of glioma and control samples," July 2020.

[32] H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, *et al.*, "Biologically informed deep neural network for prostate cancer discovery," *Nature*, vol. 598, no. 7880, pp. 348–352, 2021.

[33] A. van Hilten, S. A. Kushner, M. Kayser, M. A. Ikram, H. H. Adams, C. C. Klaver, W. J. Niessen, and G. V. Roshchupkin, "Gennet framework: interpretable deep learning for predicting phenotypes from genetic data," *Communications biology*, vol. 4, no. 1, p. 1094, 2021.

[34] J. H. Oh, W. Choi, E. Ko, M. Kang, A. Tannenbaum, and J. O. Deasy, "Pathcnn: interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i443–i450, 2021.

[35] X. Liu, Y. Tao, Z. Cai, P. Bao, H. Ma, K. Li, Y. Zhu, and Z. J. Lu, "Pathformer: biological pathway informed transformer model integrating multimodal data of cancer," *bioRxiv*, pp. 2023–05, 2023.

[36] M. T. Poon, C. L. Sudlow, J. D. Figueroa, and P. M. Brennan, “Longer-term (2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis,” *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[37] D. M. Burgenske, J. Yang, P. A. Decker, T. M. Kollmeyer, M. L. Kosel, A. C. Mladek, A. A. Caron, R. A. Vaubel, S. K. Gupta, G. J. Kitange, *et al.*, “Molecular profiling of long-term idh-wildtype glioblastoma survivors,” *Neuro-oncology*, vol. 21, no. 11, pp. 1458–1469, 2019.

[38] D. Wong and S. Yip, “Finding a four-leaf clover—identifying long-term survivors in idh-wildtype glioblastoma,” *Neuro-oncology*, vol. 21, no. 11, p. 1352, 2019.

[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, no. 1, pp. 1–12, 2018.

[40] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, “Deephit: A deep learning approach to survival analysis with competing risks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[41] H. Kvamme and Ø. Borgan, “Continuous and discrete-time survival prediction with neural networks,” *Lifetime data analysis*, vol. 27, pp. 710–736, 2021.

[42] J. Wang, N. Chen, J. Guo, X. Xu, L. Liu, and Z. Yi, “Survnet: a novel deep neural network for lung cancer survival analysis with missing values,” *Frontiers in Oncology*, vol. 10, p. 588990, 2021.

[43] S. Akoka, L. Barantin, and M. Trierweiler, “Concentration measurement by proton nmr using the eretic method,” *Analytical Chemistry*, vol. 71, no. 13, pp. 2554–2557, 1999.

[44] G. Wider and L. Dreier, “Measuring protein concentrations by nmr spectroscopy,” *Journal of the American Chemical Society*, vol. 128, no. 8, pp. 2571–2576, 2006.

[45] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[46] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in medicine*, vol. 14, no. 1, pp. 73–82, 1995.

[47] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006.

[48] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[49] M. R. Segal, "Regression trees for censored data," *Biometrics*, pp. 35–47, 1988.

[50] M. LeBlanc and J. Crowley, "Relative risk trees for censored survival data," *Biometrics*, pp. 411–425, 1992.

[51] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statistics in medicine*, vol. 24, no. 24, pp. 3927–3944, 2005.

[52] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[53] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statistics in medicine*, vol. 32, no. 13, pp. 2173–2184, 2013.

[54] B. Efron, "The efficiency of cox's likelihood function for censored data," *Journal of the American statistical Association*, vol. 72, no. 359, pp. 557–565, 1977.

[55] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, pp. 89–99, 1974.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[57] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[58] S. Firdous, R. Abid, Z. Nawaz, F. Bukhari, A. Anwer, L. L. Cheng, and S. Sadaf, "Dysregulated alanine as a potential predictive marker of glioma—an insight from untargeted hrmas-nmr and machine learning data," *Metabolites*, vol. 11, no. 8, p. 507, 2021.

[59] J. de Groot and H. Sontheimer, "Glutamate and the biology of gliomas," *Glia*, vol. 59, no. 8, pp. 1181–1189, 2011.

[60] S. K. Natarajan and S. Venneti, "Glutamine metabolism in brain tumors," *Cancers*, vol. 11, no. 11, p. 1628, 2019.

[61] S. Ekici, B. B. Risk, S. G. Neill, H.-K. Shu, and C. C. Fleischer, "Characterization of dysregulated glutamine metabolism in human glioma tissue with 1h nmr," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[62] N. Tedeschi-Blok, J. Schwartzbaum, M. Lee, R. Miike, and M. Wrensch, "Dietary calcium consumption and astrocytic glioma: the san francisco bay area adult glioma study, 1991-1995," *Nutrition and cancer*, vol. 39, no. 2, pp. 196–203, 2001.

[63] D. Akhavan, T. F. Cloughesy, and P. S. Mischel, "mtor signaling in glioblastoma: lessons learned from bench to bedside," *Neuro-oncology*, vol. 12, no. 8, pp. 882–889, 2010.

[64] H. W. S. Tan, A. Y. L. Sim, and Y. C. Long, "Glutamine metabolism regulates autophagy-dependent mtorc1 reactivation during amino acid starvation," *Nature communications*, vol. 8, no. 1, p. 338, 2017.

[65] D. Meng, Q. Yang, H. Wang, C. H. Melick, R. Navlani, A. R. Frank, and J. L. Jewell, “Glutamine and asparagine activate mtorc1 independently of rag gtpases,” *Journal of Biological Chemistry*, vol. 295, no. 10, pp. 2890–2899, 2020.

[66] I. Pull, H. McIlwain, and R. Ramsay, “Glutamate, calcium ion-chelating agents and the sodium and potassium ion contents of tissues from the brain,” *Biochemical Journal*, vol. 116, no. 2, pp. 181–187, 1970.

[67] M. Ereciska, D. Nelson, I. Nissim, Y. Daikhin, and M. Yudkoff, “Cerebral alanine transport and alanine aminotransferase reaction: alanine as a source of neuronal glutamate,” *Journal of neurochemistry*, vol. 62, no. 5, pp. 1953–1964, 1994.

[68] J. T. Brosnan and M. E. Brosnan, “Glutamate: a truly functional amino acid,” *Amino acids*, vol. 45, pp. 413–418, 2013.

[69] S. Gad, “Glutathione,” in *Encyclopedia of Toxicology (Third Edition)* (P. Wexler, ed.), p. 751, Oxford: Academic Press, third edition ed., 2014.

[70] S. Pölsterl, “scikit-survival: A library for time-to-event analysis built on top of scikit-learn,” *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020.

[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

Appendix A

Implementation Details

The codebase for this research is implemented in Python and is publicly accessible at `https://github.com/ciceklab/PiDeeL`. Specifically, the RSF, CWGB, and CoxPH models were utilized from [70], while the deep learning models were developed using the PyTorch framework [71]. All models were trained and tested on a SuperMicro SuperServer 4029GP-TRT equipped with 2 Intel Xeon Gold 6140 Processors (2.3GHz, 24.75M cache), 251GB RAM, 6 NVIDIA GeForce GTX 2080 Ti GPUs (11GB, 352Bit), and 2 NVIDIA TITAN RTX GPUs (24GB, 384Bit). The average training times for the CoxPH, RSF, and DeepSurv models were 1.5 minutes, 14 minutes, and 2 minutes, respectively. The training of PiDeeL required an average of 1.5 minutes.

Appendix B

Supplementary Figures

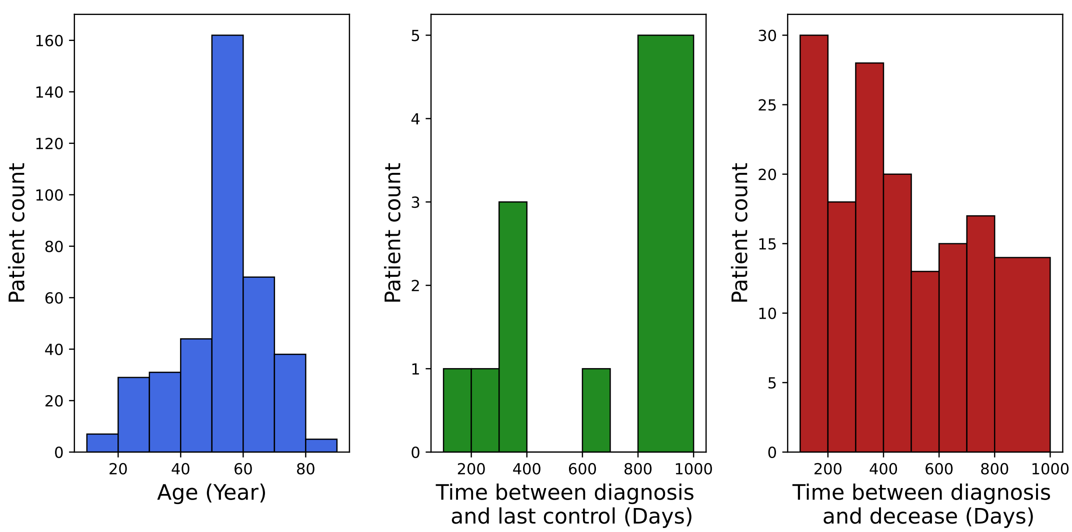

Figure B.1: (i) the age distribution of patients in the dataset, (ii) the distribution of the duration between the primary tumor removal surgery and the last follow-up, (iii) the distribution of the duration between the primary tumor removal surgery and death.

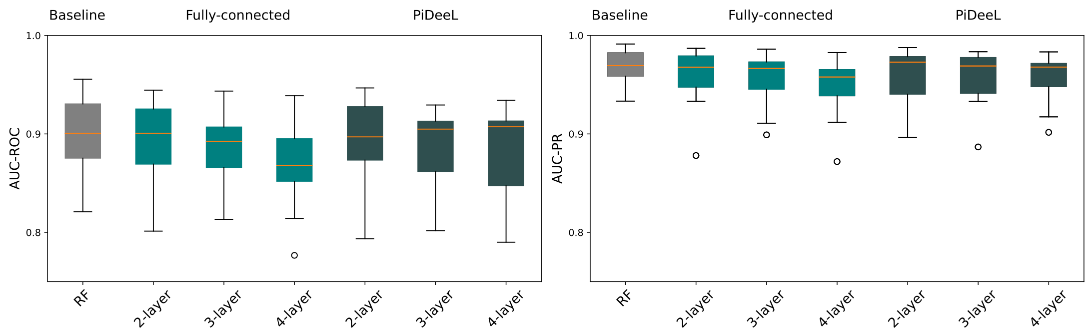

Figure B.2: Comparison of baseline models and PiDeeL on pathological classification in terms of AUC-ROC (left) and AUC-PR (right). This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

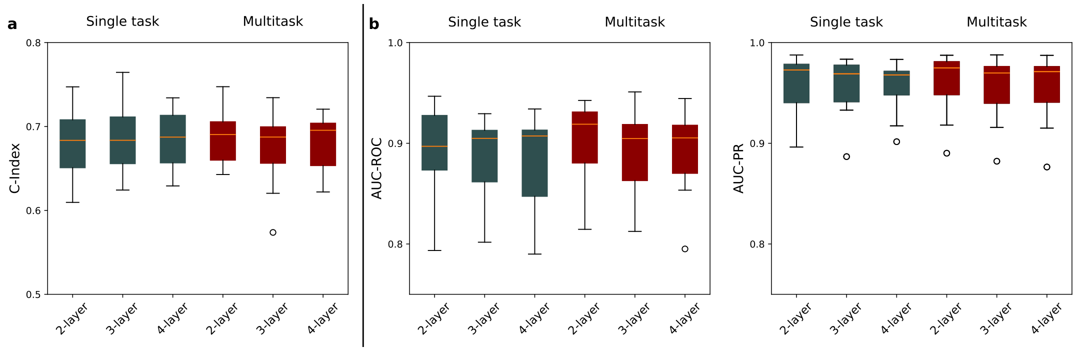

Figure B.3: Performance comparison between task-specific PiDeeL (navy blue boxplots) and multi-task learning of PiDeeL (red boxplots). The survival analysis performance is shown in Panel A, while the pathological classification performance is shown in Panel B, with AUC-ROC on the left and AUC-PR on the right. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

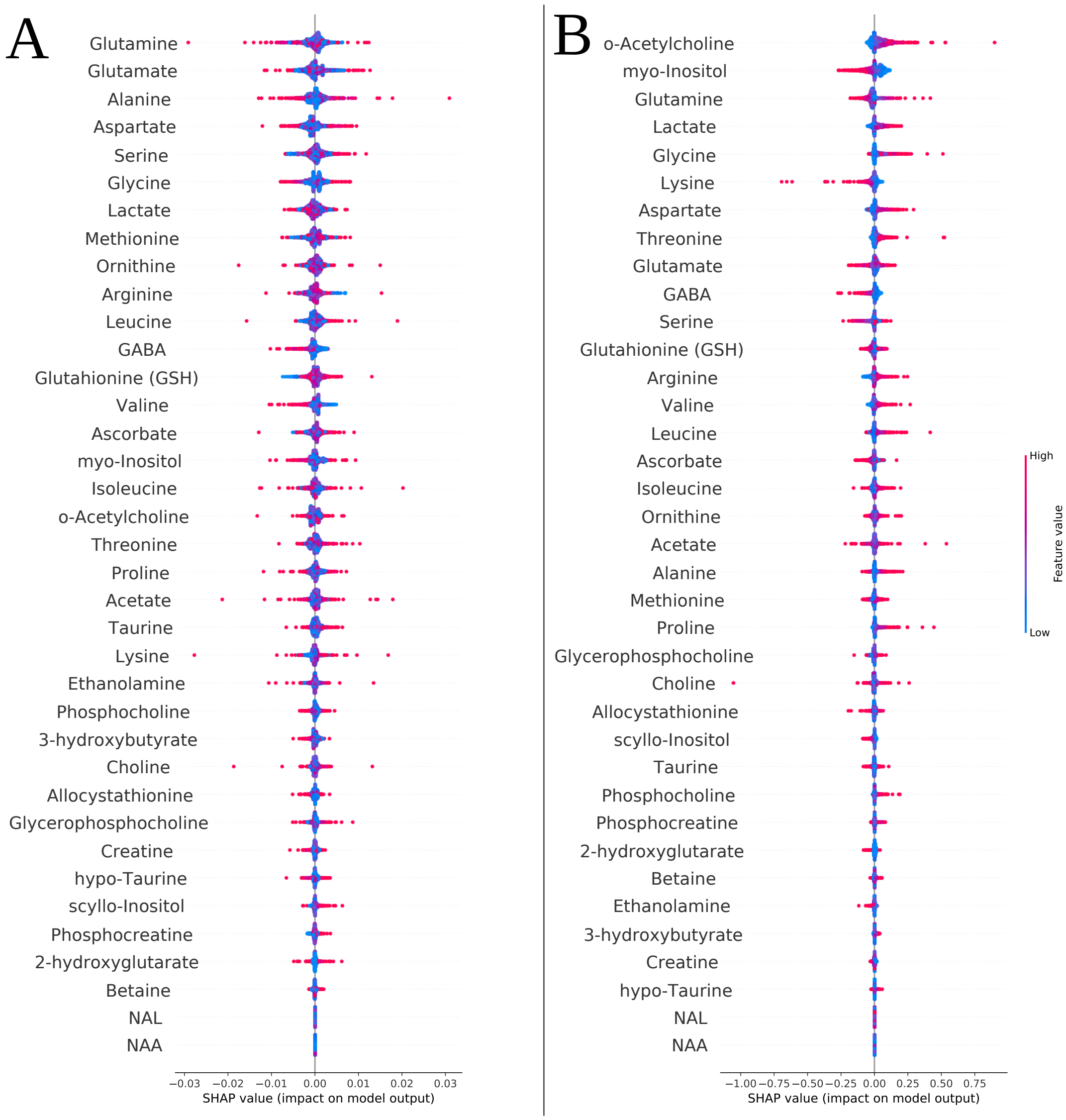

Figure B.4: UIn Panel A of the figure, the SHAP values (x-axis) for the survival analysis task of each quantified biomarker metabolite (y-axis) are displayed. In Panel B, the SHAP values (x-axis) for the pathological classification task of each quantified biomarker metabolite (y-axis) are shown. The color bar indicates that red represents a higher value and blue represents a lower value for a feature. Higher SHAP values indicate greater importance of the metabolite for the model's output towards a higher risk score in the survival analysis task and a high probability of malignant tumor in the pathological classification task.

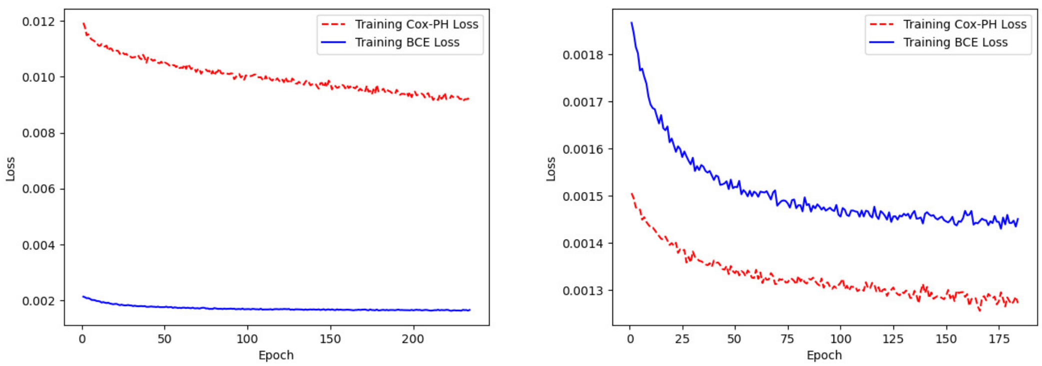

Figure B.5: The training loss convergence during training for the 3-layer PiDeeL using a random seed is depicted. On the left, the absolute values of Cox-PH and BCE losses are shown, while on the right, the scaled loss values are displayed. The weights for Cox-PH and BCE losses are 0.125 and 0.875, respectively.

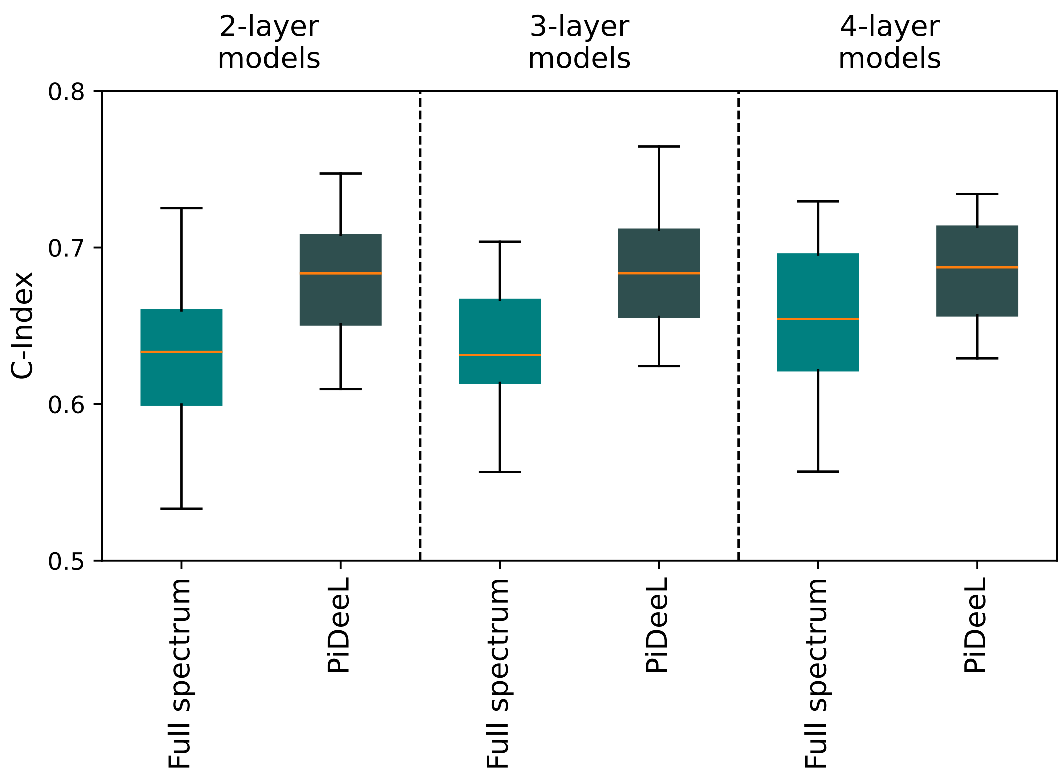

Figure B.6: Performance comparison of DeepSurv using full spectrum input and PiDeeL on survival analysis in terms of the c-index. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

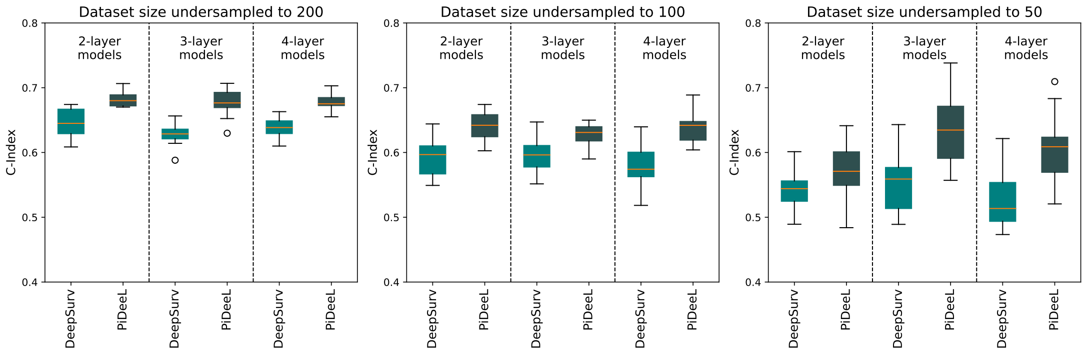

Figure B.7: Performance comparison of DeepSurv and PiDeeL on survival analysis in terms of the c-index across different dataset sizes. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

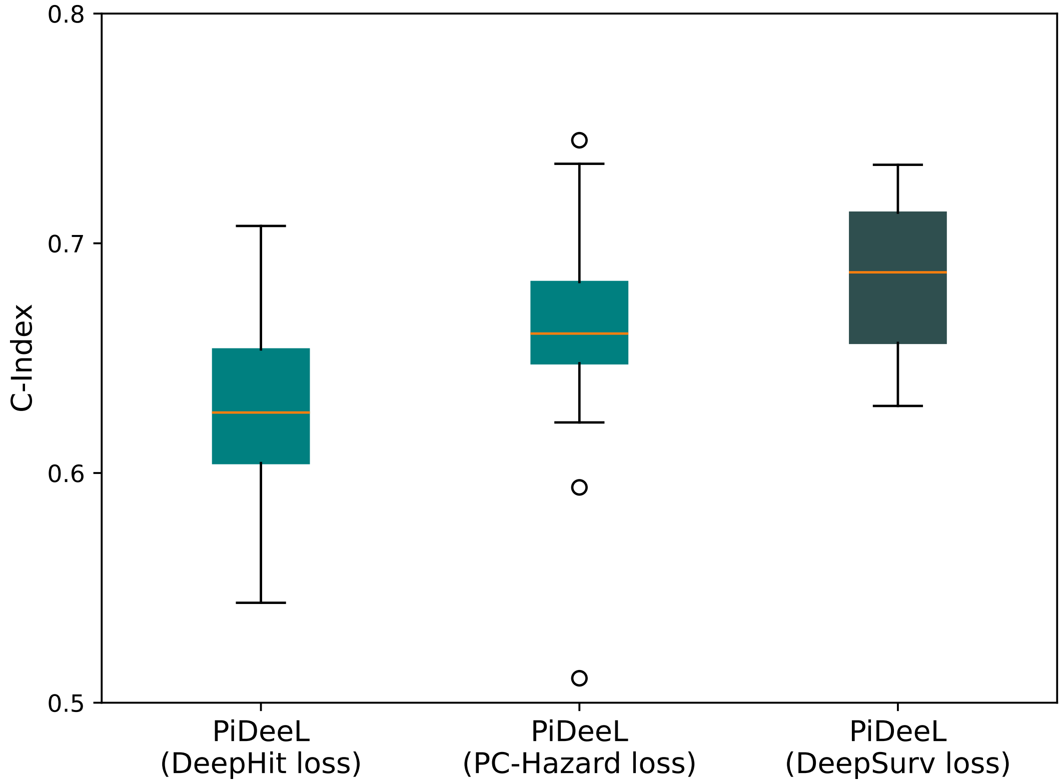

Figure B.8: Performance comparison of PiDeeL with different losses on survival analysis in terms of the c-index. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

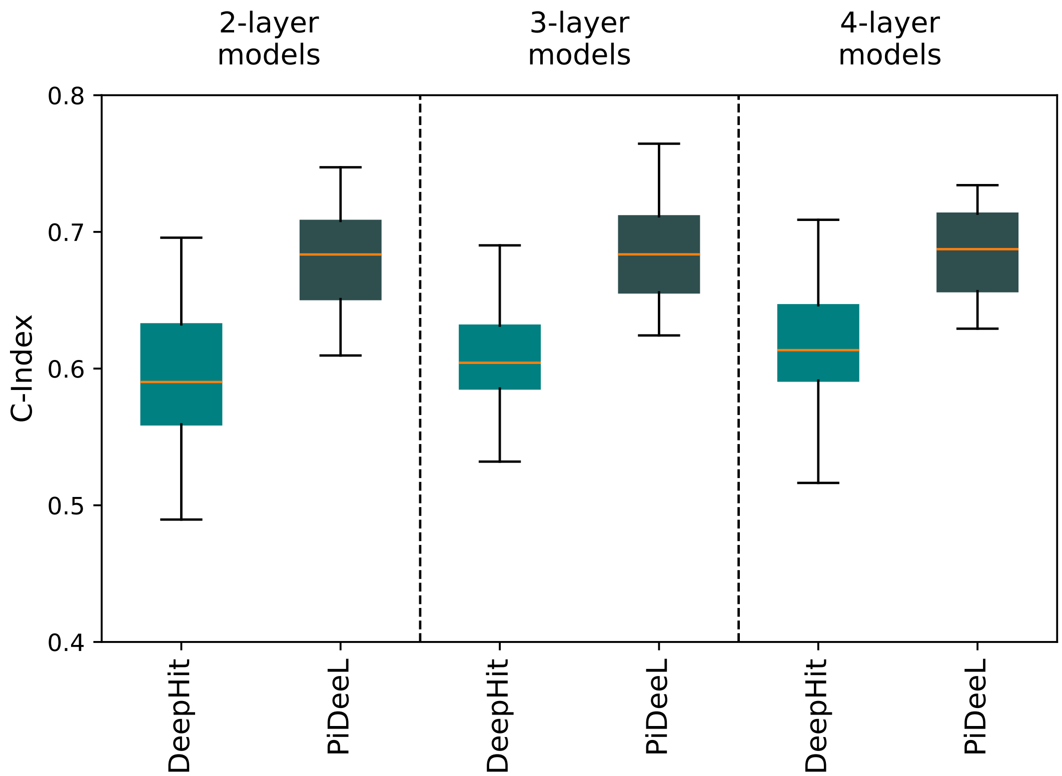

Figure B.9: Performance comparison of DeepHit and PiDeeL on survival analysis in terms of the c-index. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

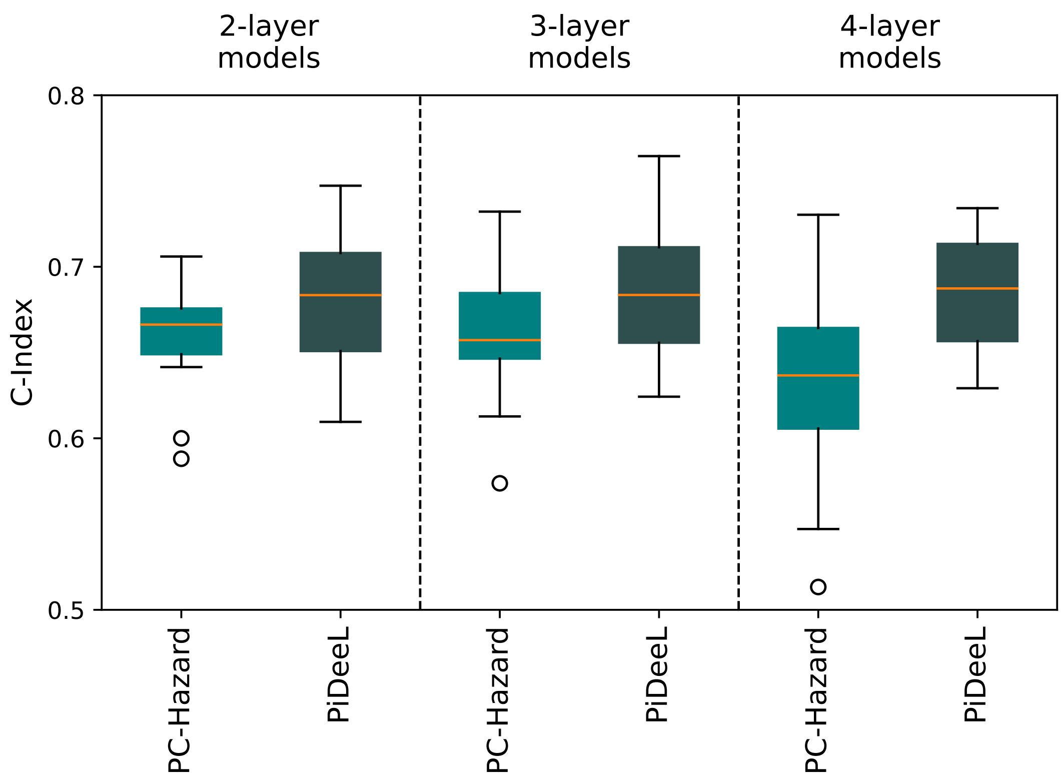

Figure B.10: Performance comparison of PC-Hazard and PiDeeL on survival analysis in terms of the c-index. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

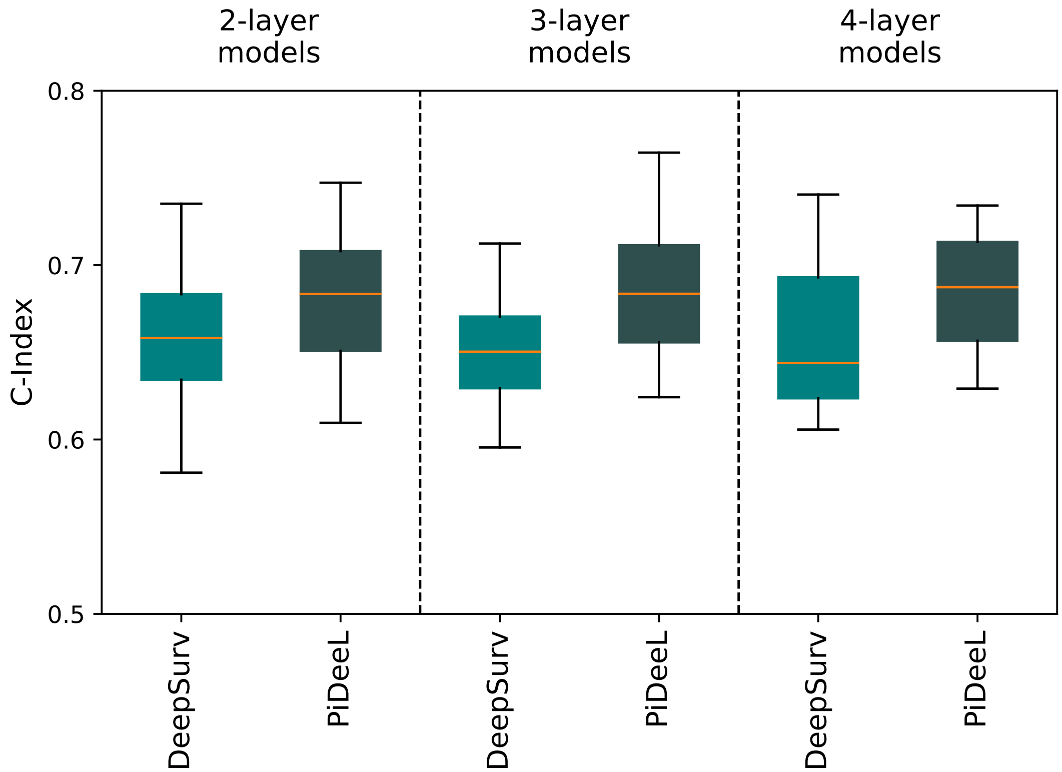

Figure B.11: Performance comparison of DeepSurv and PiDeeL on survival analysis in terms of the c-index. This evaluation is based on 5-fold cross-validation repeated 3 times, totaling 15 iterations.

Appendix C

Supplementary Table

Primary Brain Tumor Types	**Number of Samples**
Pilocytic astrocytoma (AST-I)	4
Astrocytoma grade II (AST-II)	5
Astrocytoma grade III (AST-III)	7
Glioblastoma (GBM)	178
Oligodendroglioma grade II (ODG-II)	36
Oligodendroglioma grade III[1] (ODG-III)	103
Oligoastrocytoma grade II (OAST-II)	3
Oligoastrocytoma grade III (OAST-III)	9
Ganglioma grade II (GG-II)	5
Ganglioma grade III (GG-III)	1
Dysembryoplastic Neuroepithelial Tumors (DNET)	25
Gliosarcoma (GS)	3
Unkown subtype	5

Table C.1: Histopathological classification of the tumor tissue specimens used in the thesis.

[1]Three samples had glioblastoma as well as oligodendroglioma grade III. They are included in the sample count for the latter.