Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Pooya Mobadersany Date

Predicting Time-to-Event and Clinical Outcomes from High-Dimensional Unstructured Data

By

Pooya Mobadersany
Doctor of Philosophy

Computer Science and Informatics

Lee Cooper, Ph.D.
Advisor

David Gutman, M.D., Ph.D.
Advisor

Shamim Nemati, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Predicting Time-to-Event and Clinical Outcomes from High-Dimensional Unstructured Data

By

Pooya Mobadersany
M.S., Emory University, 2018

Advisor: Lee Cooper, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Abstract

Predicting Time-to-Event and Clinical Outcomes from High-Dimensional Unstructured Data
By Pooya Mobadersany

This dissertation addresses challenges in learning to predict time-to-event outcomes such as survival and treatment response from high dimension data including whole slide images and genomic profiles that are being produced in modern pathology labs. Learning from these data requires integration of disparate data types, and the ability to attend to important signals within vast amounts of irrelevant data present in each sample. Furthermore, clinical translation of machine learning models for prognostication requires communicating the degree and types of uncertainty to clinical end users who will rely on inferences from these models.

This dissertation has addressed these challenges. To validate our developed data fusion technique, we have selected cancer histology data as it reflects underlying molecular processes and disease progression and contains rich phenotypic information predictive of patient outcomes. This study shows a computational approach for learning patient outcomes from digital pathology images using deep learning to combine the power of adaptive machine learning algorithms with survival models. We illustrate how these survival convolutional neural networks (SCNNs) can integrate information from both histology images and genomic biomarkers into a single unified framework to predict time-to-event outcomes and show prediction accuracy that surpasses the current clinical paradigm for predicting the overall survival of patients diagnosed with glioma. Next, to capture the volume of data and manage heterogeneity within the histology images, we have developed GestAltNet, which emulates human attention to high-yield areas and aggregation across regions. GestAltNet points toward a future of genuinely whole slide digital pathology by incorporating human-like behaviors of attention and gestalt formation process across massive whole slide images. We have used GestAltNet to estimate the gestational age from whole slide images of placental tissues and compared this to networks lacking attention and aggregation capabilities. To address the challenge of representing uncertainty during inference, we have developed a Bayesian survival neural network that captures the aleatoric and epistemic uncertainties when predicting clinical outcomes. These networks are the next generation of machine learning models for predicting time-to-event outcomes, where the degree and source of uncertainty are communicated to clinical end users.

Predicting Time-to-Event and Clinical Outcomes from High-Dimensional Unstructured Data

By

Pooya Mobadersany
M.S., Emory University, 2018

Advisor: Lee Cooper, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Acknowledgments

I would like to express my deepest gratitude to a person who I have the honor to call my Ph.D. advisor, Dr. Lee Cooper. I am thankful for his guidance and continued support during all these years which helped me move forward in my research. He was always willing to hear about new ideas and was always providing me the freedom in exploring new areas. I am very grateful to have him as my advisor and I have learned a lot from his excellence in research and dedication.

I want to sincerely thank the rest of my dissertation committee: Dr. Shamim Nemati from the University of California San Diego and Dr. David Gutman from Emory University. I personally have benefited a lot from their constructive feed-backs from the very beginning when I proposed the ideas, up until the end of my Ph.D. Their valuable feed-backs and presence in my thesis as my committee members have added a lot to the impact of my research.

Special thanks to Dr. Daniel Brat and Dr. Jeffery Goldstein from Northwestern University and Dr. Christopher Flowers from MD Anderson Cancer Center. None of this work would have happened without their domain-specific knowledge and feed-backs. I also want to thank the Pathology core at Northwestern University for providing institutional data for my research.

Thanks to my friends and collaborators from the Cancer Data Science Lab at Emory University: Saumya Gurbani, Sanghoon Lee, Michael Nalisnik, Mohamed Tageldin, and Safoora Yousefi. I would like to extend my sincere thanks to all my friends in the Department of Computer Science and Informatics at Emory University (all the amazing cohort in Emory CSI and BMI), Department of Mathematics at Emory University, Georgia Institute of Technology, and Department of Pathology at Northwestern University.

In the end, I would like to thank all my wonderful family and in-laws that have been always supportive of me. I would like to thank my grandparents, Akram and Mohammadreza, that passed away while I was far from home. I have learned a lot from them, and I am always carrying the deep sorrow of not getting the opportunity to say the last goodbye to them.

I want to thank my parents, Nasrin and Habib, for dedicating their whole life to me. With no hesitation, I am standing on their shoulders today and their spiritual contribution to this dissertation is as much as I did. I also thank my brother, Nima, for always encouraging me, teaching me, and guiding me through my life journey in every single aspect.

And last but not least I would like to thank my love of life, my wife, Soheyla. Words cannot explain how grateful I am to have her alongside this journey. I am thankful for her limitless support. None of this would have happened without having her hands as support on my shoulder.

To my dear wife Soheyla, who has always supported me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Histology has been an important tool in cancer diagnosis and prognostication for more than a century. Anatomic pathologists evaluate histology for characteristics like nuclear atypia, mitotic activity, cellular density, and tissue architecture, incorporating cytologic details and higher-order patterns to classify and grade lesions. Although prognostication increasingly relies on genomic biomarkers that measure genetic alterations, gene expression, and epigenetic modifications, histology remains an important tool in predicting the future course of a patient's disease. The phenotypic information present in histology reflects the aggregate effect of molecular alterations on cancer cell behavior and provides a convenient visual readout of disease aggressiveness. However, human assessments of histology are highly subjective and are not repeatable; hence, computational analysis of histology imaging has received significant attention.

Many important problems in the clinical management of cancer involve time-to-event prediction, including accurate prediction of overall survival and time to progression. Despite overwhelming success in other applications, deep learning has not been widely applied to these problems. Survival analysis has often been approached

as a binary classification problem by predicting dichotomized outcomes at a specific time point (e.g., 5-y survival) [1]. The classification approach has significant limitations, as subjects with incomplete follow-up cannot be used in training, and binary classifiers do not model the probability of survival at other times. Time-to-event models, like Cox regression, can utilize all subjects in training and model their survival probabilities for a range of times with a single model. Neural network-based Cox regression approaches were explored in early machine learning work using datasets containing tens of features, but subsequent analysis found no improvement over basic linear Cox regression [2]. More advanced "deep" neural networks that are composed of many layers were recently adapted to optimize Cox proportional hazard likelihood and were shown to have equal or superior performance in predicting survival using genomic profiles containing hundreds to tens of thousands of features and using basic clinical profiles [3].

Learning survival from histology is considerably more difficult. Time-to-event prediction faces many of the same challenges as other applications where CNNs are used to analyze histology. Compared with genomic or clinical datasets, where features have intrinsic meaning, a "feature" in an image is a pixel with meaning that depends entirely on context. Convolution operations can learn these contexts, but the resulting networks are complex, often containing more than 100 million free parameters, and thus, large cohorts are needed for training. This problem is intensified in time-to-event prediction, as clinical follow-up is often difficult to obtain for large cohorts. Data augmentation techniques have been adopted to address this problem, where randomized label-invariant positional and color transformations are used to synthesize additional training data [4, 5, 6, 7, 8, 9, 10, 11, 12]. Intratumoral heterogeneity also presents a significant challenge in time-to-event prediction, as a tissue biopsy often contains a range of histologic patterns that correspond to varying degrees of disease progression or aggressiveness. The method for integrating information from

heterogeneous regions within a sample is an important consideration in predicting outcomes. Furthermore, risk is often reflected in subtle changes in multiple histologic criteria that can require years of specialized training for human pathologists to recognize and interpret. Developing an algorithm that can learn the continuum of risks associated with histology can be more challenging than for other learning tasks, like cell or region classification.

To date, convolutional neural networks have proved to be the most powerful model for learning representation from images in different domains and applications [13, 14, 15, 16, 17, 18]. Most networks are limited to analyzing and learning from a single image patch or a batch of randomly selected image patches at a time. Conversely, the daily practice of pathologists is to examine a Whole Slide Image (WSI), identify the features present, weigh evidence supporting competing interpretations, and aggregate over those features to come to a final evaluation of the patient's status. This task of gestalt formation is common in pathology and medical imaging but is also relevant in broader domains of complicated image analysis where the whole is greater than the sum of its parts. Learning in aggregate addresses a fundamental challenge in applying convolutional networks in digital pathology and can solve a wide variety of problems.

Furthermore, when modeling outcome for patients based on their histologic or genomic features, samples with rare characteristics might yield erroneous identification of representative patterns which might have catastrophic consequences when deployed in clinical practice [19]. This makes it important to model uncertainty in survival prediction models that can yield to more reliable identification of new patterns to use in clinical practice.

1.2 Research Contributions

Pathology driven prognostication is critical for realizing treatment strategies to optimize the quality of life and survival for patients. Genomics holds promise for improving the classification and prognostication of malignancies, yet oncology practice continues to rely heavily on histopathology analysis as a fundamental tool due to its ability to provide information about cancer severity and patient status. Evaluating these pathology images remains a largely manual and subjective practice among the pathologists, where they spend a lot of time analyzing these images by zooming in and out and panning throughout these huge images, looking for some histologic characteristics that represent the patient outcome. This procedure is highly biased by the pathologist's knowledge and state of mind and can lead to highly variant and subjective prognostications. This manual procedure generally cannot account for variations in tissue processing present in multi-institution studies, cannot be reproducible, and cannot be meaningfully extended or integrated with existing resources to optimize classification and prognostication strategies. New learning tools are needed for robust histopathology analysis and for the integration of these images and clinical outcomes. There might be a lot of latent patterns in pathology images that pathologists don't recognize while studying these WSIs. However, these patterns once revealed might help improve their prediction accuracy and classification of patients.

The goal of this dissertation is to develop algorithms for quantitative analysis of Hematoxylin and Eosin (H&E) stained histopathology slides and to enable improved prognostication and outcome prediction from histology and genomics through unbiased learning algorithms. This research builds on the development of machine learning methods for prognostication in high-dimensional unstructured data and enables predicting patient survival time from Whole Slide Images (WSI), gives a better solution for incorporating more data from these large-scale images, and introduces a pipeline for Bayesian modeling of uncertainties in survival prediction. The rest of

this section highlights the details of our research contributions.

1.2.1 Survival Prediction based on Convolutional Neural Network (Chapter 3)

In this section, I consider predicting patient outcome from histology images and propose a model that enables end-to-end extraction of features inside each High Power Field (HPF) sampled from WSIs and prediction of patient outcome based on these features. This model is based on Convolutional Neural Networks (CNN) because they are a class of machine learning models that have been proved to outperform other models in many image prediction and classification tasks by capturing the visual patterns in images [13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24]. CNNs are unbiased dissimilar to how pathologists are, which makes them a great candidate to use in this field. We name our model Survival Convolutional Neural Network (SCNN), which provides highly accurate predictions of time-to-event outcomes from histology images. Our SCNN framework uses an image sampling and risk filtering technique that significantly improves prediction accuracy by mitigating the effects of intratumoral heterogeneity and deficits in the availability of labeled data for training. I use heat map visualization techniques applied to whole-slide images to show how SCNN learns to recognize important histological structures that neuropathologists use in grading diffuse gliomas and suggest relevance for patterns with prognostic significance that is not currently appreciated. I systematically validate our approaches by predicting overall survival in gliomas using data from The Cancer Genome Atlas (TCGA) Lower-Grade Glioma (LGG) and Glioblastoma (GBM) projects. To integrate both histologic and genomic data into a single unified prediction framework, I developed a Genomic Survival Convolutional Neural Network (GSCNN) that enables end-to-end learning and inference from genomic and histology data fusion. The GSCNN learns from genomics and histology simultaneously by incorporating genomic data into the

fully connected layers of the SCNN. Both data are presented to the network during training, enabling genomic variables to influence the patterns learned by the SCNN by providing molecular subtype information.

1.2.2 Architectures for Aggregate Learning (Chapter 4)

In this chapter, I consider the problem of learning the outcome from a set of images. Convolutional neural networks have proved to be the most powerful model for learning representation from images in different domains and applications. Most networks are limited to analyzing and learning from a single HPF or batch of randomly selected HPF at a time. In these traditional learning models training explicitly links clinical outcome to a single HPF (Figure 1.1.A). While this approach might be practical in tasks like tumor detection, for more challenging tasks such as predicting patient survival, the daily practice of pathologists is to examine a WSI, identify the features present, weigh evidence supporting competing interpretations, and aggregate over those features to come to a final evaluation of the patient's status. This task of gestalt formation is not only common in pathology and medical imaging but is also relevant in broader domains of complicated image analysis where the whole is greater than the sum of its parts. For these challenging applications, contrary to CNN's basic assumption of one image corresponding to one label, it is a collection of Regions of Interest (ROIs) that correspond to each patient outcome. So, in applications such as predicting patient outcome from histology, an ideal model has to learn in aggregate from a collection of images. Therefore, we need to introduce a training model that links the clinical outcome to a set of HPFs (Figure 1.1.B). In this section, I have developed a model that learns to predict patient outcome from a collection/set of HPFs. This enables incorporating more regions from each WSI during the learning procedure. Our end-to-end pipeline has 3 key features - glimpsing, attention, and aggregation which altogether emulate human attention to high-yield areas and aggre-

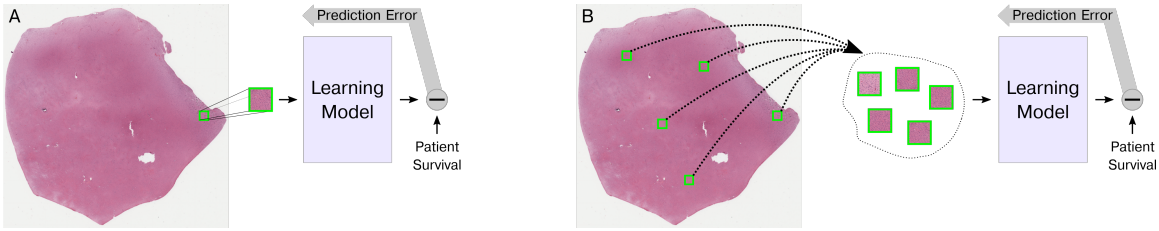

Figure 1.1: Different Learning strategies. (A) Traditional pipeline: learning the outcome from single HPF. (B) Aggregate learning: learning the outcome from the collection of HPFs.

gation across regions. I used this network to estimate the gestational age (GA) of scanned placental slides and compared it to a similar network lacking the attention and aggregation functions. Our proposed model, GestaltNet, points toward a future of genuinely whole-slide digital pathology by incorporating human-like behaviors of attention and aggregation.

1.2.3 Bayesian Neural Networks for Survival Prediction (Chapter 5)

In this chapter, I investigate the use of Bayesian neural networks in modeling the aleatoric and epistemic uncertainties in survival prediction. In general, it is important to build models that enable uncertainty quantification in tasks where the lack of confidence in predictions can have catastrophic effects. Besides, it is important to appropriately differentiate the epistemic and aleatoric uncertainties, as each of them has different underlying causes that require different handling. The aleatoric uncertainty is based on the noise or some inherent variability in the data, such as mislabeled data; this type of uncertainty cannot be addressed by incorporating more samples into model training but is reducible by measuring additional features. Whereas, the epistemic uncertainty is based on the uncertainty in model parameters and is reducible by incorporating more samples into model training. In applications such as predicting outcomes for cancer patients, underlying uncertainty in prediction can have a direct

impact on our decisions on the individualized course of treatment for patients. In such tasks, erroneous predictions might result in irreparable outcomes. In practice, we cannot deploy a model in medical settings, unless we build a model that *reliably knows what it doesn't know* and asks for extra measurements or the human specialist to intervene in the decision making process in cases where there is a lack of confidence in its predictions. Most traditional deep learning settings make it impossible to quantify the uncertainty. This includes both the applications of deep learning for regression where the output is generated as a single scalar value and classification where the output is often represented as a normalized vector obtained from a softmax layer [19]. To be able to build confidence intervals over predictions, the model has to have the ability to represent the output as a posterior distribution that is dependent on the potential noises in input features caused by the errors in measurements and/or potential variations in the data-driven model parameters caused by the variations in the number of samples at the model's search space. Dropout which has been initially introduced as a regularization method in neural networks to prevent overfitting [25] has been proposed as a means of Bernoulli approximation for the epistemic uncertainty. The Bernoulli dropout is potentially estimating the uncertainty in model parameters by applying an approximate variational Bernoulli distribution over these parameters [26]. This approach is called Monte Carlo Dropout (MCD). The quality of MCD in modeling uncertainty is contested. There have been studies illustrating that the distribution generated by MCD can be a poor approximation to most reasonable Bayesian posteriors, hence, yielding to bad decisions [27, 28]. Ian Osband in [28] notes that for a simple deep neural network the predictive uncertainty computed by MCD does not decrease with more data. This raises the question of whether MCD is a good approximation to a Bayesian posterior and whether MCD estimates epistemic uncertainty as the developers of MCD suggest. Ian Osband suggests that MCD estimate is in fact an approximation to the aleatoric uncertainty, not

to the epistemic uncertainty; hence, it seems that MCD is only applicable under only specific assumptions.

In recent years, there have been a lot of improvements in building deep survival neural networks for survival prediction [29, 3, 30]. However, there have been a few works that have addressed the problem of modeling uncertainty in survival prediction [31, 32]. Also, it is unclear which type of uncertainty most of these proposed methods for deep survival networks are modeling, and there have not been enough attempts to differentiate these different sources of uncertainties in survival neural networks, despite its critical role in building reliable survival models. Therefore, this chapter seeks to model the aleatoric and epistemic uncertainties in deep survival neural networks under a Bayesian framework.

Chapter 2

Background & Related Work

2.1 Survival Analysis

Survival analysis is the branch of statistics where the random variable under study is the time to event data. In this dissertation, we are focused on medical applications. In medical applications, the event might be the death of the patient or the recurrence of a disease. Analyzing survival in this domain is important as it can reveal critical information regarding optimized treatments for each individual patient. In this section, we will briefly explain the common issue of censoring in survival analysis and will provide a brief background over different approaches for modeling survival.

Censoring

One of the common issues in survival analysis is the censored data, where there might be some subjects for which the event is not observed because they might have left the study, lost to follow up, the study has been terminated before every subject has shown the event of interest or the true event date is not known. It is important to incorporate censored subjects into the study as it can provide information about long-term survivors. The censoring in the survival analysis context is usually assumed

to be random. This type of censoring is called non-informative censoring, where the censoring times are statistically independent of the event of interest. Non-informative censoring can be because of a predefined termination time for study, or some random loss of follow-up over patients irrespective of their disease status. In this entire dissertation, we assume the censoring is non-informative. It is important to appropriately characterize the censoring type in the study, as assuming a non-informative censoring where the study is subject to informative censoring yields inaccurate statistical inference about patients' survival.

In general, there are three types of censoring: right-censoring, left-censoring, and interval-censoring. Right-censoring, which is the most common type of censoring in survival analysis data happens where the true unobserved event is to the right of the censoring time. Left-censoring occurs when the event of interest has already happened before including the subject in the study but the true event time is not known. Interval-censoring happens when the only information about the subject event time is that the event happened between two examinations (check points). In this entire dissertation, the censoring type of data will be right-censored. Different censoring types are all illustrated in Figure 2.1.

Survival function

Let T be a non-negative continuous random variable that represents time to event data for subjects of interest. For instance, if the subjects of interest are patients with a specific disease, then T might be their time to death or time to recurrence of the disease after treatment for those patients. The survival function $S(t)$ is a function that gives the probability that the event of interest occurs after any specific time t and is given by:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \tag{2.1}$$

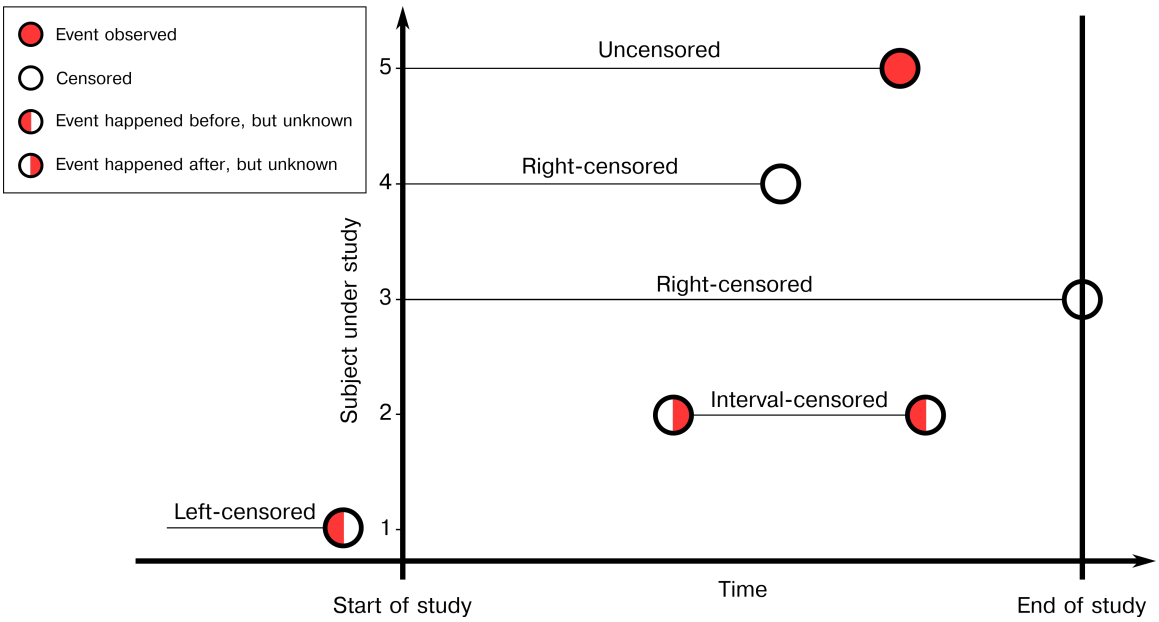

Figure 2.1: Different types of censored and uncensored subjects. Subject 1 is left-censored as the event has already happened before the start of the study but it is unknown; subject 2 is interval-censored as we only know the interval in which the event has happened; subject 3 is right censored as the event is not observed by the end of the study; subject 4 is right-censored as the subject is lost to follow up; subject 5 is uncensored as the true event is observed.

where $f(t)$ is the probability density function (PDF) for time to event random variable T, and $F(t)$ is the cumulative distribution function (CDF) which is given by:

$$F(t) = P(T \leq t) = \int_0^t f(u)du, \quad T \in [0, \infty) \tag{2.2}$$

Hazard function and cumulative hazard function

Besides the survival function, there is another way to model data distribution in survival analysis which is called hazard function. Hazard function models the chances of the event in each minuscule period of time given that the event of interest has not happened before that time frame and represents the instantaneous risk that an event of interest will happen within each very narrow time frame. Therefore, hazard function $h(t)$ is given by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t . S(t)} = \frac{f(t)}{S(t)} \tag{2.3}$$

Although hazard function is often thought of as the probability of event at each small time frame, it can exceed 1, hence, it is not a probability.

Cumulative hazard function is the accumulation of hazard over time and is an alternative to represent the hazard function. Given the hazard function $h(t)$ the cumulative hazard function is defined as follows:

$$H(t) = \int_0^t h(u)du = -log(S(t)) \tag{2.4}$$

Therefore, the survival and cumulative hazard functions are related by the following equation:

$$S(t) = e^{-H(t)}, \quad t > 0 \tag{2.5}$$

2.2 Survival models

In this section we will briefly describe different methods for modeling the distribution of survival time or time to event (T).

2.2.1 Parametric survival models

In parametric survival models, we assume a parametric form for the distribution of time to event (T). We can consider any distribution that is defined for $t \in [0, \infty)$ as a survival distribution. In general, we can convert all distributions that are defined for $x \in (-\infty, \infty)$ to a survival distribution by considering $t = e^x$, which maps the $(-\infty, \infty)$ space to $(0, \infty)$.

Exponential, Weibull and log-logistic are three of the most popular distributions for survival time or the time to event. In order to model the effect of covariates on each of these parametric models, their distribution parameters can be defined as a function of covariates.

Exponential

The exponential distribution is one of the continuous probability distributions that has only one parameter which is called rate parameter ($\lambda > 0$) [33]. The exponential distribution is the only continuous distribution that assumes a constant hazard function:

$$h(t) = \lambda, \quad \lambda > 0 \tag{2.6}$$

Considering equation 2.6, the cumulative hazard function for exponential distribution is $H(t) = \lambda t$. Plugging this into the equation 2.5, yields to the following survival function for the exponential distribution:

$$S(t) = e^{-\lambda t} \tag{2.7}$$

The PDF for the exponential distribution is:

$$f(t \mid \lambda) = \lambda e^{-\lambda t}, \quad t \geq 0 \tag{2.8}$$

Therefore, in exponential distribution,

$$P(T > t + \Delta t \mid T > t) = \int_{t+\Delta t}^{\infty} \lambda e^{-\lambda x} dx = e^{-\lambda(t+\Delta t)} = e^{-\lambda t} e^{-\lambda(\Delta t)} \tag{2.9}$$

And as we know from equation 2.7, $e^{-\lambda t}$ is the survival function at time t. Given that we know the event has happened after t, we will have $P(T > t) = 1$. Hence, $e^{-\lambda t} = 1$, which results in:

$$P(T > t + \Delta t \mid T > t) = e^{-\lambda(\Delta t)} = S(\Delta t) = P(T > \Delta t) \tag{2.10}$$

Therefore, the exponential distribution is a memoryless distribution, as $P(T > t + \Delta t \mid T > t) = P(T > \Delta t)$, and it might not be an appropriate distribution to pick where the knowledge that an event has not happened until a specific time changes the probabilities for the occurrence of that event or when the hazard varies over time. For instance, when modeling the distribution of recurrence of a disease after treatment, assuming exponential distribution is not appropriate if the knowledge that the disease has not recurred until a specific time after treatment changes the probability of its recurrence.

Weibull distribution

The Weibull distribution is another continuous probability distribution that is a generalization for the exponential distribution and is particularly popular in survival

analysis, as it can accurately model the time to events and is flexible despite having only two parameters [34]. These parameters are scale ($\lambda > 0$) and shape ($k > 0$). The hazard function for Weibull distribution is as follows:

$$h(t) = (\frac{k}{\lambda})(\frac{t}{\lambda})^{k-1} \tag{2.11}$$

Depending on the value of shape parameter, its hazard function takes different shapes. Whenever, $k < 1$ its hazard function will be monotonically decreasing over time, when $k = 1$ its hazard function becomes a constant function over time and Weibull distribution reduces to an exponential distribution. And, when $k > 1$ the hazard function will be monotonically increasing over time. The PDF for the Weibull distribution is:

$$f(t \mid \lambda, k) = 1 - e^{-(t/\lambda)^k}, \quad t \geq 0 \tag{2.12}$$

This yields to the following survival function for Weibull distribution:

$$S(t) = e^{-(t/\lambda)^k} \tag{2.13}$$

Log-logistic distribution

The log-logistic distribution is another continuous probability distribution that is widely used in survival analysis [35]. Similar to Weibull distribution, there is two parameters for a log-logistic distribution. These parameters are scale ($\alpha > 0$) and shape ($\beta > 0$). The PDF for log-logistic distribution is:

$$f(t \mid \alpha, \beta) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{(1 + (t/\alpha)^{\beta})^2}, \quad t \geq 0 \tag{2.14}$$

and its survival function is:

$$S(t) = \frac{1}{1 + (t/\alpha)^{\beta}} \tag{2.15}$$

The hazard function for log-logistic distribution is:

$$h(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^{\beta}} \tag{2.16}$$

Unlike the Weibull distribution, the log-logistic distribution can have a non-monotonic hazard function. Whenever $\beta > 1$, its hazard function becomes unimodal and when $\beta \geq 1$, the hazard function becomes monotonically decreasing over time. Therefore, the log-logistic distribution covers more possible shapes for hazard function compared to the Weibull distribution and is a reasonable choice for instances where the failure rate increases initially and decreases later, e.g. when the event of interest is the mortality of cancer patients after diagnosis or treatment.

If the censoring times are all known constants, the likelihood of the event of interest for all the aforementioned models is the multiplication of PDFs for uncensored subjects and survival functions for censored samples as follows:

$$l(t, \delta) = \prod_i f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{2.17}$$

or by replacing $f(t) = h(t)S(t)$ based on the equation 2.3 we can rewrite the likelihood function as:

$$l(t, \delta) = \prod_i h(t_i)^{\delta_i} S(t_i) \tag{2.18}$$

where i is the subject number, and δ_i is the event indicator ($\delta_i = 1$: event observed, $\delta_i = 0$: censored).

Taking the logarithm of the likelihood function will result in the following log likelihood function:

$$\mathcal{L}(t,\delta) = \sum_i \Big(\delta_i log\big(h(t_i)\big) + log\big(S(t_i)\big)\Big) \tag{2.19}$$

which by replacing the equation 2.4 results in following log likelihood function:

$$\mathcal{L}(t,\delta) = \sum_i \Big(\delta_i log\big(h(t_i)\big) - H(t_i)\Big) \tag{2.20}$$

that, if we consider the model's distribution parameters as functions of covariates, then this log likelihood function can be maximized over the coefficients of the underlying parametric model's distribution parameters to produce maximum likelihood estimates of the model parameters.

One of the limitations of the parametric models is that they might not cover all distributions of outcomes. An alternative to address this issue is using mixture distributions that we will explain briefly in the following subsection.

Mixture distributions

Mixture distributions are the convex combination of the distributions of the same family. The convex combination is a weighted sum, with non-negative weights that sum to 1 altogether, which guarantees that the mixture of the probability distributions will still be a probability distribution. The individual distributions that form the mixture distribution are called the mixture components, and the weights associated with each of these component are called the mixture weights. There is no limit in selecting the number of components in a mixture distribution. Assuming that $f_n(t)$, $F_n(t)$, $S_n(t)$, $h_n(t)$ and $H_n(t)$ are the PDF, CDF, survival function, hazard function and cumulative hazard function of the n^{th} component in the mixture, respectively, and w_n is the mixture weight of the n^{th} component we will have the following equations for a mixture distribution that consists of N components:

$$f(t) = \sum_{n=1}^{N} w_n f_n(t), \quad where \sum_{n=1}^{N} w_n = 1 \tag{2.21}$$

$$F(t) = \sum_{n=1}^{N} w_n F_n(t) \tag{2.22}$$

based on equation 2.1 this results in:

$$S(t) = 1 - F(t) = 1 - \sum_{n=1}^{N} w_n F_n(t) \tag{2.23}$$

as we know in the mixture distribution we have $\sum_{n=1}^{N} w_n = 1$, by replacing it in the equation 2.23 this results in:

$$S(t) = 1 - \sum_{n=1}^{N} w_n F_n(t) = \sum_{n=1}^{N} w_n - \sum_{n=1}^{N} w_n F_n(t) = \sum_{n=1}^{N} w_n (1 - F_n(t)) = \sum_{n=1}^{N} w_n S_n(t) \tag{2.24}$$

and based on equations 2.3 and 2.4, following will be the hazard function and cumulative hazard function for the mixture distribution:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\sum_{n=1}^{N} w_n f_n(t)}{\sum_{n=1}^{N} w_n S_n(t)} \tag{2.25}$$

and

$$H(t) = -log\big(S(t)\big) = -log\left(\sum_{n=1}^{N} w_n S_n(t)\right) \tag{2.26}$$

The likelihood and log likelihood functions of the mixture distribution will exactly be the same as in equations 2.18, 2.19 and 2.20. Similar to the single distribution models the log likelihood function can be maximized over the coefficients of the underlying mixture distribution parameters to produce maximum likelihood estimates

of the model parameters.

2.2.2 Non-parametric survival models

In this section, we will briefly describe the survival models that do not have any parametric assumption over the distribution of time to event.

Kaplan–Meier estimator

The Kaplan–Meier estimator or product limit estimator is a non-parametric step function with discontinuities in the form of jumps at the observed death times, that estimates the survival function [36] in the presence of censoring. Let d_i be the number of events at t_i and n_i be the number of individuals exposed to risk at time t_i, or in another terms n_i is the number of individuals that have not yet had an event or been censored just before time t_i. Then, the Kaplan-Meier estimator of survival function is:

$$\hat{S}(t) = \prod_{i:t_i<t} (1 - \frac{d_i}{n_i}) \tag{2.27}$$

While Kaplan-Meier estimate might be useful to examine the probability of event in the form of population level survival, or in the medical applications to compare the effectiveness of treatment methods, it is limited in its ability to estimate individual level survival that is adjusted for covariates. However, the parametric survival models and the Cox proportional hazards model which is a semi-parametric survival model that will be discussed in next sub-section, can be useful to estimate individual level survivals by taking into account the covariates for each individual.

Nelson-Aalen estimator

The Nelson-Aalen estimator is another non-parametric approach for modeling survival that is used to directly obtain an estimation of cumulative hazard function $H(t)$ [37, 38, 39]. The Nelson-Aalen estimate of cumulative hazard function is:

$$\hat{H}(t_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \tag{2.28}$$

The Nelson-Aalen estimator estimates the hazard at each distinct time of event t_j as the ratio of the number of events to the number of individuals exposed to risk. In equation 2.28, d_j is the number of events at t_j and n_j is the number of individuals exposed to risk at time t_j. Having the estimation of cumulative hazard function from equation 2.28, and plugging it into the equation 2.5 the estimation of the survival function can be obtained as follows:

$$\hat{S}(t) = e^{-\hat{H}(t)} \tag{2.29}$$

This estimation of survival function is asymptotically equivalent to Kaplan-Meier estimation of survival. Specifically, when the number of events is smaller than number of exposed individuals its results are very close to the Kaplan-Meier estimator results.

2.2.3 Semi-parametric survival models

Proportional hazards model, proportional odds models and linear transformation models are three of the most commonly used semi-parametric survival models. They are called semi-parametric because they have both the parametric and non-parametric components. Here we will briefly describe the Cox proportional hazards model which makes a parametric assumption about the effect of the covariates on the hazard function, but makes no assumption about the distribution of the hazard function $h(t)$.

Cox Proportional hazards model

In 1972 Cox observed that if we assume each covariate in the study has a multiplicative effect in the hazards function, then there is no need to consider the hazard function to describe how the hazard varies in response to explanatory covariates [40]. This approach is called Cox proportional hazards (CPH) model which assumes that each covariate has a multiplicative effect in the hazards function that is constant over time. But, unlike what we discussed for parametric models such as exponential, Weibull and log-logistic, this model does not assume any parametric form for the hazard function. The hazard function in CPH model is:

$$h(t \mid X) = \lambda_0(t)e^{\beta^T X} \tag{2.30}$$

where $\lambda_0(t)$ is the baseline hazard function which is the hazard at the baseline levels of covariates at time t and X is the matrix of all covariates where the i^{th} column is the covariate vector for the i^{th} subject in the study and β is the vector of coefficients in which each element corresponds to one of the covariates. In our notations in this thesis the vectors are column vectors by default, unless stated otherwise; also, T in β^T is the transpose sign, not the time to event.

The likelihood of the event of interest for subject i at time Y_i can be written as:

$$l_i(\beta, X) = \frac{h(Y_i \mid X_i)}{\sum\limits_{j \in \Omega_i} h(Y_i \mid X_j)} \tag{2.31}$$

where Ω_i is the set of "at-risk" samples with event or follow-up times $\Omega_i = \{j \mid Y_j \geq Y_i\}$ (where Y_i is the event or last follow-up time of subject i). By plugging the equation 2.30 in equation 2.31 and cancelling the $\lambda_0(Y_i)$ from the numerator and denominator, this yields to following likelihood for cox proportional hazards model:

$$l_i(\beta, X) = \frac{e^{\beta^T X_i}}{\sum_{j \in \Omega_i} e^{\beta^T X_j}} \tag{2.32}$$

Assuming the subjects are statistically independent of each other, the joint probability of all observed events is obtained by the following partial likelihood:

$$l(\beta, X) = \prod_{i \in U} l_i(\beta, X) \tag{2.33}$$

where U is the set of all uncensored samples. Taking the logarithm of the equation 2.33 and plugging in the equation 2.32 yields to the following log partial likelihood function of the CPH model:

$$\mathcal{L}(\beta, X) = \sum_{i \in U} \left(\beta^T X_i - log\big(\sum_{j \in \Omega_i} e^{\beta^T X_j} \big) \right) \tag{2.34}$$

that can be maximized over coefficients (β) to produce maximum partial likelihood estimates of the CPH model parameters.

Several approaches are proposed for handling tied events in CPH models. When using the log partial likelihood function in equation 2.34 as is, the Breslow's method is used by default. If there are many tied events in the study that makes handling them important when estimating the model parameters then it might be better to consider other methods such as Efron's method that yield to better results in the existence of tied events [41].

Overall, the Cox proportional hazards model is more popular than parametric models, because its non-parametric estimate of the hazard function offers more flexibility in comparison with most parametric approaches. However, a parametric model, if it is selected appropriately, offers some advantages as it provides higher efficiency by estimating fewer parameters, and enables extrapolating beyond the range of the data under study. Furthermore, the parametric models can yield more relevant in-

terpretations of the model if they are appropriately selected to match the underlying data.

Parametric alternative to the proportional hazards model (Accelerated failure time model)

Accelerated failure time model (AFT model) is a class of parametric models which is used as an alternative to the proportional hazards model [42]. In the proportional hazards model, each covariate has some constant multiplicative effect on hazard, whereas in an AFT model each covariate has some constant multiplicative (acceleration or deceleration) effect on the time to event (survival time). For any AFT model the PDF is:

$$f(t \mid \beta, X) = f_0(te^{-\beta^T X})e^{-\beta^T X} \tag{2.35}$$

where $e^{\beta^T X}$ denotes the joint effect of covariates. This PDF results in the following survival function:

$$S(t \mid \beta, X) = S_0(te^{-\beta^T X}) \tag{2.36}$$

and the hazard function will be:

$$h(t \mid \beta, X) = \lambda_0(te^{-\beta^T X})e^{-\beta^T X} \tag{2.37}$$

We can write the time to event (survival time) in AFT model as follows:

$$log(T) = \beta^T X + \epsilon \tag{2.38}$$

where $\beta^T X$ represents the fixed effects of the covariates and ϵ represents the noise. The noise is distributed as $log(T_0)$ and is independent of the covariates. This

reduces the accelerated failure time model to regression model in the absence of censored data. However, in practice we generally have censored data, so we have to extend the model to handle censoring. The censored observations provide some challenges for estimating the model, if the distribution of T_0 is unusual. Different distributions of noise (ϵ) determines the distribution of baseline survival function $S_0(t)$ of the AFT model, e.g. a Logistic distribution over ϵ implies a Log-logistic distribution for the baseline survival function S_0. Accelerated failure time models are generally named after the distribution of their baseline survival function. The log-logistic distribution is the most commonly used distribution in AFT model, because its CDF and therefore survival function $S(t)$ has a simple closed form, which is important when fitting censored data. Note that based on equation 2.38, whenever ϵ has a logistic distribution, then time to event T will have a log-logistic distribution. Similar to other classes of parametric models explained before, the log likelihood function in equation 2.19 can be maximized over the coefficients of the underlying parametric model's distribution parameters to produce maximum likelihood estimates of the AFT model parameters.

In general, unlike proportional hazards models, the coefficients estimated from AFT models are robust to omitted covariates. They are also less affected by the choice of probability distribution [43, 44]. Furthermore, the results of AFT models are easily interpreted, as they are directly modeling the effect of covariates on survival time, rather than hazard ratio which is harder to explain.

2.3 Survival Convolutional Neural Networks

Deep convolutional neural networks (CNNs) have emerged as an important image analysis tool, and have shattered performance benchmarks in many challenging applications [45]. The ability of CNNs to learn predictive features from raw image

data presents exciting new opportunities in medical imaging [46, 6, 7]. Medical image analysis applications have heavily relied on feature engineering approaches where algorithms are designed to delineate or detect structures of interest, and to measure pre-defined characteristics of these structures that are believed to be predictive. In contrast, the feature learning paradigm of CNNs does not rely on biased a priori definitions of features, and does not require the explicit delineation of anatomic structures which is often confounded by artifacts and variations in image acquisition. While feature learning has become the dominant paradigm in general image analysis tasks, medical applications present unique challenges. Large amounts of labeled data are needed to train CNNs, and medical applications often suffer from data deficits that hurt performance. As "black box" models, CNNs are also difficult or impossible to deconstruct, and so their prediction mechanisms cannot be understood. Despite these challenges, CNNs have been successfully used extensively for medical image classification and segmentation applications [4, 5, 47, 8, 48, 9, 49, 50, 51, 52, 53].

Many important problems in the clinical management of cancer involve time-to-event prediction, including accurate prediction of overall survival, time to progression, and time to metastasis. Despite overwhelming success in other applications, deep learning has not been widely applied to these problems. Survival analysis has often been approached as a machine-learning classification problem by dichotomizing outcomes (e.g. alive vs. deceased at 5 years) [1]. Neural network based Cox regression approaches were explored in early work with low-dimensional clinical datasets, but subsequent analysis found no improvement over basic Cox regression [2]. Deeper networks that are capable of feature learning were recently adapted to optimize Cox proportional hazard likelihood and were shown to have superior performance in predicting overall survival from genomic signatures and clinical datasets [54, 3]. For images, similar convolutional networks were applied to predicting overall survival using lung cancer histology but achieved only marginally better than random prediction

accuracy (0.629 c-index), and were not compared to simple models based on clinical predictors like age, sex, stage or histologic grade [55].

2.4 Learning from Sets

Most machine learning models are designed to map one input vector to one output vector where a function f transforms an input from the vector space $\mathbb{R}^d$ to the discrete space in classification tasks or a continuous space $\mathbb{R}$ in regression tasks. Unlike these models that learn to map fixed dimensional vectors, models that handle sets as their inputs (set-input) or outputs are often overlooked as it is not a trivial task [56, 57]. In these models, the model function f transforms the input that is a set $X = \{x_1, x_2, ..., x_N\}$ to an output that is permutation invariant to the order of objects in the set [58]. Developing models that can learn from unordered sets seems essential in many practical domains. For instance, when modeling the patient outcome from whole slide images (WSI) the relevant information is often present throughout the slide. But we cannot analyze the whole image with a CNN since the relevant content is mostly available at higher magnification levels, in which the dimensions of each slide might be $80K \times 80K$ which is not practical to feed in to the model in whole. Therefore, smaller image patches across the WSI are randomly sampled to feed into the CNN; however, in the basic CNN formulation each image has one label, but in reality it is a collection of the regions all over the WSI that are contributing to the outcome that makes it essential to build a model which learns to map sets of image patches to the patient outcome.

In general, a model has to satisfy two critical requirements in order to handle problems where the input is a set. First, it should be permutation invariant where the output of the model should not change under any permutation of the elements in the input set. Second, it should be able to process input sets of any size [59].

These requirements are not easily satisfied in neural-network-based models. The problem of processing permutation invariant sets by neural nets is a very general and fundamental problem and has a broad application, ranging from learning the structure of point clouds [60], to cosmology [61, 62], to estimation of population statistics [63]. Sets in general and unordered sets specifically, comprise a class of data which are challenging to address with traditional deep learning methods. A simple feed-forward neural network such as a multi-layer perceptron (MLP) [64] would need enormous amounts of data and classical feed-forward neural networks violate both of the aforementioned critical requirements [59]. Recurrent neural networks (RNN) are well known for their power on handling data that are naturally organized as a sequence or ordered sets, but they cannot be really sequence agnostic and their performance for unordered sets suffer as they are are sensitive to the input order. Even if the RNN model is trained on randomly permuted sequences, the ordering of the inputs in RNN does matter and it cannot be totally omitted [65]. [66] moved natural language processing (NLP) away from using sequence dependent RNNs as things that work for sets may work for sequences but not vice versa.

Permutation invariance is an important factor when building models to learn from unordered sets. Permutation equivariance is another concept that is closely related to the permutation invariance. In permutation invariance functions, the output of the function dos not change by permuting the input. In permutation equivariance functions in another hand, the output sequence is permuted in the same manner as the input. In other words, a permutation invariant function $f : X^N \to Y^N$ for any permutation π has the following property

$$f([x_{\pi(1)}, ..., x_{\pi(N)}]) = f([x_1, ..., x_N]) \tag{2.39}$$

while a permutation equivariance function has the following property

$$f([x_{\pi(1)}, ..., x_{\pi(N)}]) = [f_{\pi(1)}(x), ..., f_{\pi(N)}(x)] \tag{2.40}$$

A model that performs pooling over embeddings extracted from each element inside the set is a simple example of a permutation invariant model. Zaheer et al. [67] have demonstrated that all permutation invariant function f can be represented as follows

$$f([x_1, ..., x_N]) = g\Big(\sum_{i=1}^{N} h(x_i)\Big) \tag{2.41}$$

where g and h are any continuous functions and N is the number of instances inside the set. Lee et al. [59] conceptually deconstructs this into an encoder part h which independently acts on each element of a set and a decoder $g\big(\sum(.)\big)$ which aggregates the encoded features and produces the desired output. Most of the available networks for set-input problems follow this encoder-decoder structure. Ravanbakhsh et al. [68] propose a computationally efficient permutation equivariant layer and illustrate that pooling over the output of a permutation equivariant function results in permutation invariance. In other words the model in equation 2.41 remains permutation invariant even if the encoder is a stack of permutation-equivariant layers [67, 59].

Edwards and Storkey [69] and Zaheer et al. [67] propose neural network architectures which meet both critical requirements. In their approach, each element in the set is first independently fed into a feed-forward neural network that takes fixed-size inputs which results in feature-space embeddings for each element in the set. Then, these feature-space embeddings are aggregated using one of the mean, sum or max pooling operations. The final output of their model is obtained by the non-linear processing of the aggregated embedding. This model satisfies both aforementioned requirements, and more importantly, it is proven to be a universal approximator for any set function [67]; because of this property, it is possible to learn a complex map-

ping between input sets and their target outputs. However, it is not clear whether we can obtain a good approximation for complex mappings when using only instance-based feature extractors and simple pooling operations. Also, some of the information regarding interactions between elements inside the set will be discarded, because every individual element in the set is processed independently which might cause some problems that are difficult to address.

Recently, Lee et al. [59] have proposed a neural network architecture to learn from set-inputs. Their model has a self-attention mechanism to process every element in the input set which allows their model to encode pairwise or higher-order interactions between elements in the set. In order for their model to scale for large input sets they have introduced a method to reduce the $\mathcal{O}(n^2)$ computation time of full self-attention to $\mathcal{O}(nm)$ where m is a fixed hyperparameter. The self-attention mechanism in their model is used to aggregate features.

During the recent years, some deep learning architectures have been introduced to explicitly address the unique challenges proposed by sets in natural language processing domain. Iyyer et al. [70] proposed deep averaging networks (DANs) to classify text from an unordered sets of words. A DAN is basically a traditional feed-forward neural network which does an element-wise average of word embeddings in a vector space to generate its main distinguishing feature. Iyyer et al. illustrated how DANs outperform complex network architectures for the same task. However, they did not consider learning word embeddings as part of the architecture, instead they used a set of predefined embeddings. Also, their model was using only averaging to aggregate the word embeddings.

Hill et al. [71] developed a model that learns linear embeddings and does summation instead of averaging over the embeddings. However, Gardner et al. [72] show that linear embeddings are not sufficient for all tasks and indeed are unnecessary with certain pooling operations such as averaging and summing.

Gardner et al. [72] propose convolutional deep averaging networks (CDANs) for classifying and learning feature representations of datasets that contain instances with unordered features. Their model is applicable to variable-size input and is invariant to permutations of the input's order. Their model considers the effects of functions other than averaging such as taking element-wise maximums or sums.

Richard and Gall [73] have developed a neural bag-of-words model that is equivalent to a single-layer-embedding CDAN with average pooling. The embedding in their model is constrained by a softmax output, because each dimension of the embedding in their model is interpreted as a conditional probability of a Gaussian-distributed visual word given the embedded element. The instances in their model are not explicitly treated as sets, but their architecture is still permutation invariant. After pooling in their model pipeline, they have incorporated a specialized layer representing a support vector machine (SVM).

Multiple instance learning (MIL) [74, 75] is another example of problems that deal with set-structured data, where the input is a set of instances and the corresponding output is the label for the entire set (bag). MIL has this assumption of one positive instance inside a bag resulting in a positive bag; this formulation might be appropriate for tasks such as tumor detection or cancer classification based on sets of input image patches, but it might not be an appropriate approach where the goal is to predict a continuous output for the bag based on some continuous values associated with the instances inside the bag e.g. when the goal is to predict the survival of patient based on the sets of their input image patches.

Attention networks simulate selection from a set. They mimic the cognitive attention of human and are used as high-pass filter to capture the relevant context of the input data. Attention mechanism is widely used in computer vision [76] and NLP [77, 78]. Multiplicative (dot-product) attention and multi-head attention are two of the most commonly used attention techniques. Multiplicative attention obtains

the attention by performing a dot-product between vectors. Multi-head attention in another hand, combines several different attention techniques to obtain the overall attention of a network.

Xu et al. [76] have applied the attention mechanism to images in order to generate captions. In this paper they define soft vs hard attention, depending on whether the attention has access to the entire image or only a patch. In the soft attention, the attentions are learned and placed over all patches in the source image [77]. However, in the hard attention, only one patch in the whole image is selected by a hard threshold to attend at each time. Soft attention is differentiable, but it might be computationally expensive specially when the input image is large. Hard attention in another hand is computationally efficient as it needs less calculation. But, hard attention is non-differentiable which might need more complicated methods to train such as reinforcement learning [78].

Chapter 3

Predicting Cancer Outcomes From Histology and Genomics Using Convolutional Networks

In this section, we consider predicting patient outcome from histology images and propose a model that enables end-to-end extraction of features inside each image patch sampled from WSIs and prediction of patient outcome based on these features; we call these image patches High Power Field (HPF). The model we have developed combines a CNN with a Cox model to learn the image patterns that are associated with the patient outcome. We call this model Survival Convolutional Neural Network (SCNN), which provides highly accurate predictions of time-to-event outcomes from histology images. To train this model we are using a negative log partial likelihood of the Cox proportional hazards model that encourages the model to generate risk values that are associated with the patients actual survival times. Our SCNN framework uses an image sampling and risk filtering technique that significantly improves prediction accuracy by mitigating the effects of intratumoral heterogeneity and deficits in the availability of labeled data for training. We use heat map visualization

techniques applied to whole-slide images to show how SCNN learns to recognize important histological structures that neuropathologists use in grading diffuse gliomas and suggest relevance for patterns with prognostic significance that is not currently appreciated. We systematically validate our approaches by predicting overall survival in gliomas using data from The Cancer Genome Atlas (TCGA) Lower-Grade Glioma (LGG) and Glioblastoma (GBM) projects. The current World Health Organization (WHO) standard for classification of gliomas is based on a tiered procedure of molecular subtyping and histologic grading. Therefore, to integrate both histologic and genomic data into a single unified prediction framework, we developed a Genomic Survival Convolutional Neural Network (GSCNN) that enables end-to-end learning and inference from the fusion of genomic and histology data. The GSCNN learns from genomics and histology simultaneously by incorporating genomic data into the fully connected layers of the SCNN. Both data are presented to the network during training, enabling genomic variables to influence the patterns learned by the SCNN by providing molecular subtype information.

3.1 Abstract

Cancer histology reflects underlying molecular processes and disease progression, and contains rich phenotypic information that is predictive of patient outcomes. In this study, we demonstrate a computational approach for learning patient outcomes from digital pathology images using deep learning to combine the power of adaptive machine learning algorithms with traditional survival models. We illustrate how this approach can integrate information from both histology images and genomic biomarkers to predict time-to-event patient outcomes, and demonstrate performance surpassing the current clinical paradigm for predicting the survival of patients diagnosed with glioma. We also provide techniques to visualize the tissue patterns learned by these

deep learning survival models, and establish a framework for addressing intratumoral heterogeneity and training data deficits.

3.2 Introduction

Histology has been an important tool in cancer diagnosis and prognostication for more than a century. Anatomic pathologists evaluate histology for characteristics like nuclear atypia, mitotic activity, cellular density, and tissue architecture, incorporating cytologic details and higher-order patterns to classify and grade lesions. Although prognostication increasingly relies on genomic biomarkers that measure genetic alterations, gene expression, and epigenetic modifications, histology remains an important tool in predicting the future course of a patient's disease. The phenotypic information present in histology reflects the aggregate effects of molecular alterations on cancer cell behavior, and provides a convenient visual readout of disease aggressiveness. However, human assessments of histology are highly subjective and not repeatable, hence computational analysis of histology imaging has received significant attention. Aided by advances in slide scanning microscopes and computing, a number of image analysis algorithms have been developed for grading [79, 80, 81, 82], classification [83, 84, 85, 86, 4, 87], and prediction of future metastasis [5] in multiple cancer types.

Deep convolutional neural networks (CNNs) have emerged as an important image analysis tool, and have shattered performance benchmarks in many challenging applications [45]. The ability of CNNs to learn predictive features from raw image data is a paradigm shift that presents exciting new opportunities in medical imaging [46, 6, 7]. Medical image analysis applications have heavily relied on feature engineering approaches where algorithms are designed to delineate or detect structures of interest, and to measure pre-defined characteristics of these structures that are believed to be predictive. In contrast, the feature learning paradigm of CNNs does not rely on

biased a priori definitions of features, and does not require the explicit delineation of anatomic structures which is often confounded by artifacts and variations in image acquisition. While feature learning has become the dominant paradigm in general image analysis tasks, medical applications present unique challenges. Large amounts of labeled data are needed to train CNNs, and medical applications often suffer from data deficits that hurt performance. As "black box" models, CNNs are also difficult or impossible to deconstruct, and so their prediction mechanisms cannot be understood. Despite these challenges, CNNs have been successfully used extensively for medical image classification and segmentation applications [4, 5, 47, 8, 48, 9, 49, 50, 51, 52, 53].

Many important problems in the clinical management of cancer involve time-to-event prediction, including accurate prediction of overall survival, time to progression, and time to metastasis. Despite overwhelming success in other applications, deep learning has not been widely applied to these problems. Survival analysis has often been approached as a machine-learning classification problem by dichotomizing outcomes (e.g. alive vs. deceased at 5 years) [10]. Neural network based Cox regression approaches were explored in early work with low-dimensional clinical datasets, but subsequent analysis found no improvement over basic Cox regression [11]. Deeper networks that are capable of feature learning were recently adapted to optimize Cox proportional hazard likelihood and were shown to have superior performance in predicting overall survival from high-dimensional genomic signatures [1], and with low-dimensional clinical datasets [2]. For images, similar convolutional networks were applied to predicting overall survival using lung cancer histology but achieved only marginally better than random prediction accuracy (0.629 c-index).

In this section, we present a convolutional network based approach called Survival Convolutional Neural Networks (SCNN) that can predict overall survival and other time-to-event outcomes from histology images with accuracy that equals or surpasses clinical paradigms based on genomic biomarkers and manual histologic grading. We

provide a new training and prediction framework based on image resampling that significantly improves prediction accuracy by mitigating the effects of intratumoral heterogeneity and deficits in the amounts of labeled training data. We also illustrate how genomic and histology imaging data can be integrated into a single SCNN prediction model to significantly improve prognostic accuracy. Finally, we show how the prediction mechanisms of SCNN models can be interpreted using whole-slide risk heatmaps that visualize the risks associated with various regions in a histologic specimen. We systematically validate these approaches by building models to predict overall survival in gliomas using data from The Cancer Genome Atlas Lower Grade Glioma (LGG) and Glioblastoma (GBM) projects.

3.3 Learning patient outcomes with deep survival convolutional neural networks

The SCNN model architecture is depicted in Figure 3.1 and Figure 3.2 illustrates the detailed architecture of the CNN component in the SCNN model. Hematoxylin and eosin tissue sections are first digitized to large whole-slide-images. These images are reviewed using a web-based platform to identify regions-of-interest (ROIs) with representative histologic characteristics [88]. High-power fields (HPFs) from these ROIs are then used to train a deep convolutional network that is seamlessly integrated with a Cox proportional hazards model to predict patient outcomes. The network is composed of interconnected layers of image processing operations and nonlinear functions that sequentially transform the HPF image into highly-predictive prognostic features. Convolutional layers first extract visual features from the field image at multiple scales using convolutional kernels and pooling operations. These image-derived features feed into fully-connected layers that perform additional transformations, and then a final Cox model layer outputs a prediction of patient risk. The interconnection weights

and convolutional kernels are trained by comparing risk predicted by the network with survival or other time-to-event outcomes using a backpropagation technique to optimize the statistical likelihood of the network.

To improve the performance of SCNN models, we developed resampling techniques to address the limited availability of training samples and intratumoral heterogeneity (see Figure 3.3). For training, new HPFs are randomly sampled from each ROI at the start of each training iteration, providing the SCNN model with a fresh look at each patient's histology and capturing heterogeneity within the ROI. The SCNN is also trained using multiple such HPFs for each patient (one for each region) to further account for intratumoral heterogeneity across ROIs. For predicting the risk of a new patient with unknown survival, we integrate information from many HPFs by randomly sampling multiple fields in each ROI and using averaging and ranking procedures to create a robust patient-level prediction that rejects outlying risk predictions. These resampling procedures are described in detail in Methods.

3.4 Methods

3.4.1 Data and image curation

Whole slide images, clinical and genomic data were obtained from The Cancer Genome Atlas via the Genomic Data Commons (`https://gdc.cancer.gov/`). Images of diagnostic hematoxylin and eosin stained formalin-fixed paraffin-embedded sections from the Brain Lower Grade Glioma (LGG) and the Glioblastoma (GBM) cohorts were reviewed to remove images containing tissue processing artifacts including bubbles, section folds, pen markings and poor staining. Representative regions of interest containing primarily tumor nuclei were manually identified for each slide that passed quality control. In the case of grade IV disease, some regions include microvascular proliferation as this feature was exhibited throughout tumor regions. Regions con-

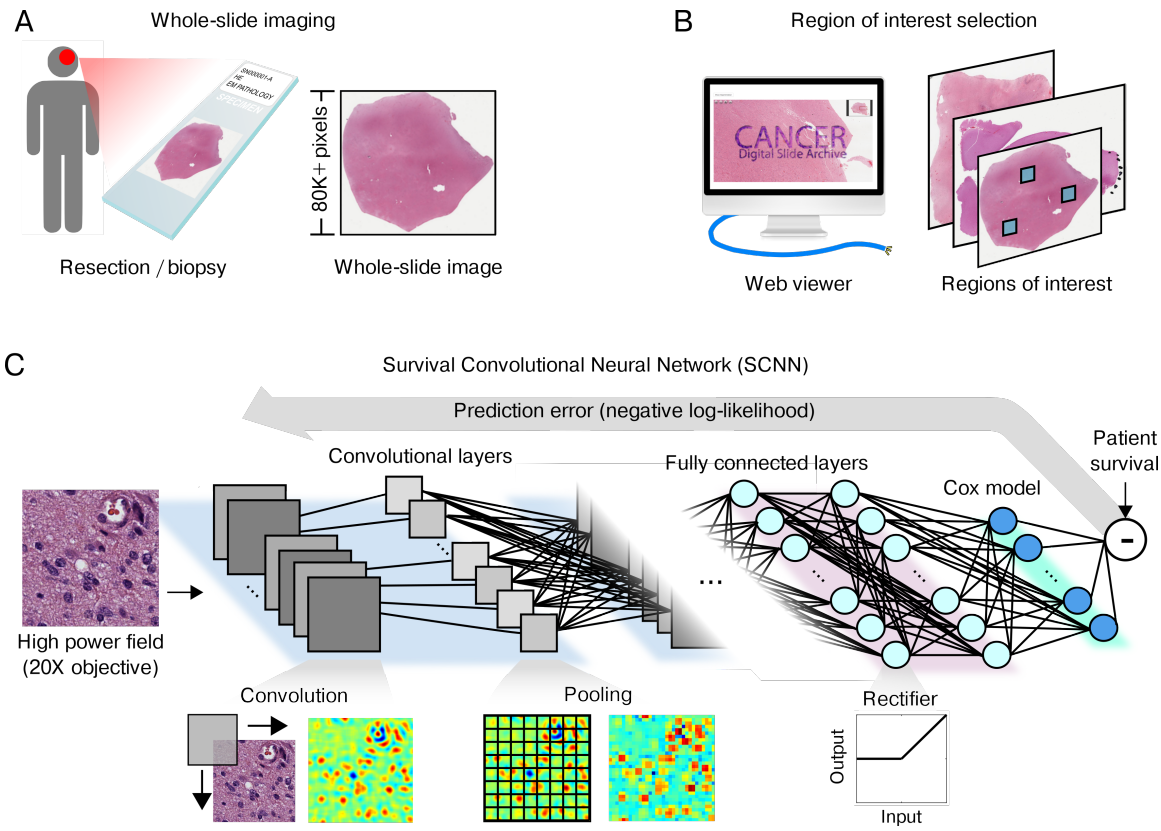

Figure 3.1: The SCNN model. The SCNN combines deep learning CNNs with traditional survival models to learn survival-related patterns from histology images. (A) Large whole-slide images are generated by digitizing H&E-stained glass slides. (B) A web-based viewer is used to manually identify representative ROIs in the image. (C) HPFs are sampled from these regions and used to train a neural network to predict patient survival. The SCNN consists of (i) convolutional layers that learn visual patterns related to survival using convolution and pooling operations, (ii) fully connected layers that provide additional nonlinear transformations of extracted image features, and (iii) a Cox proportional hazards layer that models time-to-event data, like overall survival or time to progression. Predictions are compared with patient outcomes to adaptively train the network weights that interconnect the layers.

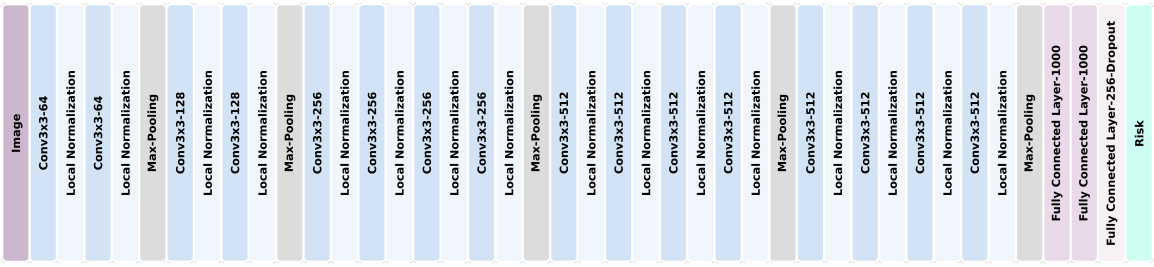

Figure 3.2: Detailed diagram of the CNN component in SCNN architecture. The architecture is a variation of the VGG19 network and combines convolutional, maximum pooling, local normalization, and fully connected layers.

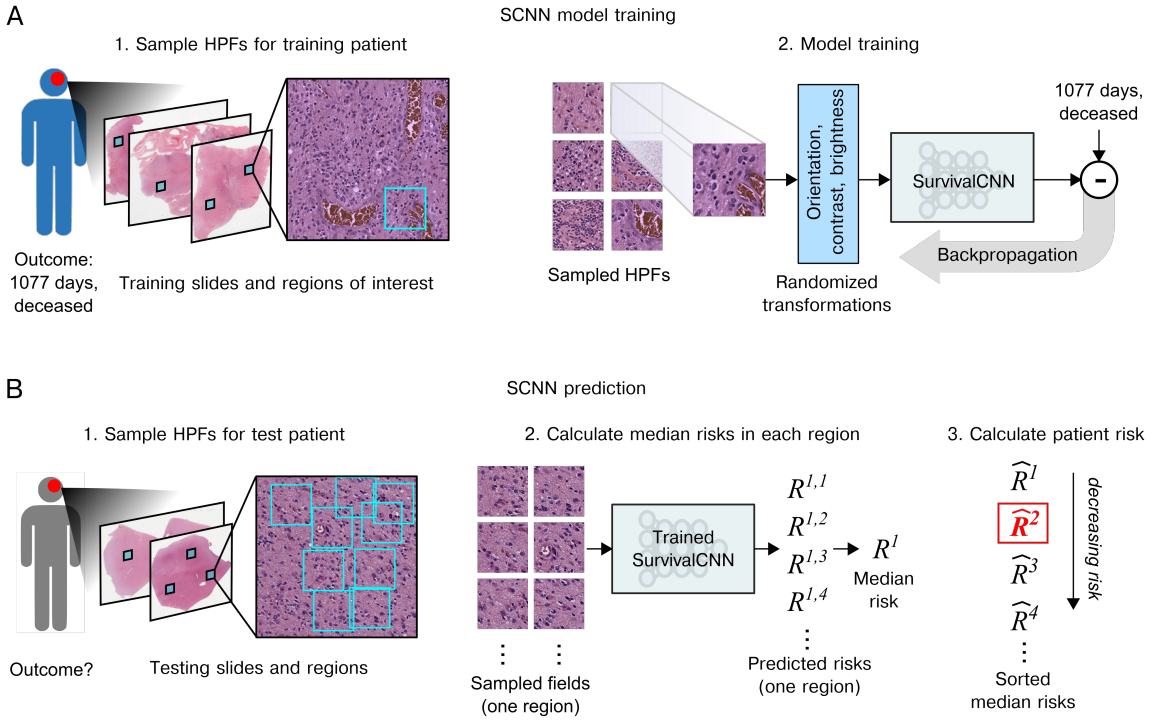

Figure 3.3: SCNN uses image re-sampling to improving the robustness of training and prediction. (A) During training, a single 256 × 256 pixel high-power field is sampled from each region, producing multiple HPFs per patient. Each HPF is subjected to a series of random transformations that simulate image acquisition variations, and is then used as an independent sample to update the network weights. New HPFs are re-sampled at each training epoch (one training pass through all patients). (B) When predicting the outcome of a newly diagnosed patient, 9 HPFs are sampled from each region of interest and a risk is predicted for each field. The median risk in each region is calculated, the median risks are sorted, and the second highest risk is selected as the risk of the patient. This process was designed to deal with tissue heterogeneity by emulating the process of histologic evaluation by a pathologist, where prognostication is based on the most malignant regions within a heterogeneous sample.

taining geographic necrosis were excluded. A total of 1061 whole-slide images from 769 unique patients were analyzed.

Regions of interest images (1024×1024 pixels) were cropped at 20X objective magnification using OpenSlide and color-normalized to a gold-standard H&E calibration image to improve consistency of color characteristics across slides. High-power fields (HPFs) at 256×256 pixels were sampled from these regions and used for training and testing as described below.

3.4.2 Network architecture and training procedures

The survival convolutional neural network combines elements of a variation of VGG19 convolutional network architecture with a Cox proportional hazard model to predict time-to-event data from images (see Figure 3.2) [14]. Image feature extraction is achieved by four groups of convolutional layers: 1) The first group contains two convolutional layers with 64 3×3 kernels, interleaved with local normalization layers, then followed with a single max-pooling layer 2) The second group contains two convolutional layers (128 3×3 kernels) interleaved with two local normalization layers, followed by a single max pooling layer 3) The third group interleaves four convolutional layers (256 3×3 kernels) with four local normalization layers, followed by a single max pooling layer 4) The fourth group contains interleaves of eight convolutional (512 3×3 kernels) and eight local normalization layers, with an intermediate and a terminal max pooling layer. These four groups are followed by a sequence of 3 fully connected layers containing 1000, 1000, and 256 nodes respectively.

The terminal fully connected layer outputs a prediction of risk $R = \beta^T X$ associated with the input image, where $\beta \in \mathbb{R}^{256 \times 1}$ are the terminal layer weights and $X \in \mathbb{R}^{256 \times 1}$ are the inputs to this layer. To provide an error signal for backpropagation, these risks are input to a Cox proportional hazards layer to calculate the negative partial log-likelihood

$$L(\beta, X) = -\sum_{i \in U} \left(\beta^T X_i - log(\sum_{j \in \Omega_i} e^{\beta^T X_j}) \right) \tag{3.1}$$

Where $\beta^T X_i$ is the risk associated with HPF i, U is the set of all uncensored samples, and Ω_i is the set of "at-risk" samples with event or follow-up times $\Omega_i = \{j \mid Y_j \geq Y_i\}$ (where Y_i is the event or last follow-up time of subject i).

The adagrad algorithm was used to minimize the negative partial log-likelihood via backpropagation to optimize model weights, biases and convolutional kernels [89]. Parameters to adagrad include the initial accumulator value = 0.1, initial learning rate = 0.001, and an exponential learning rate decay factor = 0.1. Model weights were initialized using the variance scaling method [90], and a weight decay was applied to the fully connected layers during training (decay rate = 4e-4). Models were trained for 100 epochs (1 epoch is one complete cycle through all training samples) using mini-batches consisting of 14 HPFs each. Each mini-batch produces a model update, resulting in multiple updates per epoch. Calculation of the Cox partial likelihood requires access to the predicted risks of all samples, which are not available within any single mini-batch, and so Cox likelihood was calculated locally within each mini-batch to perform updates (U and Ω_i were restricted to samples within each mini-batch). Local likelihood calculation can be very sensitive to how samples are assigned to mini-batches, and so we randomize the mini-batch sample assignments at the beginning of each epoch to improve robustness. Mild regularization was applied during training by randomly dropping out 5% of weights in the last fully connected layer in each mini-batch during training to mitigate overfitting.

3.4.3 Training resampling

Each patient has possibly multiple slides, and multiple regions within each slide that can be used to sample HPFs. During training, a single HPF was sampled from each

region, and these HPFs were treated as semi-independent training samples. Each HPF was paired with patient outcome for training, duplicating outcomes for patients containing multiple regions / HPFs. The HPFs are resampled at the beginning of each training epoch to generate an entirely new set of HPFs. Randomized transforms were also applied to these HPFs to improve robustness to tissue orientation and color variations. Since the visual patterns in tissues can often be anisotropic, we randomly apply a mirror transform to each HPF. We also generate random transformations of contrast and brightness using the "random_contrast" and "random_brightness" TensorFlow transformations to modify the HPF and simulate color variations. These resampling and transformation procedures, along with the use of multiple HPFs for each patient, has the effect of augmenting the effective size of the labeled training data. Similar approaches for training data augmentation have demonstrated considerable improvements in general imaging applications [91]. The resampling procedure during training is illustrated in Figure 3.3A.

3.4.4 Testing resampling and model averaging

Resampling was also performed to increase the robustness and stability of predictions: 1) 9 high-power fields are first sampled from each region j corresponding to patient m 2) The risk of the k^{th} HPF in region j for patient m, denoted $R_m^{j,k}$, is then calculated using the trained SCNN model 3) The median risk $R_m^j = median_k\{R_m^{j,k}\}$ is calculated for region j using the aforementioned HPFs to reject outlying risks 4) These median risks are then sorted from highest to lowest $\hat{R}_m^1 > \hat{R}_m^2 > \hat{R}_m^3 ...$, where the superscript index now corresponds to the risk rank 5) The risk prediction for patient m is then selected as the second highest risk $R_m^* = \hat{R}_m^2$. This filtering procedure was designed to emulate how a pathologist integrates information from multiple areas within a slide, determining prognosis based on the region associated with the worst prognosis. Selection of the second highest risk (as opposed to the highest risk)

introduces robustness to outliers or high risks that may occur due to some imaging or tissue-processing artifact. Since the accuracy of our models can vary significantly from one epoch to another, largely due to the training resampling and randomized mini-batch assignments, a model averaging technique was used to reduce prediction variance. To obtain final risk predictions for the testing patients that are stable, we perform model averaging using the models from epochs 96-100 to smooth variations across epochs and increase stability. Formally, the model-averaged risk for patient m is calculated as

$$\overline{R_m^*} = \frac{1}{5} \sum_{\gamma=96}^{100} R_{m(\gamma)}^* \tag{3.2}$$

where $R_{m(\gamma)}^*$ denotes the predicted risk for patient m in training epoch γ.

3.4.5 Validation procedures

Patients were randomly assigned to non-overlapping training (80%) and validation (20%) sets that were used to train models and evaluate their performance. If a patient was assigned to training, then all slides corresponding to that patient were assigned to the training set and likewise for the testing set. This ensures that no data from any one patient is represented in both training and testing sets to avoid overfitting and optimistic estimates of generalization accuracy. We repeated the randomized assignment of patients training/testing sets 15 times, and used each of these training/testing sets to train and evaluate a model. The same training/testing assignments were used in each model (SCNN, GSCNN, baseline) for comparability. Prediction accuracy was measured using Harrell's c-index (CI) to measure the concordance between predicted risk and actual survival for testing samples [92].

3.4.6 Statistical analyses

C-indexes generated by Monte Carlo cross-validation were performed using the Wilcoxon signed rank test. This paired test was chosen because each method was evaluated using identical training/testing sets. Comparisons of SCNN risk values across grade were performed using the Wilcoxon rank-sum test. Cox univariable and multivariable regression analyses were performed using predicted SCNN risk values for all training and validation samples in the randomized training/validation set 1. Analysis of the correlation of grade, molecular subtype, and SCNN risk predictions were performed by pooling predicted risks for validation samples across all experiments. SCNN risks were normalized within each experiment by z-score prior to pooling. Grade analysis was performed by determining "digital" grade thresholds for SCNN risks in each subtype. Thresholds were objectively selected to match the proportions of samples in each histologic grade in each subtype. Statistical analysis of Kaplan Meier plots was performed using the log-rank test.

3.4.7 Hardware and software

Prediction models were trained using TensorFlow (v0.12.0) on servers equipped with dual Intel(R) Xeon(R) CPU E5-2630L v2 @ 2.40GHz CPUs, 128GB RAM, and dual NVIDIA K80 graphics cards. Image data was extracted from Aperio .svs whole-slide image formats using OpenSlide (`http://openslide.org/`). Basic image analysis operations were performed using HistomicsTK (`https://github.com/DigitalSlideArchive/HistomicsTK`), a Python package for histology image analysis.

Table 3.1: Summary of dataset clinical features

Characteristic	Total, n=769	Molecular subtype		
		Astrocytoma IDH WT, n=335 (48%)	Astrocytoma IDH mutant, n=220 (32%)	Oligodendroglioma, n=142 (20%)
WHO histologic grade (%)				
II	181 (25)	14 (4)	96 (48)	69 (53)
III	205 (28)	57 (17)	88 (44)	60 (46)
IV	350 (47)	262 (79)	17 (8)	1 (1)[a]
Age at diagnosis, y				
Range	10-88	10-88	14-73	17-75
Median	51 ± 15.5	58 ± 14.0	36 ± 11.3	45.5 ± 12.7
Sex, female (%)	308 (42)	137 (41)	86 (43)	54 (42)
Median survival, y				
Grade II	-	2.1	8.2	14.2
Grade III	-	1.7	6.3	9.7
Grade IV	-	1.2	3.0	N/A[b]

[a]Grade IV is not defined for oligodendroglioma. This sample was initially classified as an astrocytoma under the older histological classification paradigm (before molecular subtyping).
[b]N/A, not applicable.

3.5 Assessing the prognostic accuracy of SCNN

The prognostic accuracy of SCNN models was assessed using Monte Carlo cross-validation. Using the Digital Slide Archive, we first identified ROIs in 1061 whole-slide images of hematoxylin and eosin stained sections obtained from 769 gliomas from the TCGA dataset. This dataset comprises lower-grade gliomas (WHO grades II and III) and glioblastomas (WHO grade IV), contains both astrocytomas and oligodendrogliomas, and has overall survivals ranging from less than 1 to 14 y or more. A summary of demographics, grades, survival, and molecular subtypes for this cohort is presented in Table 3.1. Patients were assigned to either training (80%) or validation (20%) to form 15 randomized datasets to evaluate the prognostic accuracy of methods. Accuracy was measured using Harrell's c-index, a non-parametric statistic that measures concordance between predicted risks and actual survival [92]. A c-index of 1 indicates perfect concordance between predicted risk and overall survival, and a c-index of 0.5 corresponds to random concordance.

SCNN networks demonstrated substantial prognostic power, achieving a median c-index of 0.754 (see Figure 3.4B). For comparison, we also measured the accuracy

of baseline models generated using the genomic biomarkers and manual histologic grading used in the World Health Organization (WHO) classification (see Figure 3.4A). The WHO assigns gliomas to three genomic subtypes defined by mutations in isocitrate dehydrogenase (IDH1 / IDH2) and co-deletion of chromosomes 1p and 19q. Within these molecular subtypes, gliomas are further assigned a histologic grade based on criteria that vary depending on cell of origin (either astrocytic or oligodendroglial). Subtypes with an astrocytic lineage are split by IDH mutation status, and the combination of 1p/19q codeletion and IDH mutation defines an oligodendroglioma. These lineages have histologic differences; however, histologic evaluation is not a reliable predictor of molecular subtype [93]. Histologic criteria used for grading include mitotic activity, nuclear atypia, the presence of necrosis, and the characteristics of microvascular structures. WHO baseline models based on molecular subtype and manual histologic grade had a median c-index of 0.774, outperforming SCNN networks based on machine-learning from histology images (Wilcoxon signed-rank p=2.61e-3). The manual histologic grade baseline models had a median c-index of 0.745, with performance similar to SCNN models (p=0.307). The molecular subtype baseline models had a median c-index of 0.746, and were significantly outperformed by the SCNN models (p=4.68e-2).

We also evaluated the benefits of our resampling methods in improving the performance of SCNN models. Repeating the SCNN experiments without resampling techniques reduced the median c-index to 0.696, significantly worse than for SCNN models where resampling was used (p=6.55e-4).

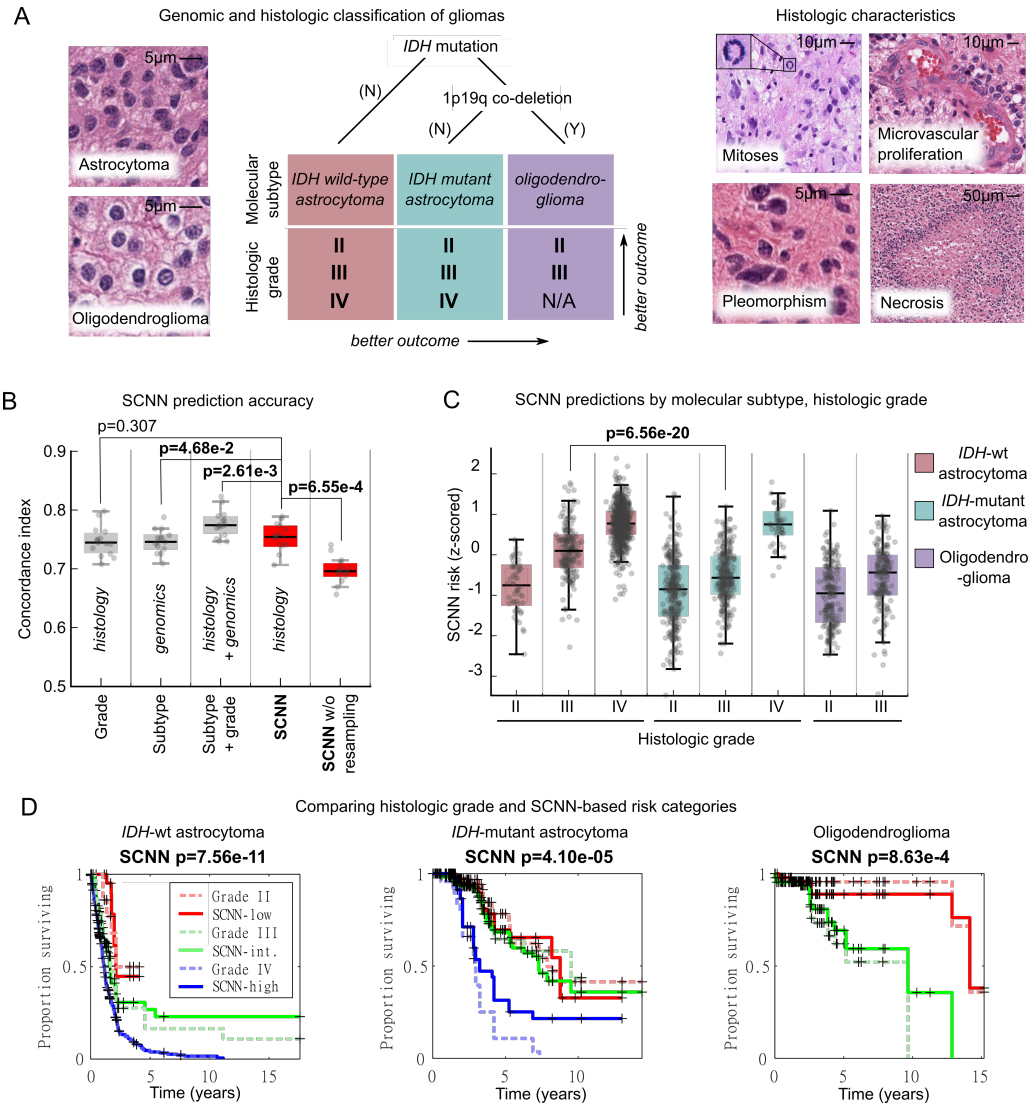

Figure 3.4: Prognostication criteria for diffuse gliomas. (A) Prognosis in the diffuse gliomas is determined by genomic classification and manual histologic grading. Diffuse gliomas are first classified into one of three molecular subtypes based on IDH1/IDH2 mutations and the codeletion of chromosomes 1p and 19q. Grade is then determined within each subtype using histologic characteristics. (B) Comparison of the prognostic accuracy of SCNN models with that of baseline models based on molecular subtype or molecular subtype and histologic grade. Models were evaluated over 15 independent training/testing sets with randomized patient assignments and with/without training and testing sampling. (C) The risks predicted by the SCNN models correlate with both histologic grade and molecular subtype, decreasing with grade and generally trending with the clinical aggressiveness of genomic subtypes. (D) Kaplan–Meier plots comparing manual histologic grading and SCNN predictions. Risk categories (low, intermediate, high) were generated by thresholding SCNN risks. N/A, not applicable.

3.6 SCNN predictions correlate with genomic subtypes and manual histologic grade

To further investigate the relationship between SCNN predictions and the WHO paradigm, we visualized how risks predicted by SCNN networks are distributed across molecular subtype and histologic grade (see Figure 3.4C). SCNN predictions were highly correlated with both subtype and grade, and were consistent with expected patient outcomes in each category. Firstly, within each molecular subtype, the risks predicted by SCNN increase with histologic grade. Secondly, predicted risks are consistent with the published expected overall survivals associated with genomic subtypes [93]. Astrocytomas with wild-type IDH are highly aggressive with a median survival of 18 months, and the collective risks for these patients is higher than for patients from other subtypes. Astrocytomas having IDH mutations are another subtype with considerably better overall survival ranging from 3-8 years, and the predicted risks for patients in this subtype are more moderate. Notably, in this subtype, SCNN risks are not well separated for grades II and III, consistent with reports of histologic grade being an inadequate predictor of outcome in this subtype [94]. Gliomas with mutations in IDH and co-deletion of chromosomes 1p/19q are described as oligodendroglioma, have a distinct differentiation, and have the lowest overall predicted risks consistent with survivals of 10+ years for this subtype. Finally, we noted a significant difference in predicted risks for grade III gliomas in the astrocytic subtypes (rank-sum p=6.56e-20). These subtypes share an astrocytic lineage, are graded using identical histologic criteria, and are not known to have any distinguishing histological characteristics. Despite the inability of pathologists to discriminate between these subtypes using histology, SCNN can predict risks that are consistent with worse outcomes for grade III IDH wild-type astrocytomas (median survival 1.7 years) compared to grade III IDH mutant astrocytomas (median survival 6.3 years).

To illustrate how SCNN risks can be used to assign a categorical "digital" grade, we performed a Kaplan Meier analysis to stratify patients based on SCNN risks (see Figure 3.4D). Risk thresholds defining digital grades were established for each molecular subtype separately. The proportions of each histologic grade in each subtype were used as a guideline to set thresholds on SCNN risks. In each subtype, the digital grades capture survival differences in a manner analogous to manual histologic grading. A comparison to stratification by histologic grade is presented in Figure 3.4D. Based on these results we observed that digital and manual histologic grades have similar prognostic power in IDH wild-type astrocytomas (log-rank p=1.23e-12 versus p=7.56e-11 respectively). In IDH mutant astrocytomas, both digital and manual histologic grades have difficulty separating Kaplan Meier curves for grades II and III, yet both clearly distinguish grade IV as being associated with worse outcomes. Discrimination for oligodendroglioma survival is also similar between digital and manual histologic grades (log-rank p=9.73e-7 versus p=8.63e-4 respectively).

3.7 Improving prognostic accuracy by integrating genomic biomarkers

To leverage both histologic and genomic data in predicting survival, we developed a hybrid Genomic-SCNN model (GSCNN). The GSCNN learns prognosis from both genomics and histology by incorporating genomic variables into the fully-connected layers of the SCNN to improve prognostic accuracy (see Figure 3.5). This configuration enables the genomic variables to influence the patterns learned from histology by providing information on molecular subtype near the terminal network layers.

We repeated our experiments using GSCNN models with histology images, IDH mutation status, and 1p/19q co-deletion as inputs, and found that the median c-index improved to 0.801. The addition of genomic variables improved the performance by

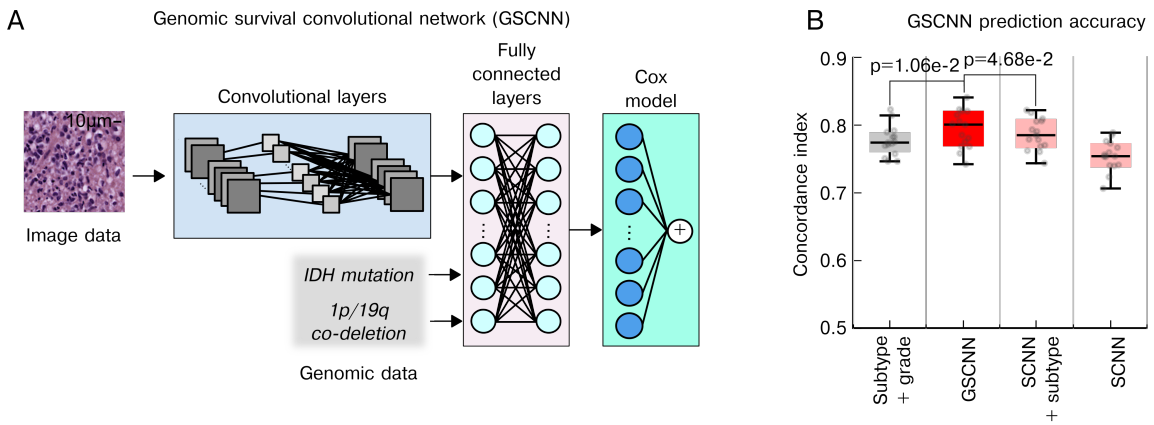

Figure 3.5: Genomic-SCNN models integrate genomic and imaging data for improved performance. (A) A hybrid architecture was developed to combine histology image and genomic data to make integrated predictions of patient survival. These models incorporate genomic variables as inputs to their fully-connected layers. Here, we show the incorporation of genomic variables for gliomas, however any number of genomic or proteomic measurements can be similarly used. (B) The GSCNN models significantly outperform SCNN models, as well as the WHO paradigm based on genomic subtype and histologic grading.

5% on average over SCNN models that are trained on histology images alone. The GSCNN models also significantly outperform the WHO baseline subtype-grade model trained on equivalent data (signed-rank p=1.06e-2). We compared GSCNN with a more superficial integration approach, where an SCNN model was first trained using histology images, and then, the risks from this model were combined with IDH and 1p/19q variables in a simple three-variable Cox model as illustrated in Figure 3.6. This superficial approach did not perform as well as GSCNN, with a median c-index of 0.785 (signed-rank p=4.68e-2), illustrated as "SCNN + subtype" in Figure 3.5B.

To evaluate the independent prognostic power of risks predicted by SCNN and GSCNN, we performed a multivariable Cox regression analysis (see Table 3.2). In a multivariable regression that included SCNN risks, subtype, grade, age, and sex, SCNN risks were prognostic when correcting for all other features including manual grade and molecular subtype (p=2.71e-12). Manual histologic grade was not significant in this regression analysis. We also performed a similar multivariable regression

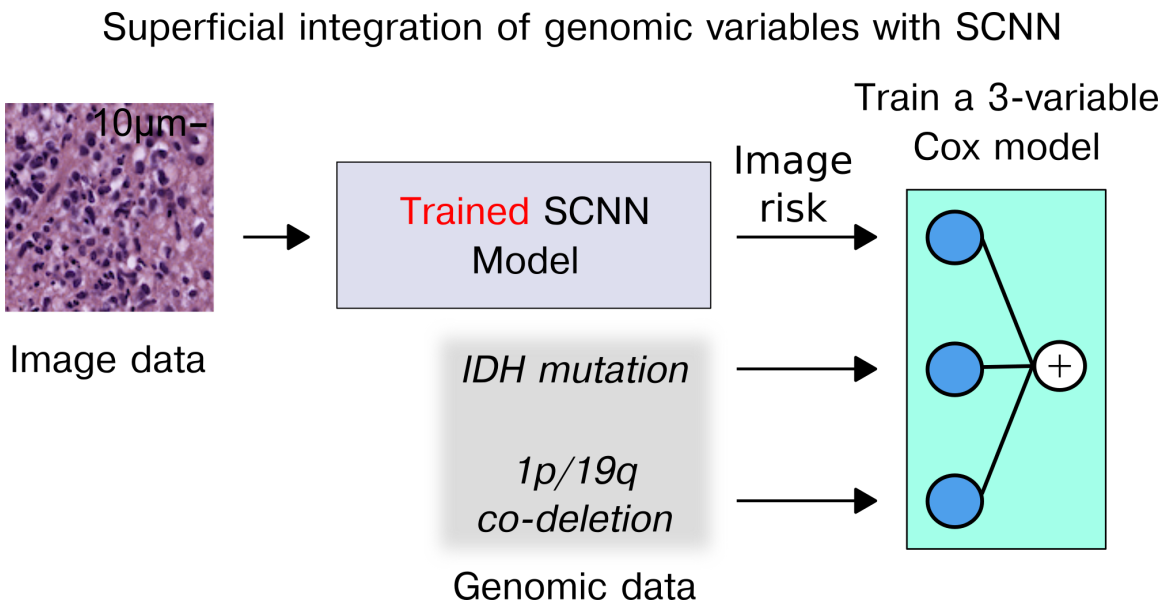

Figure 3.6: Superficial integration of histology and genomic biomarkers. We evaluated the benefit of including genomic biomarkers in GSCNN training by evaluating the accuracy of a more superficial integration approach. We first trained an SCNN using histology images alone (step 1). After this training, we combined the risks produced by this SCNN with genomic variables using a simple linear Cox regression model. This Cox model was trained using the training samples and was evaluated on testing samples to measure prediction accuracy.

Table 3.2: Hazard ratios for univariable and multivariable Cox regression models.

	Single Variable				Multivariable (SCNN)			Multivariable (GSCNN)		
Variable	c-index	Hazard ratio	95% CI	*P value*	Hazard ratio	95% CI	*P value*	Hazard ratio	95% CI	*P value*
SCNN	**0.741**	**7.15**	**5.64, 9.07**	**2.08e-61**[a]	**3.05**	**2.22, 4.19**	**2.71e-12**	-	-	-
GSCNN	0.781	12.60	9.34, 17.0	**3.08e-64**	-	-	-	8.83	4.66, 16.74	**9.69e-12**
IDH WT astrocytoma	0.726	9.21	6.88, 12.34	**3.48e-52**	4.73	2.57, 8.70	**3.49e-7**	0.97	0.43, 2.17	0.93
IDH mutant astrocytoma	-	0.23	0.170, 0.324	**2.70e-19**	2.35	1.27, 4.34	**5.36e-3**	1.67	0.90, 3.12	0.10
Histologic grade IV	0.721	7.25	5.58, 9.43	**2.68e-51**	1.52	0.839, 2.743	0.159	1.98	1.11, 3.51	**0.017**
Histologic grade III	-	0.44	0.332, 0.591	**1.66e-08**	1.57	0.934, 2.638	0.0820	1.78	1.07, 2.97	**0.024**
Age	0.744	1.77	1.63, 1.93	**2.52e-42**	1.33	1.20, 1.47	**9.57e-9**	1.34	1.22, 1.48	**9.30e-10**
Sex, female	0.552	0.89	0.706, 1.112	0.29	0.85	0.67, 1.08	0.168	0.86	0.68, 1.08	0.18

[a]Bold indicates statistical significance (p<5e-2)

with GSCNN risks, and found GSCNN to be significant (p=9.69e-12). In this multivariable regression, molecular subtype was not significant, and histologic grade was only marginally significant. We also used Kaplan–Meier analysis to compare risk categories generated from SCNN and GSCNN (Figure 3.7). Survival curves for SCNN and GSCNN were very similar when evaluated on the entire cohort. In contrast, their abilities to discriminate survival within molecular subtypes were notably different.

3.8 Visualizing prognosis with SCNN heatmaps

Deep learning networks are often criticized for being "black-box" approaches that do not reveal insights into their prediction mechanisms. To investigate the visual patterns SCNN models learn from histology images, we created risk heatmap overlays to visualize the risks associated with different regions in a whole-slide image. These heatmaps were generated by predicting risk for each non-overlapping HPF in a whole-slide image. The predicted risks of each HPF were used to generate a color-coded transparent overlay that represents the SCNN risk predictions across the entire slide.

A selection of risk heatmaps from three patients is presented in Figure 3.8, with inlays demonstrating how SCNNs associate risk with important pathologic phenomena. For TCGA-DB-5273 (Grade III, IDH mutant astrocytoma), the SCNN heatmap clearly highlights regions of early microvascular proliferation, an advanced form of angiogenesis that is a hallmark of malignant progression, as being associated with

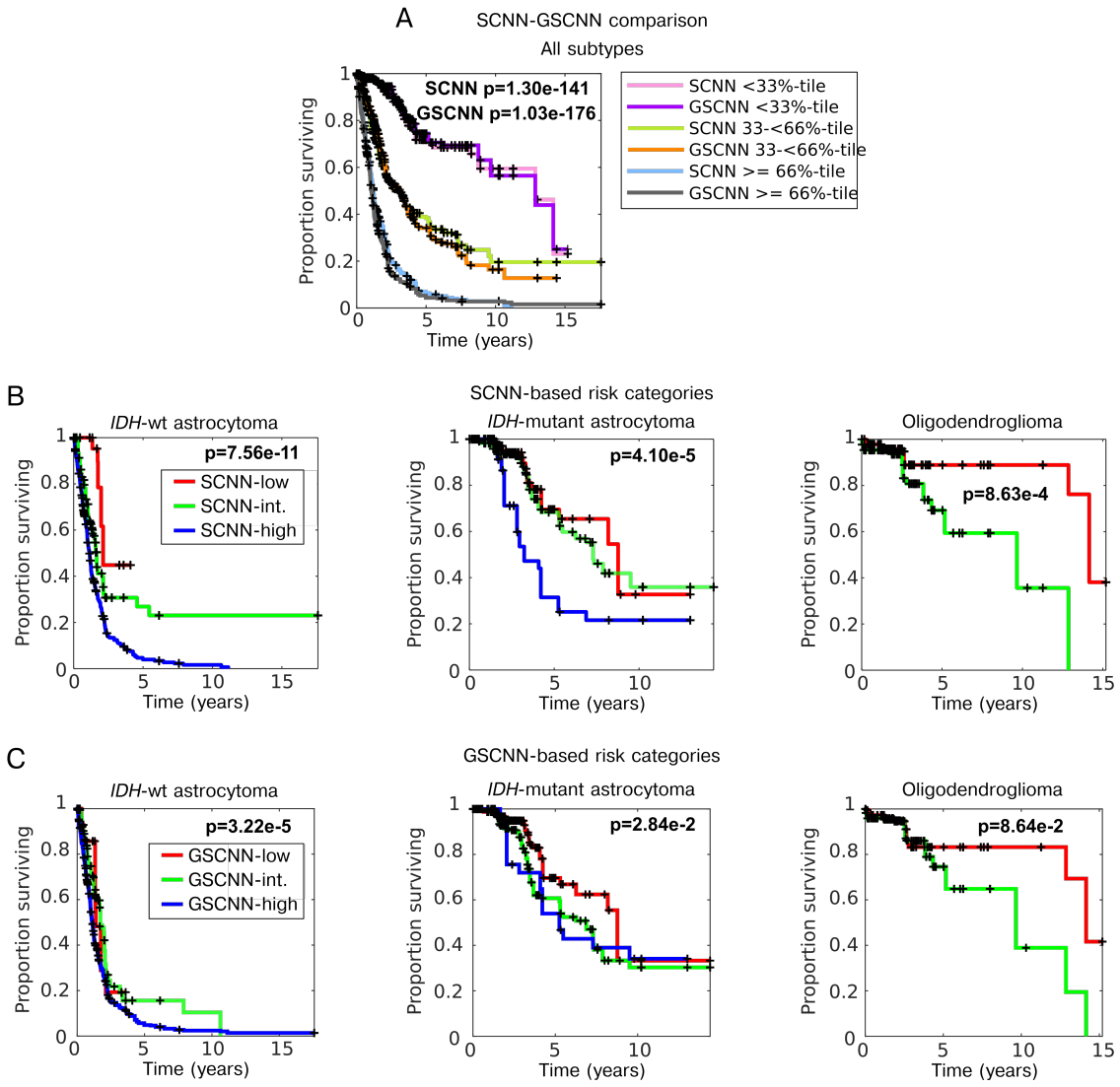

Figure 3.7: Kaplan–Meier analysis of SCNN and GSCNN. (A) We compared the overall prediction power of SCNN and GSCNN in the samples from all subtypes using tertiles. Although the log rank test for GSCNN indicates slightly better separation of survival curves, visually, the curves for SCNN and GSCNN are remarkably similar. (B) SCNN risk categories perform well when examined within each molecular subtype. SCNN is not able to assign patients to these subtypes reliably, however, since its predictions are based entirely on histology. (C) GSCNN risk categories overlap significantly when examined in each molecular subtype. Although some separation is apparent, most of the predictive power of GSCNN comes from its ability to reliably assign patients to molecular subtypes.

high risk. Risk in this heatmap also increases with cellularity, heterogeneity in nuclear shape and size (pleomorphism), and the presence of abnormal microvascular structures. Regions in TCGA-S9-A7J0 have varying extents of tumor infiltration, ranging from normal brain, to sparsely infiltrated adjacent normal regions exhibiting satellitosis (where neoplastic cells cluster around neurons), to moderately and highly infiltrated regions. This heatmap correctly associates the lowest risks to normal brain regions, and can distinguish normal brain from adjacent regions that are sparsely infiltrated. Interestingly, higher risks are assigned to sparsely infiltrated regions (region 1, upper panel) than to regions containing relatively more tumor infiltration (region 2). We observed a similar pattern in TCGA-TM-A84G, where edematous regions (region 1, lower panel) adjacent to moderately cellular tumor regions (region 1, upper panel) are also assigned higher risks. These latter examples provide novel risk features embedded within histologic sections that have been previously unrecognized and could inform and improve pathology practice.

3.9 Discussion

We developed a deep learning algorithm for learning survival directly from histological images, and systematically evaluate its prognostic accuracy in the context of the current clinical standard based on genomic classification and histologic grading. In contrast to a previous study that achieved only very marginal prediction accuracy, SCNN rivals or exceeds the performance of highly trained human experts in assessing prognosis. Our study provides new insights into applications of artificial intelligence in medicine, and also new technical approaches for dealing with intratumoral heterogeneity and training data deficits. We also developed a visualization technique that allows pathologists to explore the associations between histological patterns and prognosis over large whole slide images that inevitably exhibit significant heterogene-

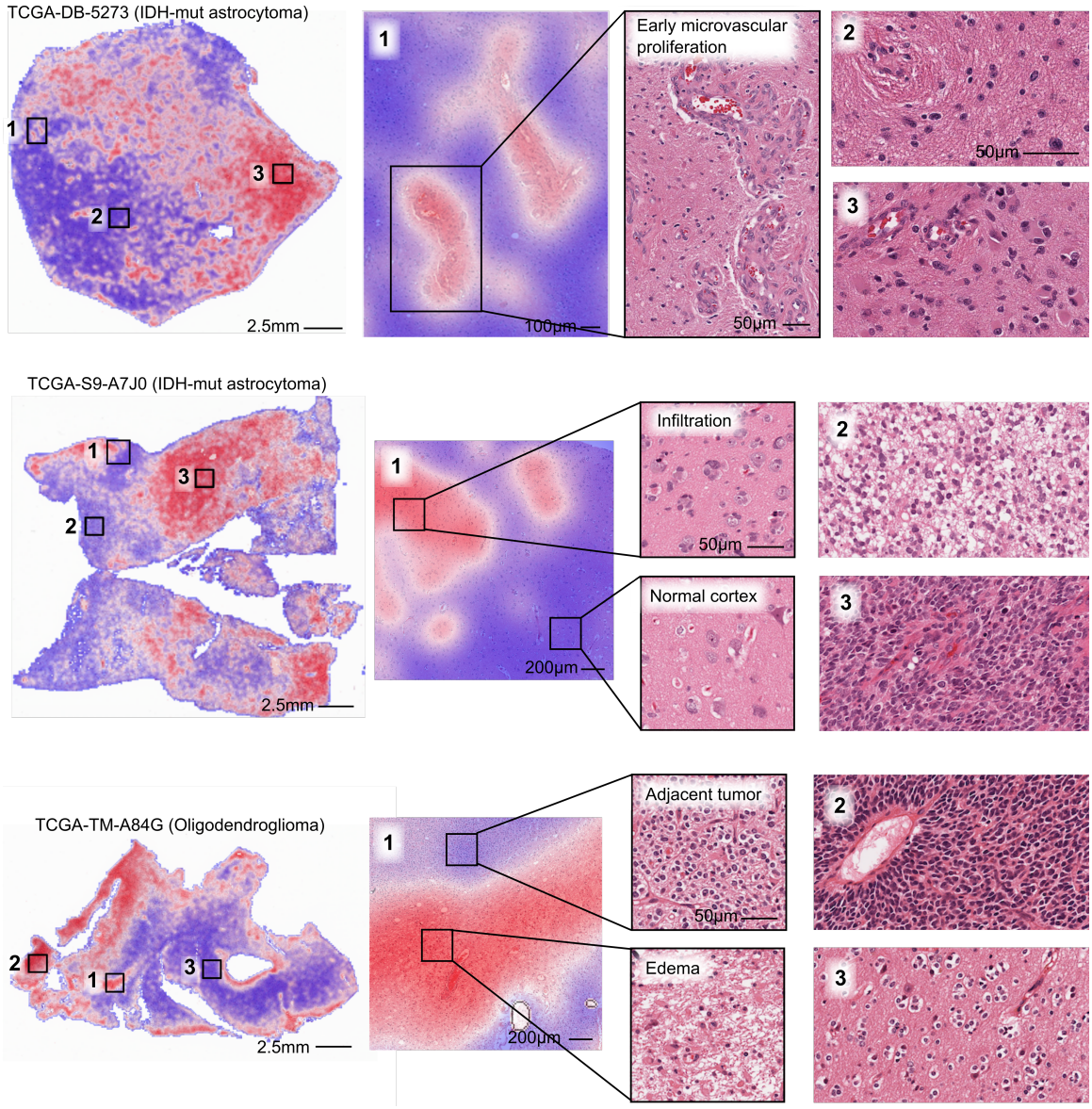

Figure 3.8: Visualizing risk with whole-slide SCNN heatmaps. We performed SCNN predictions exhaustively within whole slide images to generate heatmap overlays of the risks that SCNN associates with different histologic patterns. Red indicates relatively higher risk, and blue lower risk (the scale for each slide is different). (Top) In TCGA-DB-5273, SCNN clearly and specifically associates early microvascular proliferation with high-risks (region 1), and also higher risks with increasing tumor infiltration and cell density (region 2 versus 3). (Middle) In TCGA-S9-A7J0, SCNN can appropriately discriminate between normal cortex (region 1, lower panel) and adjacent regions infiltrated by tumor (region 1, upper panel). Highly cellular regions containing prominent microvascular structures (region 3) are again assigned higher risks than lower density regions of tumor (region 2). Interestingly, low density infiltrate in the cortex was associated with high risk (region 1, upper panel) (Bottom) In TCGA-TM-A84G, SCNN assigns high risks to edematous regions (region 1, lower panel) that are adjacent to tumor (region 1, upper panel).

ity.

Our study investigated the ability to predict overall survival in gliomas, a disease with wide variations in outcomes, and an ideal test case where histologic grading and genomic classifications have independent prognostic power. Remarkably, SCNN performed as well as manual histologic grading or molecular subtyping in predicting overall survival in our dataset. Further investigation of the associations between SCNN risk predictions, genomic subtypes, and histologic grades revealed that SCNN can effectively discriminate outcomes in each subtype, effectively performing digital histologic grading. Furthermore, SCNN can effectively recognize differences in images that associate with genomic subtypes, and predict risks accordingly. Oligodendrogliomas have a distinct histology, and so the ability to discriminate this subtype is not unexpected. For astrocytomas, the SCNN network could correctly predict higher risks for grade III IDH wild-type astrocytomas than for grade III IDH mutant astrocytomas, suggesting that SCNN may be able to recognize subtle histologic differences associated with IDH mutations that are not yet appreciated by pathologists. The broad hypermethylation induced by IDH mutations could plausibly affect nuclear appearance, providing a possible explanation of visual differences that are detectable by SCNN, but more investigation of this topic is needed.

To integrate genomic information in prognostication, we developed a hybrid approach that learns from both histology images and genomic biomarkers. The GSCNN model presented in our study significantly outperforms the WHO standard based on identical inputs. By providing molecular subtype data directly to the network, instead of relying on inferences from histology, GSCNN can focus more attention on learning histologic patterns associated with disease progression in each subtype. This result illustrates how complementary genomic and image data can be practically integrated into a single prediction framework, an issue that presents a significant barrier in the clinical implementation of computational prognostication. Our previous work

in developing deep-learning survival models from genomic data has shown that accurate survival predictions can be learned from high-dimensional genomic and protein expression signatures (34). Incorporating additional genomic variables into GSCNN models is an area for future research, and requires larger data sets with both histologic images and rich genomic annotations.

While deep learning methods frequently deliver outstanding performance, the interpretability of these models is limited, and remains a significant barrier in their validation and adoption. The risk heatmap provides insights into the histologic patterns associated with increased risk, and can also serve as a practical tool to guide pathologists to tissue regions associated with worse prognosis. This approach suggests that our network can learn visual patterns associated with histologic criteria used in grading including microvascular proliferation, cell density, and nuclear morphology. Microvascular prominence and proliferation are associated with disease progression in all forms of diffuse glioma, and these features are clearly delineated as high-risk in the heatmap presented for TCGA-DB-5273. Likewise, increases in cell density and nuclear pleomorphism are also associated with increased risk in all examples. In addition to these results, the heatmap analysis provided some interesting results that need to be further investigated. In region 1 of TCGA-S9-A7J0, SCNN assigns higher risk to sparsely infiltrated cerebral cortex than to region 2 that is infiltrated by a higher density of tumor cells (adjacent normal cortex in region 1 is properly assigned a very low risk). Widespread infiltration into distant sites of the brain is a hallmark of gliomas, and results in treatment failure since surgical resection of visible tumor leaves residual neoplastic infiltrates. It is not clear that this is the reason for SCNN assigning high risk to sparsely infiltrated regions, but nonetheless, it is an interesting finding worth pursuing. Similarly, region 1 of TCGA-TM-A84G illustrates a high risk associated with low cellularity edematous regions, compared to adjacent oligodendroglioma with much higher cellularity. Edema is frequently observed within

gliomas and in adjacent brain and its degree may be related to the rate of growth [95] yet its histologic presence has not been previously recognized as a feature of aggressive behavior or incorporated into grading schemes. These observations confirm that risks predicted by SCNN are not purely a function of cellular density or nuclear atypia and demonstrate that these methods can identify novel, potentially practice changing features associated with increased risk embedded within pathology images.

3.10 Limitations and future work

Although our study provides insights into deep learning for precision medicine, it has some important limitations. A relatively small portion of each slide was used for training, and the selection of regions of interest requires expert guidance. More advanced methods are needed for automatically selecting regions and for incorporating more of the slide into the learning and prediction process. A single whole slide image can contain remarkable heterogeneity, and so incorporating more of the slide into the learning process will require more advanced training methods. Our method currently produces a dimensionless risk by optimizing partial likelihood, and learning of the baseline hazard would permit calibrated prediction of actual survival times. Finally, while we have applied our techniques to gliomas, validation of these approaches in other diseases is needed and could provide additional insights. Furthermore, our methods are not specific to histology imaging or cancer applications, and could be adapted to other medical imaging modalities and biomedical applications.

Chapter 4

Architectures for Aggregate Learning

The models we developed in previous chapter learn to map each single input image (HPF) to the patient outcome during the training. But, in reality it is not only one HPF that is contributing to the patient outcome, but it is a pool of HPFs which altogether with different significance are contributing to the patient outcome. To address this challenge, in this chapter, we have developed an model that learns to predict patient outcome from a collection/set of HPFs. Our end-to-end pipeline has 3 key features - glimpsing, attention, and aggregation which altogether emulate human attention to high-yield areas and aggregation across regions. To evaluate our model performance we have shifted focus from glioma data to placenta as there is a clear application here for learning from aggregation as different spatial locations and patterns inside the placenta data has different implications. We used this network to estimate the gestational age (GA) of scanned placental slides and compared it to a similar network lacking the attention and aggregation functions. Our proposed model, GestAltNet, points toward a future of genuinely whole-slide digital pathology by incorporating human-like behaviors of attention and aggregation.

4.1 Abstract

The placenta is the first organ to form and performs the functions of the lung, gut, kidney, and endocrine systems. Abnormalities in the placenta cause or reflect most abnormalities in gestation and can have life-long consequences for the mother and infant.

Placental villi undergo a complex but reproducible sequence of maturation across the 3^{rd}-trimester. Abnormalities of villous maturation are a feature of gestational diabetes and preeclampsia, among others, but there is significant interobserver variability in their diagnosis. Machine learning has emerged as a powerful tool for research in pathology. To capture the volume of data and manage heterogeneity within the placenta, we developed GestAltNet, which emulates human attention to high-yield areas and aggregation across regions. We used this network to estimate the gestational age (GA) of scanned placental slides and compared it to a baseline model lacking the attention and aggregation functions.

In the test set, GestAltNet showed a higher r^2 (0.9444 vs. 0.9220) than the baseline model. The mean absolute error (MAE) between the estimated and actual GA was also better in the GestAltNet (1.0847 weeks vs. 1.4505 weeks). On whole slide images, we found the attention sub-network discriminates areas of terminal villi from other placental structures. Using this behavior, we estimated GA for 36 whole slides not previously seen by the model. In this task, similar to that faced by human pathologists, the model showed an r^2 of 0.8859 with an MAE of 1.3671 weeks.

We show that villous maturation is machine-recognizable. Machine-estimated GA could be useful when GA is unknown or to study abnormalities of villous maturation, including those in gestational diabetes or preeclampsia. GestAltNet points toward a future of genuinely whole-slide digital pathology by incorporating human-like behaviors of attention and aggregation.

4.2 Introduction

The placenta is the first organ to form and functions as the fetal lung, gut, kidney, endocrine, and immune systems. As an active participant in gestation, it consumes as much oxygen at term as the entire fetus [96]. Placental pathology causes and reflects adverse events in pregnancy [97, 98]. Pathology in the placenta can have lifelong consequences for mothers and offspring, including increased risk of cardiovascular disease [99], bronchopulmonary dysplasia [100], cerebral palsy [101], colorectal carcinoma [102], and asthma [103]. Therefore, the examination of the placenta can yield considerable benefit. Yet, less than 20% of placentas are examined in the United States, and significant lesions are frequently unrecognized [104, 105].

Digital pathology has the potential to revolutionize our understanding of placental function and disease [106]. Routine diagnostic pathology relies on qualitative assessment and pattern recognition. Research studies on human placentas usually rely on these assessments or quantitative measurements of selected regions done by hand. A more quantitative, thorough examination may identify new biology and pathophysiology. The sheer volume of archived glass slides of placentas, 120,000 at our institution alone, with 500,000 cells in each whole slide image (WSI), provides an enormous untapped reservoir of material for hypothesis development and testing.

In comparison, clinical examination captures only a fraction of the information from each slide, and the quality is dependent on the examiner. Despite the accessibility of placentas at the time of birth, that information is discarded in most cases. Once an AI system is operating, increasing the scale, adding new populations or diseases is simple. This could include placentas from low-resource or international settings, patients with specific sociodemographic factors, or patients with emerging diseases of pregnancy, like COVID-19.

4.2.1 Changes over time

Over the course of the 2^{nd}- and 3^{rd}-trimesters, the placental disc increases approximately 10-fold in size. The most significant microscopic changes are within the terminal villi, with increased numbers of small villi with decreased cellularity, increased stromal density, migration of capillaries to below the syncytial membrane, and collection of syncytiotrophoblast nuclei into knots. These changes have the overall effect of minimizing the distance between maternal and fetal blood [107, 108, 109]. In analogy with the lung, this results in maximum surface area with minimum diffusion distance for oxygen and nutrients (Figure 4.1). Determination of the appropriateness of villous maturation is a key step in assessing a placenta. This task is daunting, as it involves the integration of the factors mentioned above across multiple slides to form a single gestalt. Accordingly, interobserver variability is high [110, 111, 112].

Gestational age (GA) is the single most important factor in perinatal well-being. The probability of a newborn successfully transitioning from womb to nursery to home increases markedly with GA, and the probability of adverse outcomes including hypoxic ischemic encephalitis, necrotizing enterocolitis, and bronchopulmonary dysplasia markedly decrease [113]. Accurate identification of GA most commonly relies on sonographic measurements made in the 1^{st}- or 2^{nd}-trimester [114, 115, 116]. These measurements may not be available in low resource settings or when prenatal care is inadequate. Other methods, such as the recalled date of the last menstrual period or sonographic measurements made in the 3^{rd}-trimester, are less accurate.

4.2.2 The placenta and digital pathology

Compared to neoplasia, the placenta is relatively understudied by digital pathology. Studies using photomicrographs of single fields and manual annotation show the potential for scientific discovery using deep, image-based phenotyping of the placenta. Manual measurement of villous and vascular surface area has shown changes over

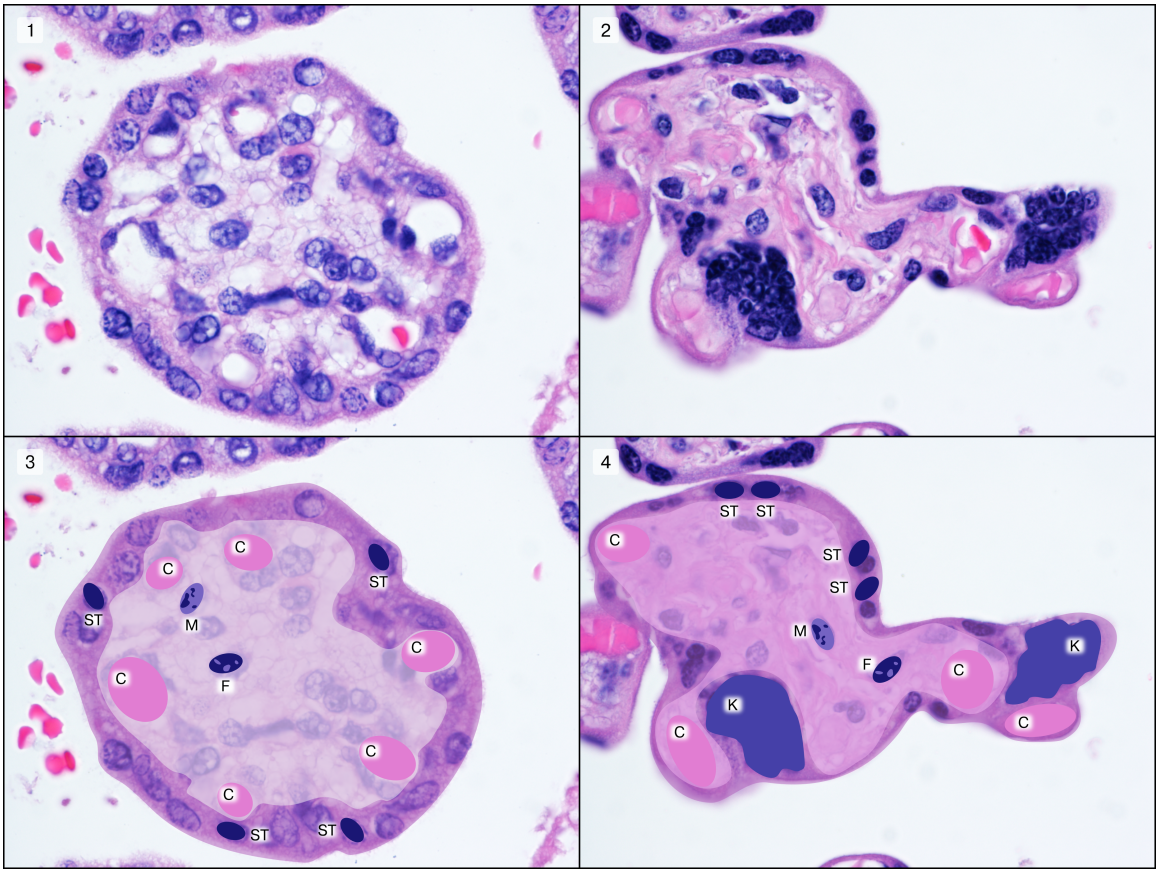

Figure 4.1: Changes in terminal villi over gestation. In the early 3^{rd}-trimester (24 weeks GA, panels 1 and 3), syncytiotrophoblast (ST) nuclei are evenly spaced. Capillaries (C) are distant from maternal blood, which bathes the villi. The stroma consists of loose extracellular matrix proteins with frequent macrophages and fibroblasts (M and F). At term (40 weeks GA, panels 2 and 4), the villi are smaller. Syncytiotrophoblast nuclei are gathered into knots (K), thinning the vasculo-syncytial membrane. Capillaries are directly beneath the syncytiotrophoblast layer. Stroma is denser with lower cellularity.

pregnancy [107, 108]. Preeclampsia (PreE) has been associated with changes in villous count, area, diameter, capillary count, and degree of capillarization in the villous core [109]. Gestational diabetes has been associated with decreased villous vascular volume [117]. Abnormal villous maturation has a genetic expression signature - placentas with a diagnosis of accelerated maturation have gene expression more appropriate for placentas delivered 4.7 weeks later with normal maturation [118].

More recent studies support the feasibility of applying modern machine learning and digital pathology techniques to the placenta. Studies have shown the ability to segment villi from scanned slides and measure their stromal density and vessel numbers [119, 120]. Published algorithms exist for identifying cytotrophoblast, fibroblast, macrophage, syncytiotrophoblast, and vascular endothelial cells in the placenta [121].

Deep learning models employing convolutional neural networks (CNN) have shown impressive performance for identifying image content in multiple domains and tasks, including digital pathology [29, 20, 21, 22, 23, 24]. In training, networks commonly learn to associate a single image or HPF to an outcome or finding of interest. Contrary to CNN's implicit assumption of one image corresponding to one label, a single WSI contains thousands of HPF with considerable heterogeneity. Practicing pathologists must examine all HPF, attend to fields they consider representative, and aggregate their findings to produce a single diagnosis. The gap between algorithm development and practice reduces the clinical relevance of many AI studies including those in the broader medical imaging field. We propose an algorithm that learns the patient outcome from a collection or set of images in training. This helps to incorporate more regions from each WSI during the learning procedure.

The problem of aggregation extends beyond digital pathology and is present whenever a model receives multiple inputs. Practitioners must decide at which stage of the pipeline data are incorporated, how they are weighted, and the extent to which aggregation is trainable. In non-image tasks, data are routinely input as a single

vector allowing complex trainable interactions. Conversely, ensemble strategies may aggregate results from multiple separately trained models without back propagation. Choices in aggregation strategy are liable to be suboptimal if practitioners are unaware that a choice is being made.

This study aims to develop a deep learning model that incorporates and predicts across whole slides and demonstrates the utility of that model in estimation of gestational age in placenta - a low concordance task in a notoriously heterogeneous tissue.

4.3 Materials, subjects, and methods

4.3.1 Patients and materials

Pathology reports from patients delivering 1/1/2010 to 10/31/2019 were retrieved from the laboratory information system (Cerner Build List Id: 2014.08.1.36). GA, clinical history, and diagnoses, including accelerated, delayed, and appropriate maturation, were extracted using regular expressions (6.2) and the Natural Language Toolkit (NLTK, version 3.3) on Python (version 3.6.9) as described [122, 123].

We identified cases with an obstetrically determined GA of 24-42 weeks with an original pathologic diagnosis of appropriate villous maturation, confirmed through a review by a practicing perinatal pathologist at Department of Pathology at Northwestern University (Dr. Jeffery A Goldstein). This GA was considered ground truth for each case.

Clinical examination of placentas at our institution includes 1 cassette of membranes, 1 of umbilical cord sections, 1 with three incisional biopsies of the placental disc's maternal surface (basal plate plus villi), 2 cassettes of representative nonlesional full-thickness placental disc, and additional cassettes containing any lesions. The maternal surface biopsies and full-thickness sections are selected from the in-

ner $\frac{2}{3}$ of the radius of the placental disc were reviewed for possible scanning. We selected a slide with morphology consistent with clinically determined GA without mass-forming lesions or villous abnormalities. Given low counts in the earliest GA, we allowed cases with decidual or chorionic plate pathology (e.g. chorioamnionitis).

One slide per patient with villous tissue, either basal villous wedges or full-thickness placental disc, was selected and scanned at the institutional Pathology Core Facility using a Hamamatsu Nanozoomer 2.0 HT scanner at 20X objective magnification. 154 slides were split randomly, stratified by GA, into training, validation, and test sets with proportions of 70% (107 slides), 15% (23 slides), and 15% (24 slides), respectively. Because deliveries are not evenly distributed across the GA and maturation anomalies are more prevalent at earlier GA, the training, validation, and test sets are not precisely balanced at each GA. The number of cases and corresponding ROIs and HPFs is presented in Table 4.1.

Table 4.1: Number of cases and corresponding ROIs and HPFs

	WSI (train; valid; test)	Annotations (ROIs)	512×512 Patches (HPFs)
Annotated Data	154 (107; 23; 24)	1918	26555
Non-annotated Data (WSI-level testing)	36 (0; 0; 36)	0	152289

Regions of terminal villi with villous maturation consistent with GA were box annotated by the pathologist. Stem villi, areas of fibrin deposition, and septae were avoided. On full-thickness sections, parabasal areas were preferentially annotated. In total, 1918 region annotations (at least 10 per slide) were made. Regions were extracted with OpenSlide (1.1.1) on Python (3.6.9) and were color normalized using the method from Macenko et al. [124]. Regions were tiled into 512×512 pixel high power fields (HPF) at 20X magnification level and shrunk to 256×256 (effective magnification 10X), for a total of 26,555 HPF (Table 4.1). During training, HPF are

augmented by random rotations and changes in brightness and contrast [125].

4.3.2 Baseline model

HPF are input into a feature extraction convolutional neural network based on VGG19 [14] with trainable weights initialized by a pre-trained model on ImageNet [126] in Keras (Tensorflow 2.3.0). The network is modified by replacing the fully connected layers in the original VGG19 architecture with a single fully connected layer of size 1024 with ReLU activation function and a dropout with rate 0.5. The extracted feature map is submitted to the representation learning sub-network, which consists of sequential fully connected layers of size 1024 and 256 with ReLU activation functions and a dropout with rate 0.5 after the first fully connected layer, and one linear node at the end to produce a single value - the estimated gestational age (EGA). The mean squared error loss between EGA and clinically determined GA (as the ground truth) is used to train the model. The baseline model was trained for 2000 epochs. To aggregate across a WSI for inference, the median EGA for all HPF is determined post hoc.

4.3.3 GestAltNet - input & glimpsing mechanism

In the base model, training explicitly links the clinical outcome to a single HPF. We propose an alternative network for estimating gestational age, GestAltNet (Figures 4.2 and 4.3). GestAltNet learns in aggregate from a collection of images and relates the clinical outcome to a set of HPF during training. While the baseline model trains using a single HPF as input, GestAltNet uses a glimpse as input in training. Each glimpse consists of 16 randomly selected HPF from a single WSI, generally representing multiple regions. Glimpses are examined in batches of 64 and consumption of all batches represents one epoch. HPF and glimpses are resampled as needed to maintain glimpse and batch sizes. HPF are randomly assigned to glimpses at initialization and

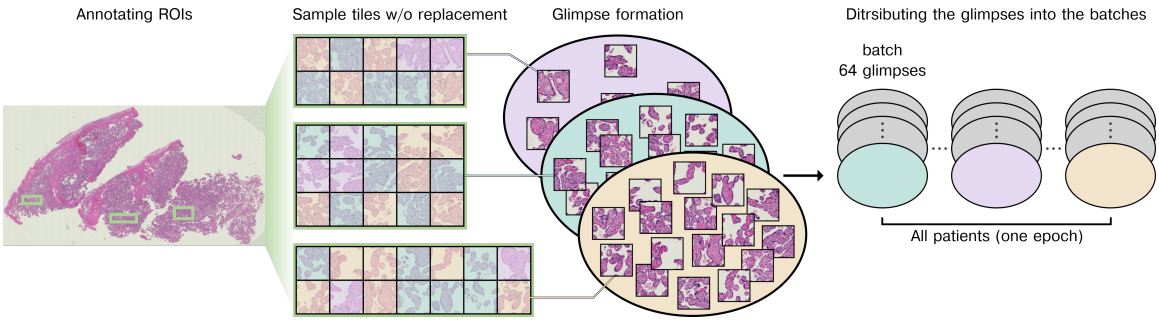

Figure 4.2: Glimpse and batch formation: Scanned whole slide images are annotated, and ROI are extracted (left panel). ROI are tiled into HPF (2^{nd} panel, black lines). HPF are randomly sampled without replacement across all ROI of each patient to form a glimpse (3^{rd} panel, HPF shading indicates glimpse) 2^{nd} panel from left, colored HPF indicate their corresponding glimpse. Glimpses are constant size (16) except the last glimpse (purple oval), which takes the remainder. Glimpses from one patient are distributed across batches (4^{th} panel, gray ovals are glimpses from other patients).

after every 50 epochs (chosen based on the performance in the validation set).

4.3.4 GestAltNet - pipeline & attention and aggregation

As in the baseline model, images are input into a VGG19 derived network. The intermediate output of VGG19 at block3, consisting of 256 3×3 kernels (Figure 4.3, red squares), is input to the attention sub-network. This sub-network is a feedforward neural network with two fully connected layers of size 256, 256 with ReLU activation functions, a dropout with rate 0.5 after the first fully connected layer [25], and one linear node at the end. The linear node results in a single scalar value for each HPF in the glimpse, representing its attention. To limit extreme values, attentions are transformed using softmax.

A single aggregate feature map ($\bar{f}$ in Figure 4.3) is obtained through weighted averaging over the feature maps of the 16 HPF within the glimpse, where weights are the corresponding HPF attentions. The aggregate feature map is submitted to the representation learning sub-network as in the baseline network to compute EGA. During training, mean squared error between EGA and clinically determined GA

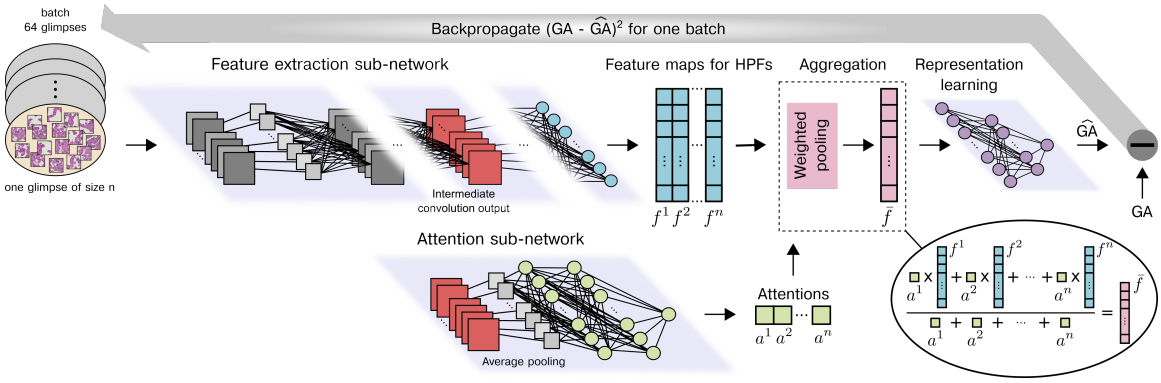

Figure 4.3: Model pipeline: Glimpses are submitted as a batch to a convolutional neural network (feature extraction sub-network). Intermediate outputs (red boxes) are input to an attention sub-network. Feature maps ($f^1 - f^n$) are weighted by their attentions ($a^1 - a^n$) and aggregated via weighted averaging (oval). The representation learning sub-network estimates the gestational age ($\widehat{GA}$) based on the aggregated feature map $\bar{f}$. The mean squared error $(\widehat{GA} - GA)^2$ inside an entire batch of 64 glimpses is used in backpropagation. The whole learning procedure is done in an end-to-end manner.

(ground truth) is used as the loss function, and backpropagation is performed end-to-end across the entire network. GestAltNet was trained for 500 epochs. For whole slide inference, the median EGA, computed across glimpses, was determined.

4.3.5 Evaluation metrics

To assess the overall accuracy, we measured the coefficient of determination (r^2) and the absolute error in weeks. For test and unannotated slides, EGA was calibrated using the linear regression of EGA vs. GA for validation regions and whole slides (respectively). We considered an absolute error of greater than 3 weeks as clinically significant because 1) accelerated villous maturation has been diagnosed based on an apparent GA of $\geq$ 37 weeks with chronologic GA of $\leq$ 34 weeks, i.e., 3 weeks [127]; 2) gene expression study showing accelerated villous maturation equates to 4.7 weeks ahead, and delayed maturation equates to 1.5 weeks behind normal gestation (average 3.1 weeks) [118]; 3) Using the placental weight reference of Pinar et al., [128] a placenta of average weight at one GA is considered large or small for gestational

age (LGA, SGA) 3 - 5 weeks earlier or later. For example, a placenta with the mean weight for 24 weeks, 189 grams, is considered LGA at 21 weeks (expected 114 - 172 grams) and SGA at 27 weeks (expected 192 - 305 grams).

4.3.6 Attention and whole slide estimation of GA

For the whole slide level inference 36 new slides, neither previously annotated nor part of the training, validation and testing sets were used. The non-tissue area of the WSI was masked out by first applying gaussian smoothing to the slide's grayscale thumbnail, and then applying Otsu's image binarization method [129] to the thumbnail. Attention was determined and GA was estimated on a per-HPF basis for all HPF. To determine appropriate attention thresholds for the selection of representative HPF in WSI level inference, we examined the per-HPF attention and accuracy over the non-overlapping HPF inside the tissue area of the WSI in our validation set. We set the lower threshold at the median attention of HPF with absolute errors of ≤ 3 weeks and the upper threshold at the 99^{th} percentile of attention for HPF with absolute errors of ≤ 3 weeks in the validation set.

For generating heat maps, 87.5% overlapping HPF were extracted, and attention and EGA values were produced on a per-HPF basis. Attention was colored with minimum and maximum values scaled based on variation in the validation set. EGA was colored as H&E (appearing pink at low power) for absolute error ≤ 3 weeks, red if > 3 weeks high and blue if > 3 weeks low.

4.4 Results

4.4.1 Interobserver variability

29,943 placentas were examined over 9.5 years by 8 pathologists. Given a GA determined by clinical parameters, pathologists diagnose whether maturation is appro-

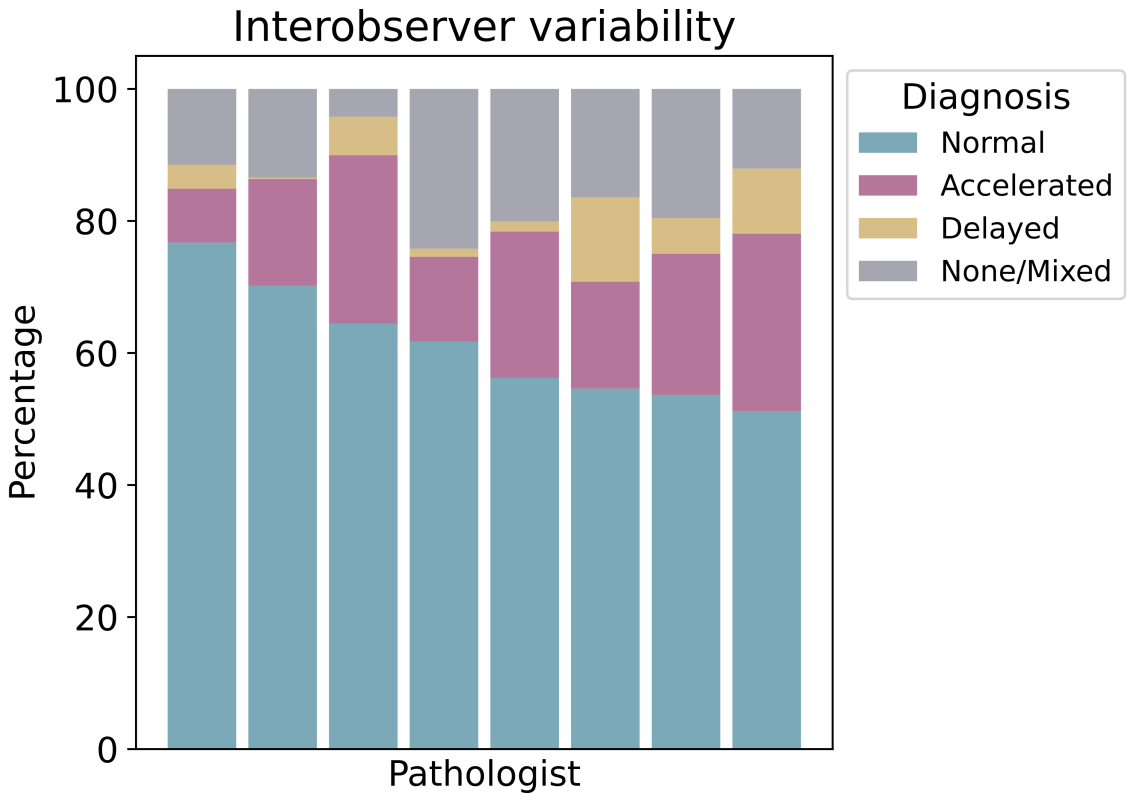

Figure 4.4: Interobserver variability in clinical diagnoses. Despite well-defined patterns of maturation, pathologists are inconsistent in their diagnoses of whether the villous maturation is normal (green), accelerated (red), or delayed (yellow) for the stated gestational age. Each column represents one pathologist.

priate, accelerated, or delayed for the stated GA. Overall, 17,806 (60%) placentas were diagnosed with appropriate maturation, 5,108 (17%) with accelerated maturation and 1024 (3.4%) with delayed maturation (Figure 4.4). 6,005 placentas (20%) received multiple diagnoses, for example, "appropriate for gestational age with regionally delayed maturation," or had no description of maturation, which may occur when maturation is obscured by other findings like chorangiosis or post-mortem changes. The percentage of cases diagnosed as normal varied from 51% to 77%, as accelerated from 8.2% to 27%, and as delayed from 0.2% to 13%. Assuming a random distribution of placentas among pathologists, this represents significant interobserver variability.

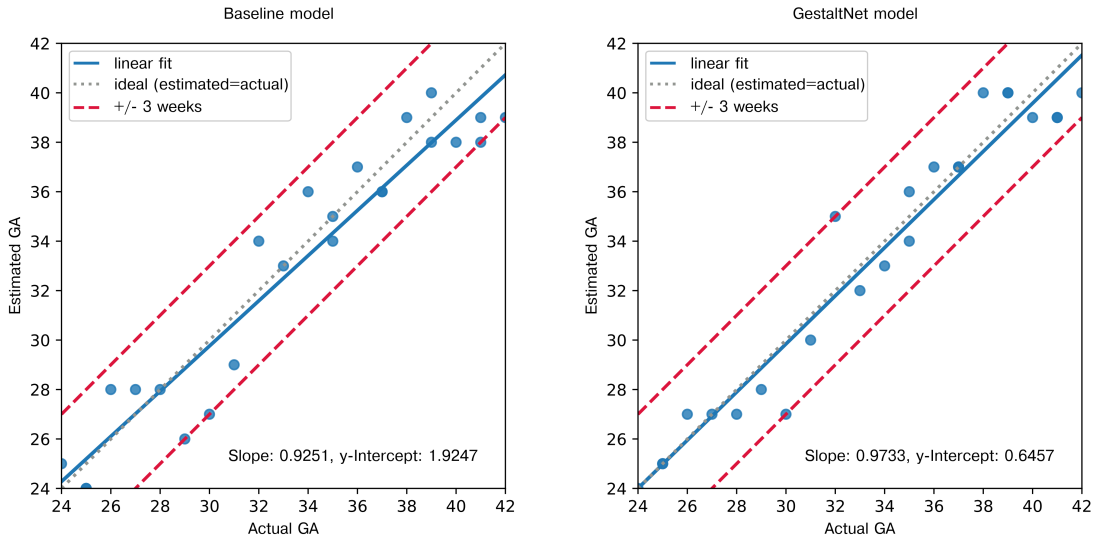

Figure 4.5: Test Results: (a) In the test set, the baseline model shows an r^2 of 0.9220 with an MAE of 1.4505 weeks. (b)The GestAltNet shows an r^2 of 0.9444 with an MAE of 1.0847 weeks.

4.4.2 Deep learning model performance

In the test set, the GestAltNet and baseline models showed r^2 of 0.9444 and 0.9220, respectively (Figure 4.5a-b). After calibration, the mean absolute error (MAE) was 1.0847 weeks for the GestAltNet model and 1.4505 for the baseline model. An error of ≥ 3 weeks is significant in evaluating GA. By this standard, both the GestAltNet and baseline models adequately estimated GA 24/24 test cases.

4.4.3 Attention and estimation of GA across whole slides

The GestAltNet technique simulates a pathologist's cognitive process of incorporating information across multiple regions of interest. However, it still relies on hand-annotated regions of interest selected to include representative, high-quality areas of tissue. To explore variation across tissue and emulate the pathologist attention and gestalt formation process across the whole slide, we obtained attention and EGA across 36 WSI that were unannotated and not part of the existing training, validation,

or test sets. This resulted in an r^2 of 0.8859 and an MAE of 1.3671 weeks. The model estimated GA was within 3 weeks of the actual GA in 35/36 (97.22%) cases (Figure 4.6). To illustrate and further examine how WSI attention and prediction relate, we generated whole-slide attention and predictions for one WSI using overlapping HPF (Figure 4.7). Perhaps surprisingly, given that we did not train our model to discriminate between different regions of the placenta, terminal villi show the highest attention, while stem villi, basal plate, and chorionic plate showed lower attention. GA estimation was variable within the villous region; however, the most accurate areas tended to be away from large stem villi or other masses. Some non-villous areas, including chorionic vessels, are attended to with divergent and inaccurate predictions.

4.5 Discussion

GA is the most significant factor in neonatal well-being. However, practicing pathologists rely on GA derived from other factors and show considerable inter-rater variability even in identifying whether the villous appearance is appropriate for the stated GA. We show that GA can be predicted with extraordinary accuracy from the beginning of viability (24 weeks) to post-term (42 weeks) using a deep learning approach. In practice, pathologists examine several regions across multiple whole slides, looking for different features that are either concordant or discordant with the chronological GA.

Developing a model for this task requires a solution to what we call "The Problem of Aggregation." Our solution is to analyze multiple HPF in a glimpse. Aggregation occurs at the feature map stage. Feature maps are weighted based on the attention generated by an independent multilayer perceptron. The model takes the form of a single end-to-end network in which all sub-networks are trainable. We show that the integration of image features at an early stage with weighting and end-to-end

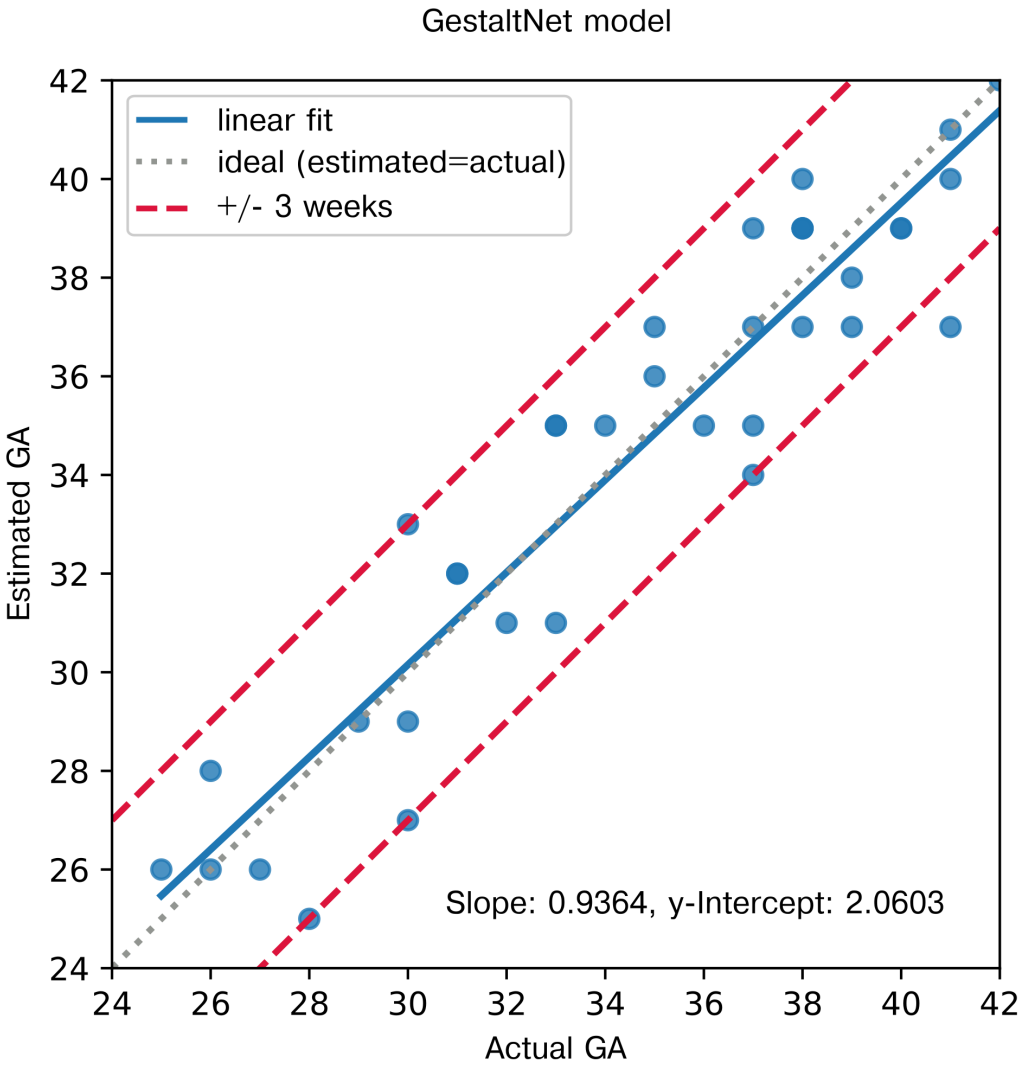

Figure 4.6: WSI Level Test Results on unannotated set: In this set of not previously seen slides, the model estimates GA with an r^2 of 0.8859 with an MAE of 1.3671 weeks. 35 of 36 cases were called correctly within +/- 3 weeks (red lines).

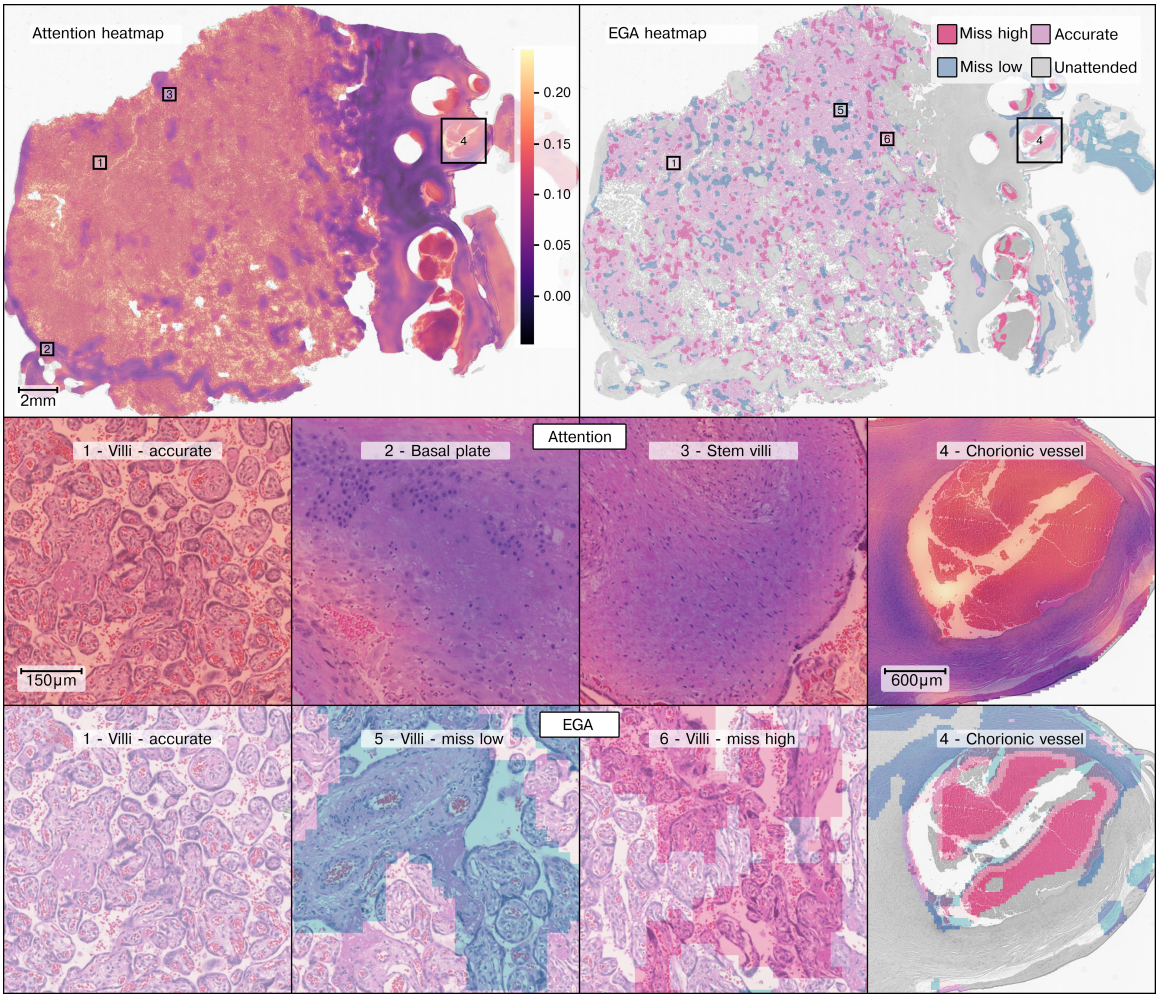

Figure 4.7: Example whole-slide attention (top left, detail - middle row) and EGA (top right, detail - bottom row). Terminal villi are primarily high attention (yellow, regions 1). Basal plate (left side of WSI and region 2), stem villi (region 3, intermixed with villous areas), and chorionic plate (right side of WSI and region 4) are generally low attention (purple). Estimated gestational age shows variegation with accurate areas (region 1) intermixed with areas with inaccurate low (blue, region 5) and high (red, region 6) estimates. Areas with low attention are disregarded (grayscale). The model is not explicitly trained to recognize tissue types and shows erroneous high attention to some areas. For example, one chorionic plate vessel (region 4) is part high- and part low-attention. The attended part of the vessel wall gives an estimate that misses low. Intravascular blood is attended and misses high.

trainability provides superior accuracy compared to post hoc averaging used in the baseline model. The improvement is highlighted by the stress test of calculating EGA without the regularization provided by human annotation.

One of the characteristics of deep learning algorithms that has made them so successful in digital pathology is their end-to-end learning approach. These adaptive algorithms learn to predict labels directly from pixel values in contrast to prior approaches that seek to incorporate a-priori knowledge in algorithm design. The unbiased end-to-end learning method is often credited as enabling deep learning models to learn latent predictive features in histology that may not be appreciated by human pathologists, but at the cost of algorithm interpretability.

End-to-end learning becomes practically difficult when labels correspond to an entire slide or a large region rather than a high-power field due to the scale of data corresponding to a single label and the limitations of computer hardware used to train deep learning algorithms. In this scenario, end-to-end learning requires that the mechanism for aggregating over multiple fields be incorporated into the learning model and be adaptive. In applications like tumor detection, a single positive field gives the whole slide label, and have been solved using approaches like multiple instance learning. Other applications may be more compositional, requiring the interpretation and weighting of several tissue patterns, or learning to perform a weighted averaging over regions of the slide.

This paper provides a solution involving exhaustive random sampling of HPF representing a single case with the differential weighting of HPF by attention. This strategy is broadly applicable to any scenario when large amounts of data are consumed for each sample. However, it is particularly relevant for image analysis, where the interpretation of one portion of the image depends on context from other portions. For example, a pedestrian waving to another pedestrian on the other side of a street is more likely to enter the street than one waving to a departing car.

In pathology, injured liver adjacent to a liver tumor represents mass effect, not cirrhosis. GestAltNet assigns attention weights on a per-HPF basis. This reflects the variability in information content between HPF, even within human-annotated ROI. Within-image attention, for example Grad-CAM, has been proposed to address the problem of interpretability in AI [130]. Theoretically, our attention could be used in a similar fashion, analogous to the use of dotting pens in pathology practice to annotate key areas for diagnosis. Within-image attention has been criticized for focusing on edges or complex structures and using similar patterns of attention to explain correct and incorrect answers [131]. It is not clear that a by-HPF system, such as GestAltNet, is immune from this problem, and the observation that it assigns similar attention to correct, miss-high, and miss-low regions (Figure 4.7) is concerning.

Our choice of a single end-to-end network is also appealing in that it reflects human cognition, and all operations are potentially trainable. This mimics human thought patterns of aggregating impressions rather than diagnoses. Features may also be a more worthy area of focus as they are representations of biological phenomena, while HPF are arbitrary grids imposed by computer memory limitations. Other authors have addressed the aggregation problem in placenta with success. Clymer et al. use the multiple-resolution pyramid of images found in scanned slide files to identify vessels within placental membranes followed by clustering to produce a slide-level diagnosis as either containing healthy or pathologic maternal vessels [132]. However, this study did not use end-to-end training.

4.6 Limitations and future work

From a generalizability standpoint, the most significant limitations of this work are the use of a single site with consistent protocols and a single pathologist reviewer. Further work is necessary to develop and demonstrate generalizability across institu-

tions and practitioners. Our demonstration of interobserver variability is limited in that pathologists are not reviewing the same placenta, but rather placentas submitted more or less randomly from the same population. The remainder of this work suggests that human-machine collaboration to overcome this variability will be more productive than perseverating on the precise degree of heterogeneity.

This is among the first studies using machine learning in placental pathology and demonstrates the potential of this field. The extremely high accuracy in detecting normal morphology across gestation will allow the classification of many abnormalities, some currently unknown or with too low interobserver reliability to be useful.

In high-resource settings, GA is usually determined by 1st-trimester ultrasound. The system demonstrated is unlikely to replace this method but could be useful in cases where the dating of the pregnancy is unclear, or there is a discrepancy between the stated and apparent GA. In low or middle-income settings, photomicrographs of relevant areas taken using a smartphone and adapter could be used in lieu of whole slide images [133]. In this use-case of human-machine cooperation, the small size of captured images means that a cloud-based network could provide estimated GA in real-time.

Accelerated and delayed villous maturation are among the most commonly reported placental findings in large data sets [122]. Nonetheless, they show poor interrater reliability, decreasing the significance of these findings. AI could be used in a quality assurance/improvement paradigm to improve interobserver variability in practice and is likely useful in identifying maturation abnormalities.

Our solutions to the problem of aggregation, as used in GestAltNet, will have applications far beyond the placenta. Intratumoral heterogeneity complicates neoplasia classification and is a marker for adverse outcomes [134, 135, 136]. In other non-neoplastic diseases, such as idiopathic pulmonary fibrosis, heterogeneity itself may be a criterion [137]. Beyond digital pathology, attention and aggregation within large

and complex images remain fundamental challenges of image analysis.

4.7 Conclusion

In conclusion, we report the machine learning-based estimation of GA from scanned histologic slides of the placenta. This demonstrates the tractability of this system and may be useful in diagnostic, quality, and research settings. We present a novel aggregation and attention model to manage and utilize the vast quantity of data present in whole slides.

Chapter 5

Bayesian Survival Neural Networks

5.1 Abstract

Applying Bayesian learning to neural networks results in powerful and flexible non-linear models that can be used in many applications from regression to prediction and classification. A Bayesian learning framework enables studying all sources of uncertainties by probabilities. In this chapter, I investigate the use of Bayesian neural networks in modeling the aleatoric and epistemic uncertainties in survival prediction. In general, it is important to build models that enable uncertainty quantification in tasks where the lack of confidence in predictions can have catastrophic effects. In applications such as predicting outcomes for cancer patients, this can have a direct impact on the decision-making process for treating patients. In such tasks, erroneous predictions might result in irreparable outcomes. In practice, we cannot deploy a model in medical settings, unless we build a model that *reliably knows what it doesn't know* and asks for extra measurements or for the human specialist to intervene in the decision-making process in cases where there is a lack of confidence in its predictions. In recent years, there have been a lot of improvements in building deep neural networks for survival prediction from high dimensional data [138, 22, 139, 29, 3].

However, there have been few works that have addressed the problem of modeling uncertainty in survival prediction [31, 140]. Also, it is unclear which type of uncertainty most of these proposed methods for survival prediction are modeling, and there have not been enough attempts to differentiate different sources of uncertainties in survival neural networks, despite its critical role in building reliable survival models and in the decision-making process. Two main types of uncertainty are aleatoric and epistemic uncertainty. The aleatoric uncertainty is based on the noise or some inherent variability in the data, such as mislabeled data; this type of uncertainty cannot be addressed by incorporating more samples into model training but is reducible by measuring additional features. Whereas, the epistemic uncertainty is based on the uncertainty in model parameters and is reducible by incorporating more samples into model training. It is important to appropriately differentiate the epistemic and aleatoric uncertainties, as each of them has different underlying causes that require different treating. Therefore, this chapter seeks to model the aleatoric and epistemic uncertainties in survival neural networks under a Bayesian framework.

5.2 Introduction

5.2.1 Uncertainty analysis

Deep learning is known to be a powerful tool for learning representations from high dimensional data. The predictions made by deep learning methods are often assumed to be accurate. However, during the past years, this assumption has resulted in some catastrophic consequences such as a fatal accident caused by the confusion of a perception mechanism in an assisted driving system [19]. If the algorithm was able to assign high uncertainty to its erroneous predictions in such a system, it would have been possible to avoid its disastrous consequence by making a better decision.

Uncertainty analysis is used to help the decision-making process through quantify-

ing the uncertainties in the relevant variables. Aleatoric and epistemic uncertainties are two of the main types of uncertainties one can quantify in Bayesian modeling [141, 19] (Figure 5.1). Both of the aleatoric and epistemic uncertainties are present in real life applications. The goal in quantifying uncertainties is to express each of them separately, which is helpful in designing the study, data accumulation, selecting appropriate measurements and during inference. Following is the brief explanation of each of the uncertainties.

Aleatoric uncertainty

Aleatoric uncertainty is the type of the uncertainty that arises from the noise that is inherent in the observations. For instance, this can be the noise inherent in a measurement process. This type of uncertainty is not reducible by incorporating more samples, but it can be addressed by additional measurements. Aleatoric uncertainty is further categorized into homoscedastic and heteroscedastic uncertainties. Homoscedastic uncertainty is constant for different values of patient features, and heteroscedastic uncertainty is dependent on the inputs (i.e. patient features) to the model; in other words the heteroscedastic uncertainty would be conditional on patient features, and different patients would have different uncertainties in this case. This is important because some types of disease might have more variability in outcomes than others.

Epistemic uncertainty

Epistemic uncertainty which is also known as model uncertainty originates from the uncertainty in the model parameters. This type of uncertainty is because of limited data and hence knowledge and is reducible by incorporating more samples (data).

Our goal in this chapter is to build a *Bayesian neural network* for survival prediction that quantifies the aleatoric and epistemic uncertainties.

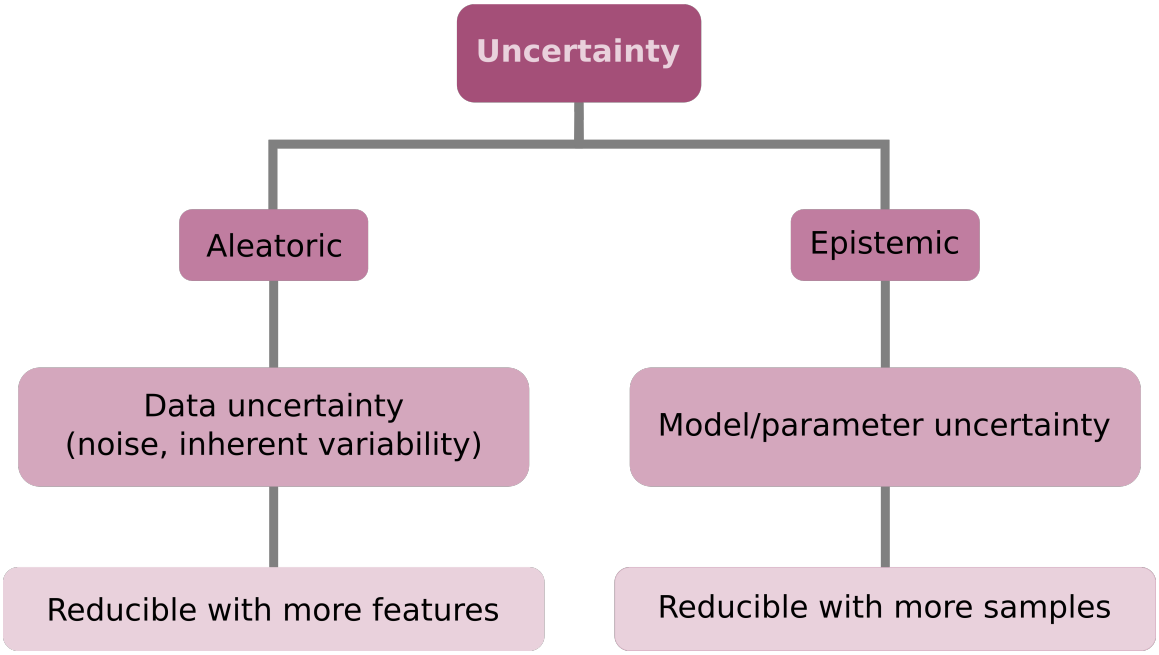

Figure 5.1: Different types uncertainty. Aleatoric uncertainty is resulted from the inherent noise in data and is only reducible with obtaining more features. Epistemic uncertainty is the uncertainty that originates from the underlying uncertainty in model parameters and is reducible with obtaining more samples.

5.3 Methods

5.3.1 Bayesian neural network

Bayesian neural networks combine the universal function approximation power of neural networks with the benefits of probabilistic modeling in building confidence intervals over the predictions. There have been a lot of advances in this field over the past three decades, from introducing the foundations of Bayesian neural networks to the recent advances in leveraging Bayesian neural networks to model uncertainties [142, 143, 144]. In probabilistic modeling, first a prior is assumed for the model parameters based on some prior knowledge, and then this prior parameters are combined with likelihood to obtain the posterior based on the Bayes' theorem as follows

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{5.1}$$

Where θ is the set of model parameters, $\mathcal{D}$ is the observed data, $p(\theta|\mathcal{D})$ is the posterior probability, $p(\mathcal{D}|\theta)$ is the likelihood, $p(\theta)$ is the prior, and $p(\mathcal{D})$ is the marginal probability of data which is an unknown constant and acts as a scaling factor in this equation.

Note that the inference in probabilistic modeling gives a complete probability distribution rather than point estimates; hence, it enables building confidence intervals around the predictions.

When training the traditional neural networks with maximum likelihood estimation (MLE), in fact we model $p(\mathcal{D}|\theta)$, by finding model parameters θ that maximize the probability of the observed data $\mathcal{D}$. However, in the context of probabilistic modeling we first assume some prior distribution over model parameters θ before observing the data $\mathcal{D}$, also we assume some interaction between the model parameters and the data by specifying a likelihood. Then, we can estimate the posterior distribution of model parameters $p(\theta|\mathcal{D})$ based on the Bayes' theorem, by which it is proportional to the multiplication of the assumed prior and the specified likelihood.

One of the key points during specification of probabilistic models, is that obtaining the posterior distribution is analytically a challenging task for most of the real-world priors. Therefore, some sampling methods that mostly include Markov Chain Monte Carlo (MCMC) approaches or approximation techniques like variational inference (VI) methods are used to obtain the posterior value of the model parameters.

In MCMC algorithms, the model parameters are sampled in proportion to their probabilities. Metropolis Hastings, Hamiltonian Monte Carlo or its adaptive extension No U-Turn Sampler (NUTS) are among the popular MCMC methods [145, 146, 147, 148]. MCMC methods assume no model for the posterior. Hence, these methods have a low bias but a high variance.

Variational inference methods in other hand, approximate the posterior under an optimization framework. For this, a surrogate posterior (variational posterior)

$q(\theta|\mathcal{V})$ is assumed to approximate the true posterior $p(\theta|\mathcal{D})$, where $\mathcal{V}$ defines the parameter space of the underlying distribution assumed for the surrogate posterior. The surrogate posterior needs to cover a wide variety of distributions to be able to capture the true posterior. Then, the Kullback-Leibler (KL) divergence between the surrogate posterior q and true posterior p is used as the measure of distance between surrogate and true posteriors; hence, the optimization objective in variational inference is minimizing the following KL divergence between p and q, over q

$$KL(q||p) = -\Big(\mathbb{E}_q\big(\log p(\theta, \mathcal{D})\big) - \mathbb{E}_q(\log q)\Big) + \log p(\mathcal{D}) \tag{5.2}$$

in equation 5.2, $\mathbb{E}$ represents the expectation operator, $\mathbb{E}_q\big(\log p(\theta, \mathcal{D})\big) - \mathbb{E}_q(\log q)$ is called evidence lower bound ($ELBO$) and $\log p(\mathcal{D})$ is an unknown constant; Therefore equation 5.2 can be re-written as follows

$$KL(q||p) = -ELBO + c \tag{5.3}$$

where c denotes the unknown constant value ($\log p(\mathcal{D})$). Hence, minimizing the KL divergence is equivalent to maximizing the $ELBO$.

In order to apply variational inference, we need derivatives and implementations that are specific to each model. This process is mathematically and computationally challenging. Automatic Differentiation Variational Inference (ADVI) is one of the methods that automates the procedure for obtaining the solutions in variational inference. The inputs in ADVI are the probabilistic model and the observed data, and the outputs are the posterior inferences about the latent variables in the model [149].

The MCMC and VI methods have different properties that make each of them appropriate for different use cases. On one hand, the sampling process of MCMC methods is generally computationally expensive and heavy but instead, it has no bias;

hence, these methods are preferred when obtaining more accurate results is important than the processing time it takes. On the other hand, for the VI methods, although the choice of the surrogate posterior can clearly introduce a bias, their optimization process makes these methods particularly appropriate for very large-scale inference problems where fast computations are the priority. Our ultimate goal is to apply our models for survival prediction from high dimensional data; therefore, we have selected VI for our use case.

We have developed a Bayesian parametric survival network by combining a parametric survival model with a Bayesian neural network. In order to incorporate the censoring to the model loss when optimizing through the variational inference, the log likelihood component of the $ELBO$ is obtained through the following equation derived from equation 2.17 in Chapter 2

$$\mathcal{L}(t,\delta) = \sum_i \Big(\delta_i log\big(f(t_i)\big) + (1-\delta_i) log\big(S(t_i)\big) \Big) \tag{5.4}$$

where $f(t_i)$ and $S(t_i)$, as explained in Chapter 2, are the PDF and survival function at time t_i for the corresponding distribution of time-to-event, respectively. In our model we have assumed a log-logistic distribution for the time-to-event; hence we have considered a log-logistic survival model as the basis of the Bayesian survival neural network, where we are using a Bayesian neural network with two outputs that are the location and scale parameters of the underlying logistic distribution in the log-logistic model. The output of the log-logistic model is the time-to-event distribution (Figure 5.2). Depending on the problem and the dataset, the log-logistic distribution can be replaced by any other parametric survival models, e.g. the Weibull distribution or even the mixture parametric models. The reason we use parametric survival models is that they model the actual time-to-event, enable building confidence/credible intervals over the inferences of time-to-event, and provide more options for estimation of uncertainty in time-to-event prediction. For our data and application, we have

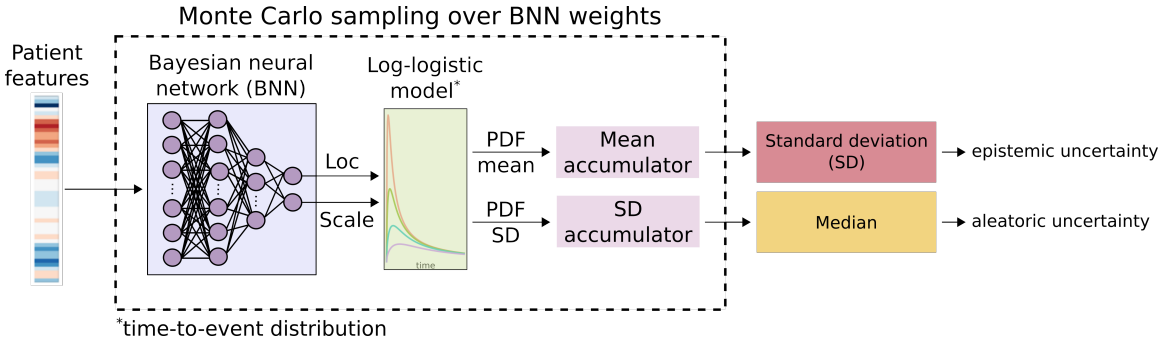

Figure 5.2: The pipeline for Bayesian survival neural network. The patient features are fed into the Bayesian neural network that outputs the location and scale parameters of the underlying logistic distribution in log-logistic model. Log-logistic model is assumed as the basis for time-to-event prediction. Aleatoric and epistemic uncertainties are obtained through a Monte Carlo sampling over the Bayesian neural network's weights and biases. At each round of Monte Carlo sampling, mean and standard deviation of the time-to-event distribution, obtained from log-logistic model, is accumulated; once the Monte Carlo sampling is over, the standard deviation of accumulated mean values is considered as the epistemic uncertainty, and the median of accumulated standard deviations is considered as the aleatoric uncertainty.

opted for log-logistic distribution for the time-to-event, as it appropriately models hazard rate from cancer following diagnosis or treatment where the hazard rate increases initially and decreases later. Besides, we want to allow people with the same mean survival to have different variances (conditional on features). Furthermore, the coupling of mean and variance impacts the modeling of aleatoric uncertainty. Some models have a strong coupling, for example in exponential distribution which is a single parameter model, the mean and variance are strongly coupled as the mean and standard deviation are equal. So, it is important to use a model in which the mean and variance are less coupled.

This model learns and maps the non-linear relations of the covariates to the patients survival times. We use VI to approximate the parameters of this model. Our goal is to quantify the aleatoric and epistemic uncertainties based on this model.

5.3.2 Quantifying uncertainties in Bayesian survival neural networks

Most of the existing approaches for quantifying uncertainties in Bayesian neural networks, measure the epistemic and aleatoric uncertainty separately. In order to capture epistemic uncertainty, they place a prior distribution over the model parameters and measure how the model parameters vary by feeding the data. But, on the other hand, in order to measure the aleatoric uncertainty they corrupt the model output by placing some probability distribution over the output. In case of regression, depending on whether they are modeling the aleatoric uncertainty as either a homoscedastic or heteroscedastic uncertainty, they corrupt the output by either a constant noise for all data points or a Gaussian random noise (generally), respectively. Homoscedastic uncertainty is less of an interest in practice as in reality the uncertainty generally varies for different values of patient features. To capture the heteroscedastic aleatoric uncertainty they measure how the variance of the applied random noise varies over different values of sample (e.g. patient) features [150].

Inspired by [19] our aim in this chapter is to develop a Bayesian neural network for survival prediction that captures both the aleatoric and epistemic uncertainty simultaneously. In order to capture the epistemic uncertainty in this neural network, we put a Gaussian prior distribution over its parameters which converts the traditional neural network with deterministic weights to a Bayesian neural network [151, 152] that has distributions over its weights that follow a Gaussian distribution $\theta \sim \mathcal{N}(0, I)$, where θ represents the weights of the neural network. Then, we measure the epistemic uncertainty by studying the variance of model parameters by altering the number of input samples. In order to simultaneously capture the heteroscedastic aleatoric uncertainty we have to corrupt the output by placing some probability distribution over it. As mentioned before, one of the common approaches is placing a Gaussian random noise and studying its variance over different values of patient

features. However, in tasks such as predicting overall survival for cancer patients Gaussian distribution seems less practical to capture the uncertainty over the predicted survival times. Therefore, we decided to place an output noise in the form of a log-logistic distribution instead. Firstly, this converts the model to a parametric survival model with log-logistic distribution that is proved to be one of the powerful parametric distributions in modeling survival in real-life applications. Secondly, the variance of this noise (in the form of log-logistic distribution) will enable us to capture the underlying heteroscedastic aleatoric uncertainty. The parameters of this parametric survival model or more specifically the location and scale parameters of the underlying logistic distribution is learned through a Bayesian neural network with two outputs.

We use variational inference to approximate the posterior distribution $p(\theta|\mathcal{D})$ in this *Bayesian survival neural network* (BSNN) which learns to predict the survival and captures the epistemic and aleatoric uncertainties simultaneously.

More formally, if the outputs of this Bayesian neural network for each random sampling over the distributions of parameters (θ) are denoted as $f_\theta^\mu(X)$ and $f_\theta^s(X)$ for location and scale parameters of the underlying logistic distribution, respectively, we define the model likelihood as

$$p\big(\log(Y)|f_\theta^\mu(X), f_\theta^s(X)\big) = Logistic\big(f_\theta^\mu(X), f_\theta^s(X)\big) \tag{5.5}$$

where $X = \{x_1, ..., x_N\}$ and $Y = \{y_1, ..., y_N\}$ are the set of inputs (features) and outputs (observed times) in the dataset, respectively. Note that, as explained in Chapter 2, having a logistic distribution for $\log(Y)$ is analogous to having a log-logistic distribution for Y. Based on this model, the aleatoric and epistemic uncertainties are defined as follows

$$\mathcal{A} = \widetilde{f_{\theta_i}^s(X)}, \quad i = 1, ..., M \tag{5.6}$$

$$\mathcal{E} = \sigma\big(f_{\theta_i}^{\mu}(X)\big), \quad i = 1, ..., M \tag{5.7}$$

Where $\mathcal{A}$ and $\mathcal{E}$ are the aleatoric and epistemic uncertainty indicators, respectively. M is the number of samplings done over the parameter space of the Bayesian neural network, $\sigma(.)$ denotes standard deviation, and $\tilde{}$ is the median operator. The whole pipeline for the BSNN model is illustrated in Figure 5.2. This model quantifies the aleatoric and epistemic uncertainties, while predicting the time-to-event. Equation 5.6 quantifies the aleatoric uncertainty (inherent variability in the data) based on the variability in time-to-event, and equation 5.7 quantifies the epistemic uncertainty (underlying uncertainty of model weights and biases) based on the variability of these parameters (weights and biases). In order to validate our hypothesis about the aleatoric and epistemic uncertainties we have generated synthetic survival data. The procedure for generating this synthetic data is explained below.

5.3.3 Generating synthetic survival data

We did an initial validation of our methods on the synthetic data because it gives freedom in altering the number of samples, adding or reducing the features, and even perturbing the features to do a more fine-grained analysis and validation of methods. Besides, the simulation with data that consists of known properties of interest is an essential step in developing and validating statistical methods. However, simulating survival data is more challenging in comparison to most of the simulation tasks. One of the key factors to consider when generating survival data is the generation process of censoring times. In particular, when generating synthetic survival data to use in randomized trials, it is critical to make sure the non-informative censoring assumption is not violated. Censoring is called non-informative when the reason for censoring is not related to the event of interest. Therefore, the non-informative

censoring assumption is violated when the censoring criteria is related to the failure process of subjects. When the non-informative censoring assumption is violated, the standard statistical methods may give invalid inferences [153, 154]. In order to produce synthetic survival data of size N, where N is the number of synthetic samples, we first generate N samples for each random covariate X_i from independent continuous uniform distributions $\mathcal{U}$.

$$X_i \sim \mathcal{U}(c_1^{(i)}, c_2^{(i)}), \quad -\infty < c_1^{(i)} < c_2^{(i)} < \infty \tag{5.8}$$

where $c_1^{(i)}, c_2^{(i)} \in \mathbb{R}$ are the randomly selected intervals for i^{th} covariate. Also, we generate two random coefficients β_μ and β_s for each of the covariates from another set of independent continuous uniform distributions. These coefficients will be the corresponding coefficients for the location μ and scale s parameters of the underlying logistic distribution that will be used to generate the survival times in the next step.

$$\beta_{i,\mu} \sim \mathcal{U}(l_1^{(i)}, l_2^{(i)}), \quad -\infty < l_1^{(i)} < l_2^{(i)} < \infty \tag{5.9}$$

$$\beta_{i,s} \sim \mathcal{U}(s_1^{(i)}, s_2^{(i)}), \quad -\infty < s_1^{(i)} < s_2^{(i)} < \infty \tag{5.10}$$

$\beta_{i,\mu}$ and $\beta_{i,s}$ are the coefficients of the location and scale parameters corresponding to the i^{th} covariate, respectively. Then, the *location* (μ) and *scale* (s) parameters of underlying logistic distribution is obtained as $\mu = \beta_\mu^T X$ and $s = \beta_s^T X$.

Next, the initial synthetic survival times T are obtained by sampling from a log-logistic distribution as follows

$$\log(T) \sim Logistic(\mu, s) \tag{5.11}$$

To apply non-informative censoring to these times and to simulate the randomness

in real-world censoring phenomenon, first independent random censoring times (C) are sampled from an exponential distribution as follows

$$C \sim Exp(1/\mu) \tag{5.12}$$

Then, the initial event indicator for these samples is assigned as followed

$$\delta_0 = \begin{cases} 0 & \text{if } C < T \\ 1 & \text{otherwise} \end{cases} \tag{5.13}$$

Next, c samples are randomly sampled without replacement from a discrete uniform distribution $c \sim \mathcal{U}\{1, N\}$ where N is the total number of generated samples in synthetic data. The event indicator based on this uniform sampling process (δ_u) is assigned to all N samples as follows

$$\delta_u = \begin{cases} 0 & \text{if } n \in c \\ 1 & \text{otherwise} \end{cases} \tag{5.14}$$

Then, the final event indicator for the samples in this synthetic data is defined as follows

$$\delta = \delta_0 \cap \delta_u \tag{5.15}$$

Any sample with $\delta = 0$ in the synthetic data is a censored sample and all others are uncensored. This whole process of random censoring reasonably guarantees that the non-informative censoring assumption is *not* violated.

5.4 Results

5.4.1 Synthetic data

In order to validate the methods, we have generated synthetic survival data with one feature, with the same procedure explained in the previous section. Incorporating larger number of features do not change the interpretation behind the results, but only makes it harder to visualize, so we opted for a single feature experiments in order to validate our hypothesis. The distribution of our single feature synthetic data with $20,000$ samples is illustrated in Figure 5.3. In this dataset approximately 30.37% of samples are censored.

In our experiments, we manipulated the distribution of training samples to study the epistemic and aleatoric uncertainties on the test set. In these experiments we gradually extended the training samples to more areas and studied how it changes the aleatoric and epistemic uncertainties. Based on our results in Figure 5.4 we observed that areas with less training samples get higher epistemic uncertainty (red area between curves). Also, we observed that areas with higher noise get higher aleatoric uncertainty (yellow area between curves). Furthermore, we observed that the epistemic uncertainty decreases when increasing the number of training samples in sparse areas (Figures 5.4 and 5.5), in other hand the distribution of the train set does not change the aleatoric uncertainty, and it remains roughly constant for different train set distributions (Figures 5.4 and 5.6). These results are obtained through 100 samplings over the weights and biases of the Bayesian neural network and clearly illustrate the expected behaviours for aleatoric and epistemic uncertainties.

5.4.2 Survival prediction for glioma patients

In this section we have applied our Bayesian survival neural network to quantify uncertainty in predicting glioma patients' overall survival from their protein expression

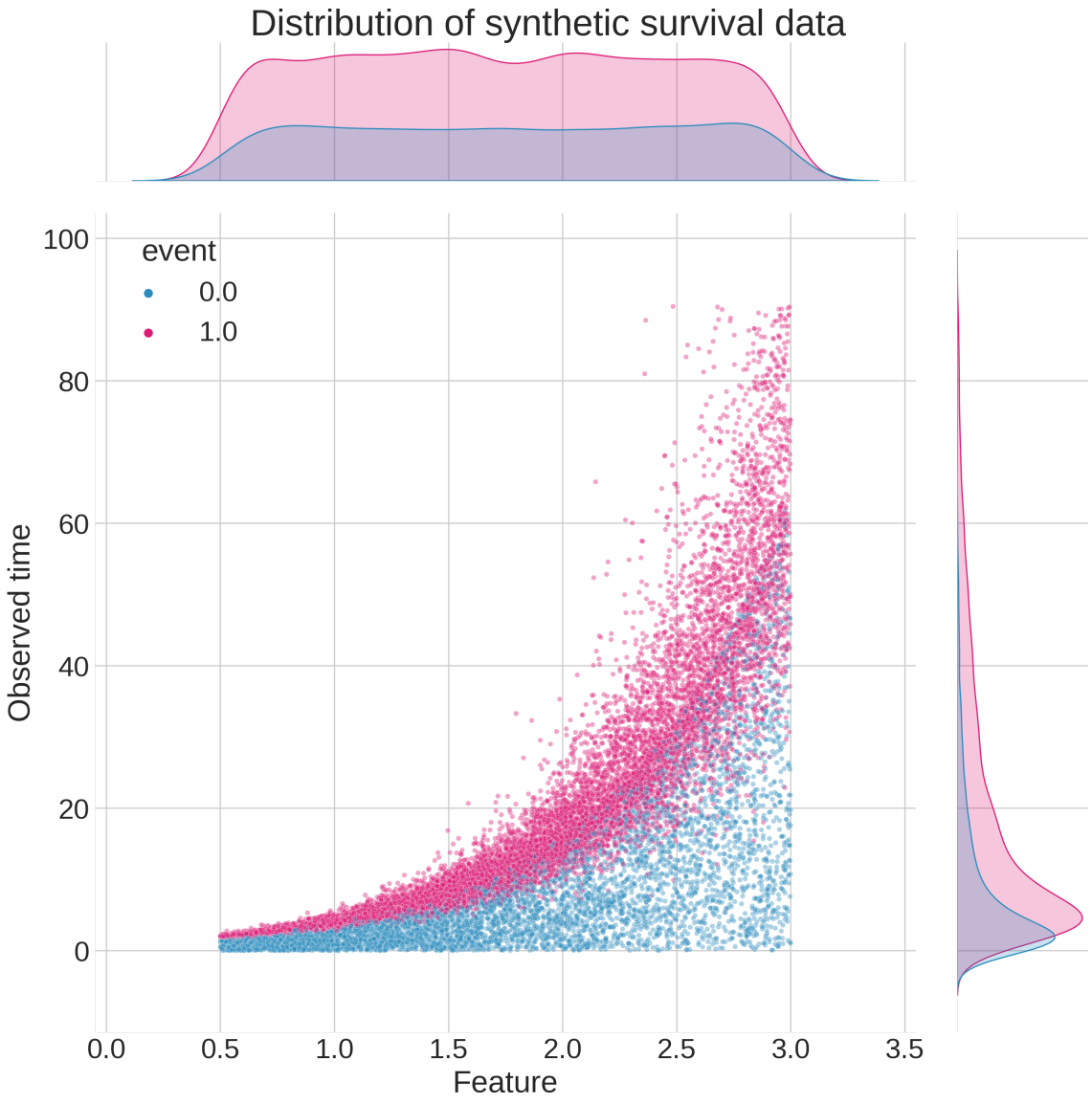

Figure 5.3: Distribution of the generated synthetic survival data of $20,000$ samples. Pink points show the uncensored samples and blue points show the censored samples. $\sim 30.37\%$ of samples are censored.

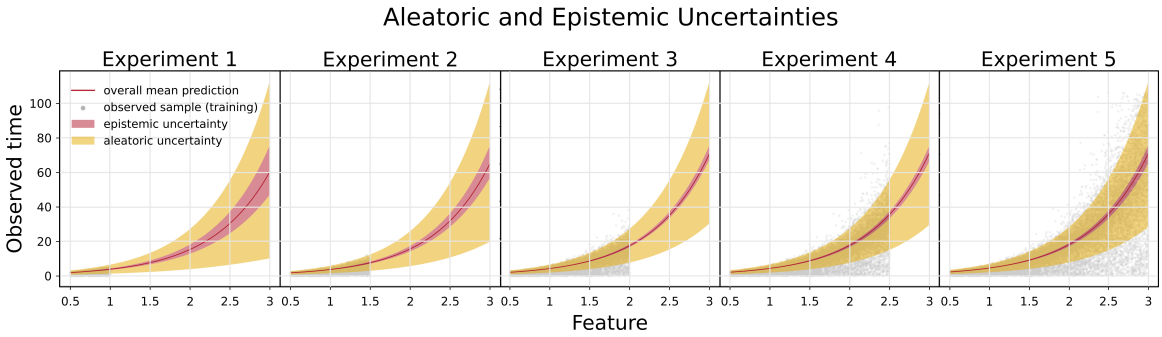

Figure 5.4: Aleatoric and epistemic uncertainties for test set when trained on different train set distributions. Increasing the number of train set samples from Experiment 1 (left) to Experiment 5 (right). Areas with no training samples get higher epistemic uncertainty. Areas with higher feature noise in train set get higher aleatoric uncertainty.

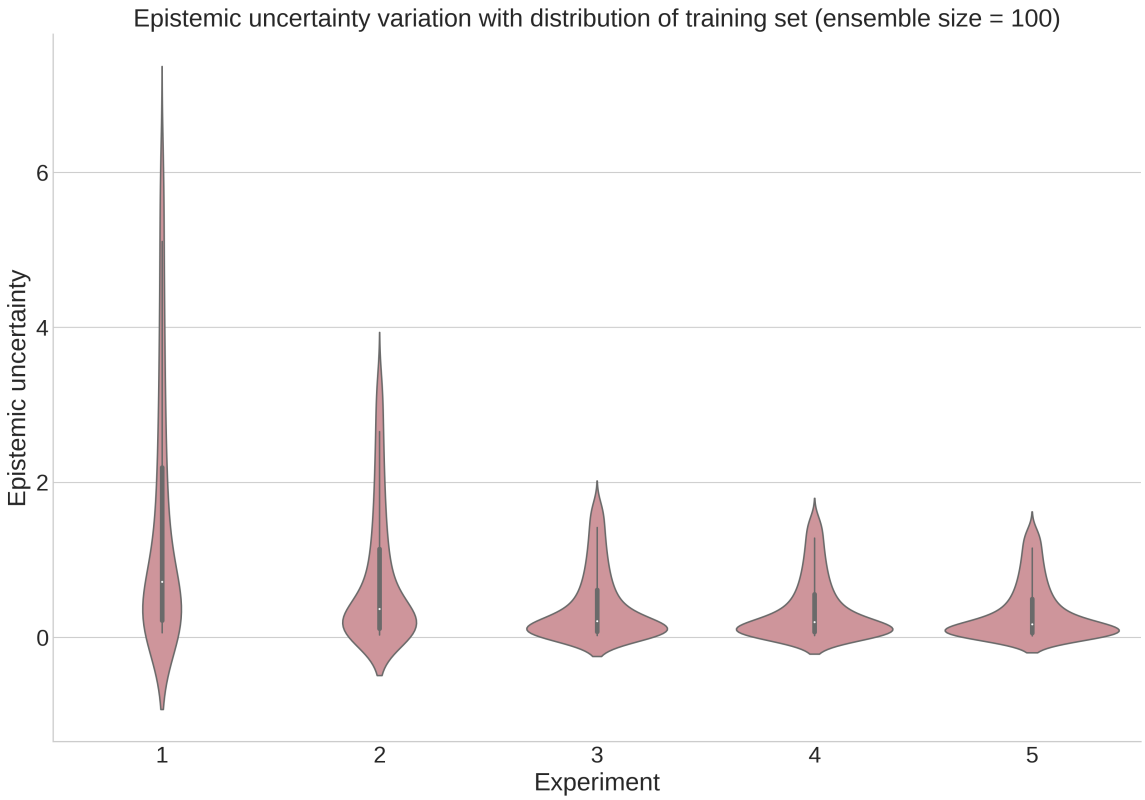

Figure 5.5: Synthetic data - Epistemic uncertainty decreases by gradually increasing the number of samples in areas where there was no training sample.

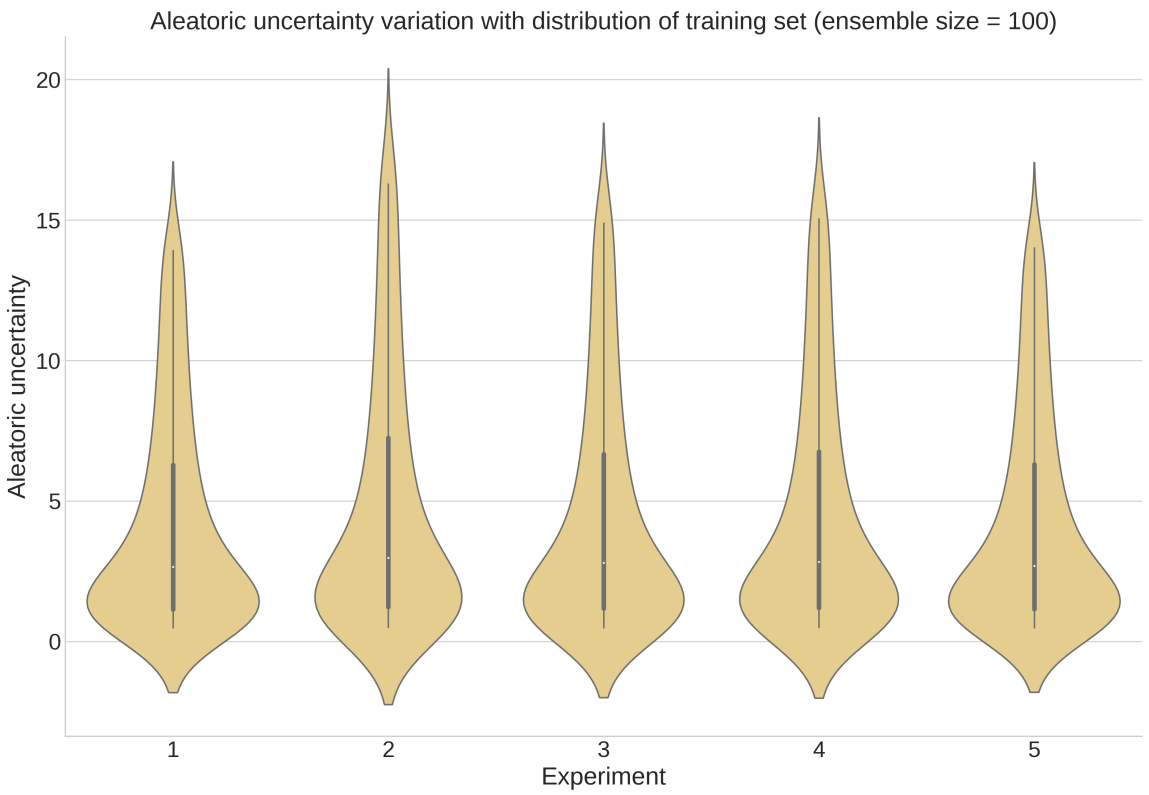

Figure 5.6: Synthetic data - Distribution of the train set does not change the aleatoric uncertainty.

Table 5.1: Summary of dataset clinical features

Characteristic	Total, n=595	Molecular subtype		
		Astrocytoma IDH WT, n=267 (45%)	Astrocytoma IDH mutant, n=196 (33%)	Oligodendroglioma, n=132 (22%)
WHO histologic grade (%)				
II	186 (31)	17 (6)	97 (49)	72 (55)
III	200 (34)	52 (20)	88 (45)	60 (45)
IV	209 (35)	198 (74)	11 (6)	N/A[a]
Age at diagnosis, y				
Range	14-88	18-88	14-73	17-75
Median	49 ± 15.9	59 ± 14.7	36.5 ± 11.3	45 ± 13.1
Sex, female (%)	254 (43)	113 (42)	86 (44)	55 (42)

[a]N/A, not applicable.

data. For this purpose, we obtained protein expression and clinical follow-up data for 595 patients from glioma data generated by The Cancer Genome Atlas (TCGA) Research Netwrok (https://www.cancer.gov/tcga). This dataset comprises lower-grade gliomas (WHO grades II and III) and glioblastomas (WHO grade IV), contains both astrocytomas and oligodendrogliomas, and has overall survivals ranging from less than 1 to 14 years or more. A summary of demographics, grades, and molecular subtypes for this cohort is presented in Table 5.1. We did a random stratified split of patients into training (70%), validation (15%), and testing (15%) sets. After selecting the hyperparameters of the model, the training and validation sets were merged to build the ultimate training set (85%) and the model performance was reported on the held-out test set (15%). We used histologic grade, IDH mutation, codeletion of 1p/19q chromosomes, and event indicator (censorship status) as our stratification criteria when splitting the data to train, validation, and test sets. This stratification process guarantees getting similar cohorts inside each split.

We performed different experiments to study the variation of aleatoric and epistemic uncertainties when predicting survival for glioma patients. In these experiments, we gradually increased the number of training samples (each time by 20% of the total number of training samples) and quantified the epistemic and aleatoric uncertainties on the test set. Based on our results in Figure 5.7 we observed that similar

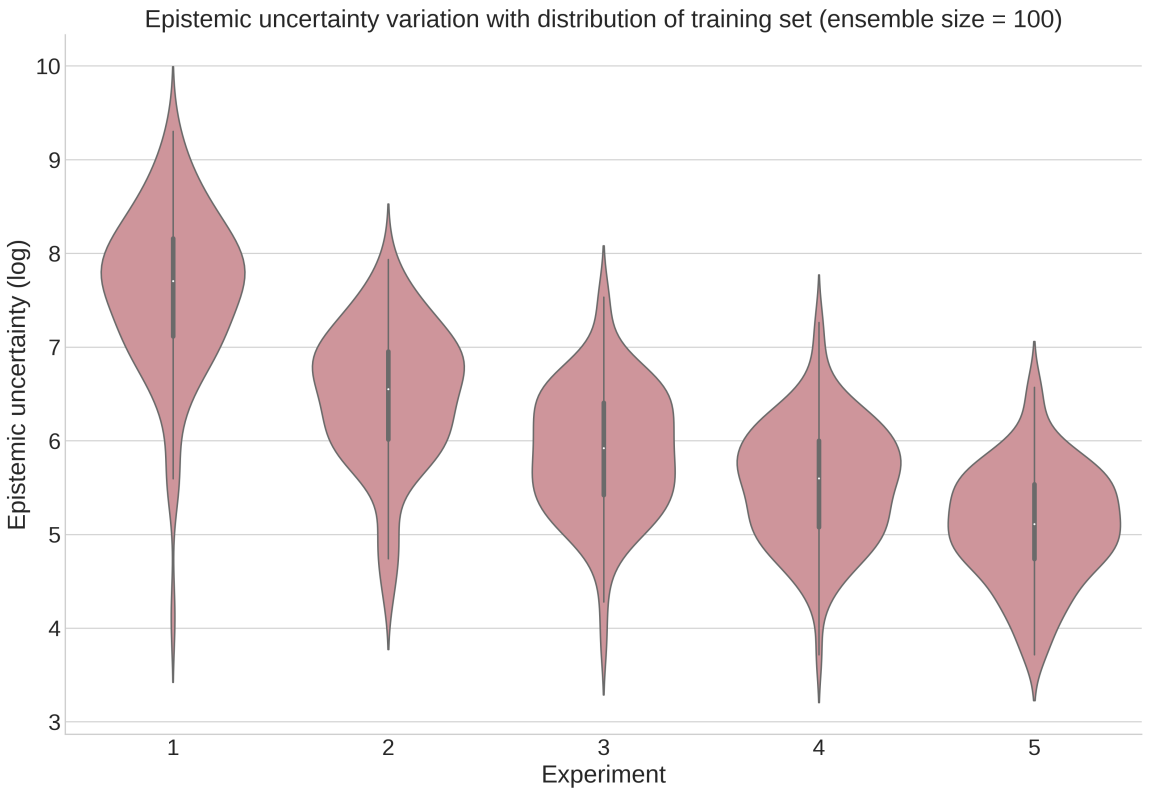

Figure 5.7: Epistemic uncertainty decreases by gradually increasing the number of samples in train set. Increasing the number of samples from Experiment 1 to 5, each time by 20% of the total number of training samples.

to our expectation, the epistemic uncertainty decreases by increasing the number of training samples from experiment 1 to 5. Furthermore, we observed that changing the number of training samples does not change the aleatoric uncertainty, and it remains roughly constant for different number of samples in train set, similar to what we are expecting for aleatoric uncertainty (Figure 5.8). These results are obtained through 100 samplings over the model's parameter space and clearly illustrate the expected behaviours for aleatoric and epistemic uncertainties. For a more clear visualization and comparison of underlying uncertainties in different experiments, we are illustrating the logarithm of the epistemic and aleatoric uncertainty indicators in Figures 5.7 and 5.8.

Also, in order to do a more in-depth analysis of the uncertainties, we compared

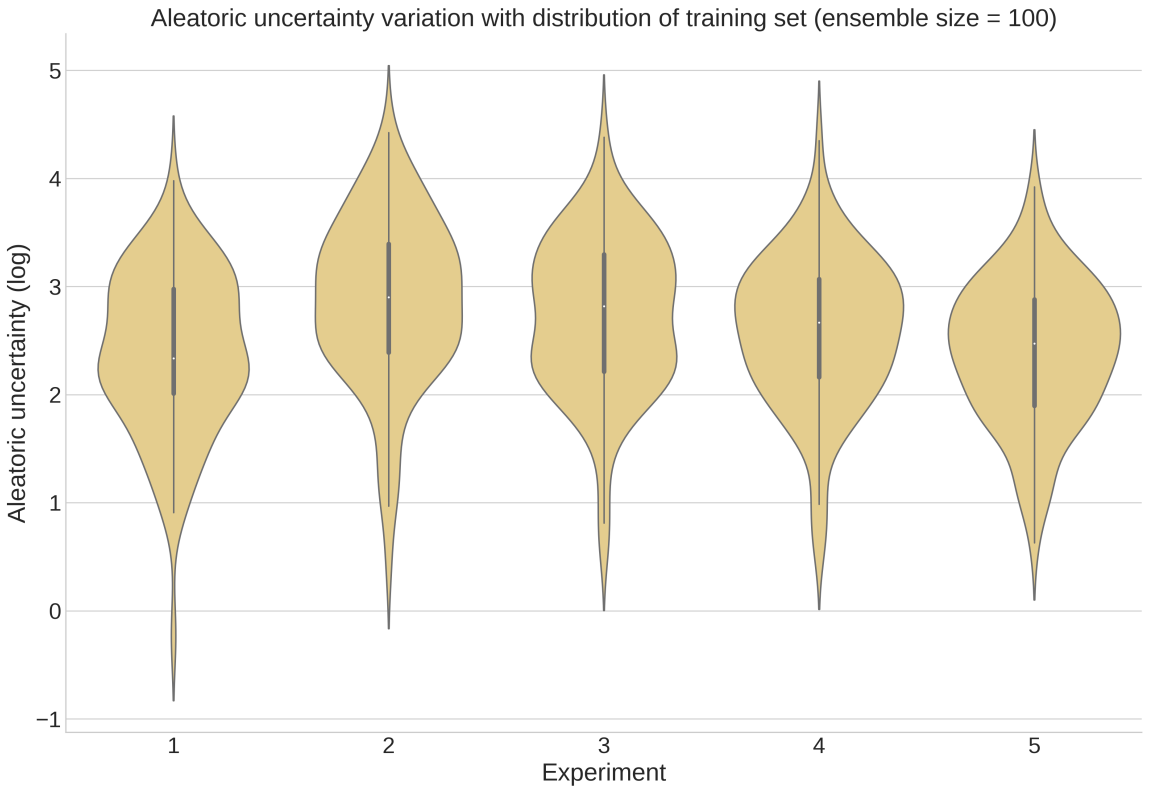

Figure 5.8: The aleatoric uncertainty remains roughly constant by changing the number of samples in train set. Increasing the number of samples from Experiment 1 to 5, each time by 20% of the total number of training samples.

the uncertainty levels between three different subtypes for a single model. The KM curves for these different molecular subtypes of glioma are shown in Figure 5.9. Figures 5.10 and 5.11 illustrate how the epistemic and aleatoric uncertainties vary over three different molecular subtypes. Based on Table 5.1, the IDH-wildtype astrocytoma has larger number of samples compared to the IDH-mutant astrocytoma and oligodendroglioma; so we are expecting lower epistemic uncertainty for this subtype which is evident from the lower values of the corresponding red box in Figure 5.10. In other hand, higher range of survival times, along with higher percentage of censored samples for oligodendroglioma ($\sim 88\%$ censored in train set) and IDH-mutant astrocytoma ($\sim 80\%$ censored in train set) acts as having more noisy features and introduces more aleatoric uncertainty for these two subtypes compared to the IDH-wildtype astrocytoma that has lower percentage of censored samples ($\sim 30\%$ censored in train set) and a limited variation of survival times (mostly focused on about < 1500 as opposed to IDH-mutant astrocytoma and oligodendroglioma where survival times are expanded smoothly from 0 to > 4000) as evident from corresponding curves in Figure 5.9.

We also evaluated the Kolmogorov-Smirnov statistic (KS statistic) on the test set, which is illustrated in Figure 5.12. This is a tool used in gene set enrichment analysis, where a curve below the straight line suggests enrichment of low uncertainty samples and above the straight line indicates enrichment of high uncertainty samples. In order to obtain the KS statistic we rank samples by uncertainty (aleatoric or epistemic, separately), from lowest uncertainty to highest, and then look at the cumulative distribution of each molecular subtype in this ranked list.

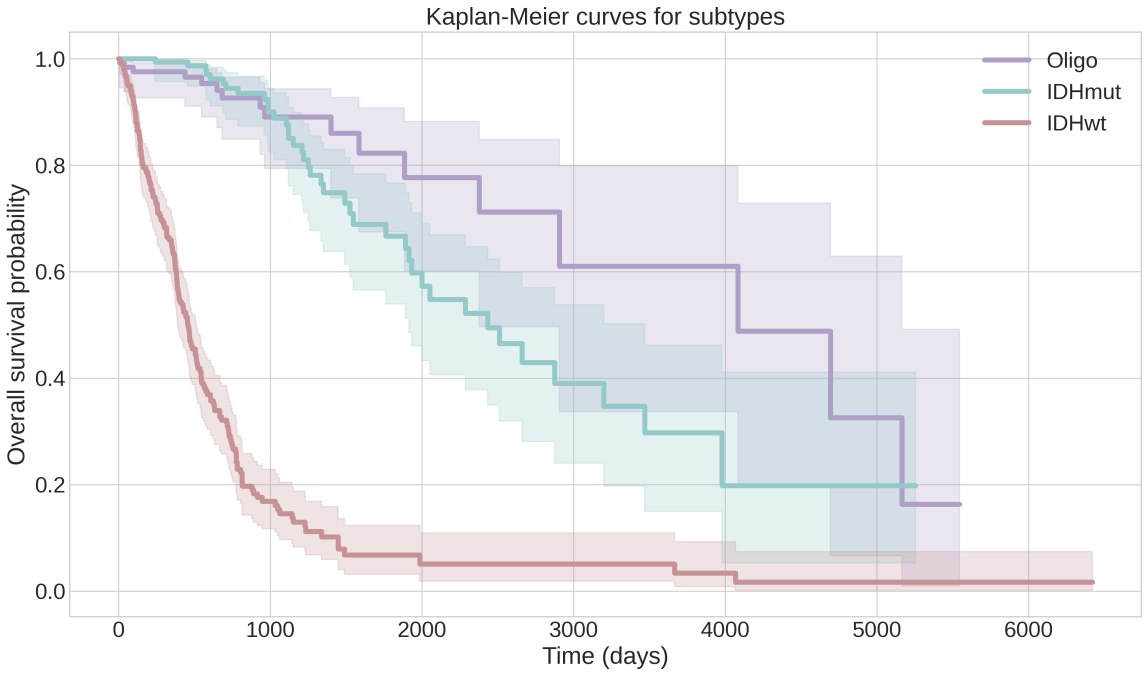

Figure 5.9: Kaplan-Meier curves for oligodendroglioma (Oligo), IDH-mutant astrocytoma (IDHmut), and IDH-wildtype astrocytoma (IDHwt).

5.5 Conclusion and future work

In this chapter we have looked at the formulation of survival analysis using parametric models and Bayesian neural networks. We have developed a Bayesian survival neural network that predicts the survival and estimates the aleatoric and epistemic uncertainties. To validate our formulation of aleatoric and epistemic uncertainties we generated a synthetic survival data with non-informative censoring. Our results on the synthetic data validates our formulation of epistemic and aleatoric uncertainties at Bayesian survival neural networks, and suggests that our choice of parametric distribution for the survival models has a direct impact on the quality of the modeled aleatoric uncertainty. Because, based on our experimental results aleatoric uncertainty in these models is directly proportional to the the standard deviation of the underlying survival distribution. We validated our hypothesis about the quantification of uncertainties by doing experiments on the synthetic survival data. These

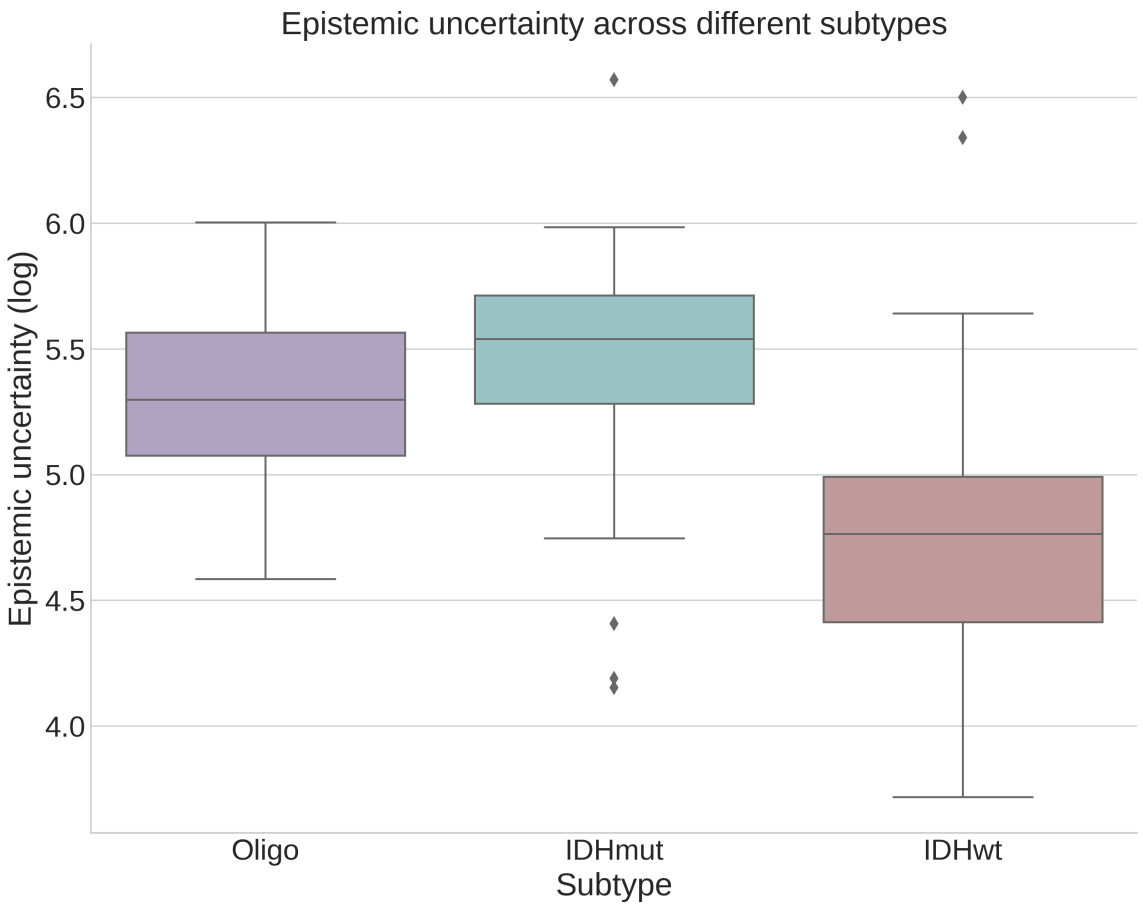

Figure 5.10: Epistemic uncertainty across three different subtypes. Oligodendroglioma (Oligo), IDH-mutant astrocytoma (IDHmut), and IDH-wildtype astrocytoma (IDHwt). Among all the available training samples $\sim$ 22% are oligodendroglioma, $\sim$ 33% are IDHmut-astrocytoma, and $\sim$ 45% are IDHwt-astrocytoma.

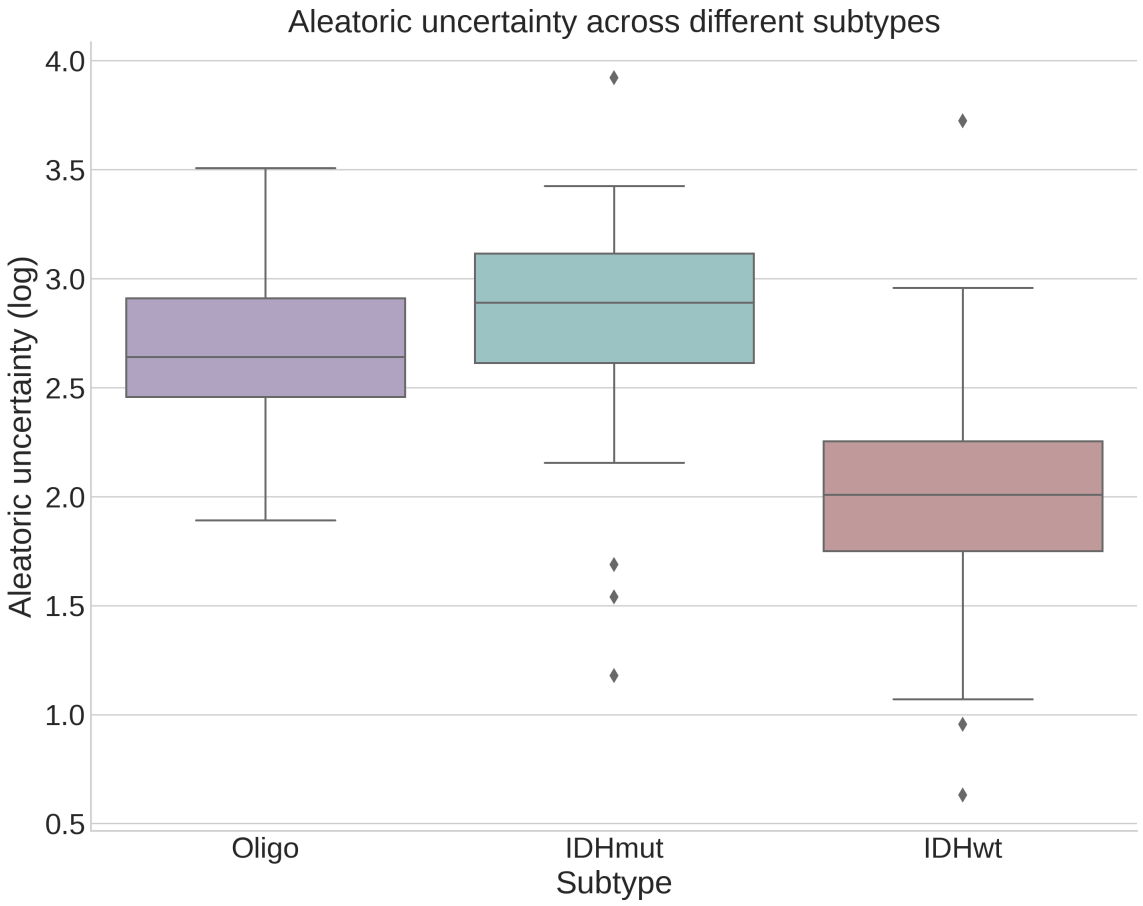

Figure 5.11: Aleatoric uncertainty across three different subtypes. Oligodendroglioma (Oligo), IDH-mutant astrocytoma (IDHmut), and IDH-wildtype astrocytoma (IDHwt). $\sim 88\%$ of training samples in oligodendroglioma, $\sim 80\%$ of training samples in IDHmut-astrocytoma, and $\sim 30\%$ of training samples in IDHwt-astrocytoma are censored.

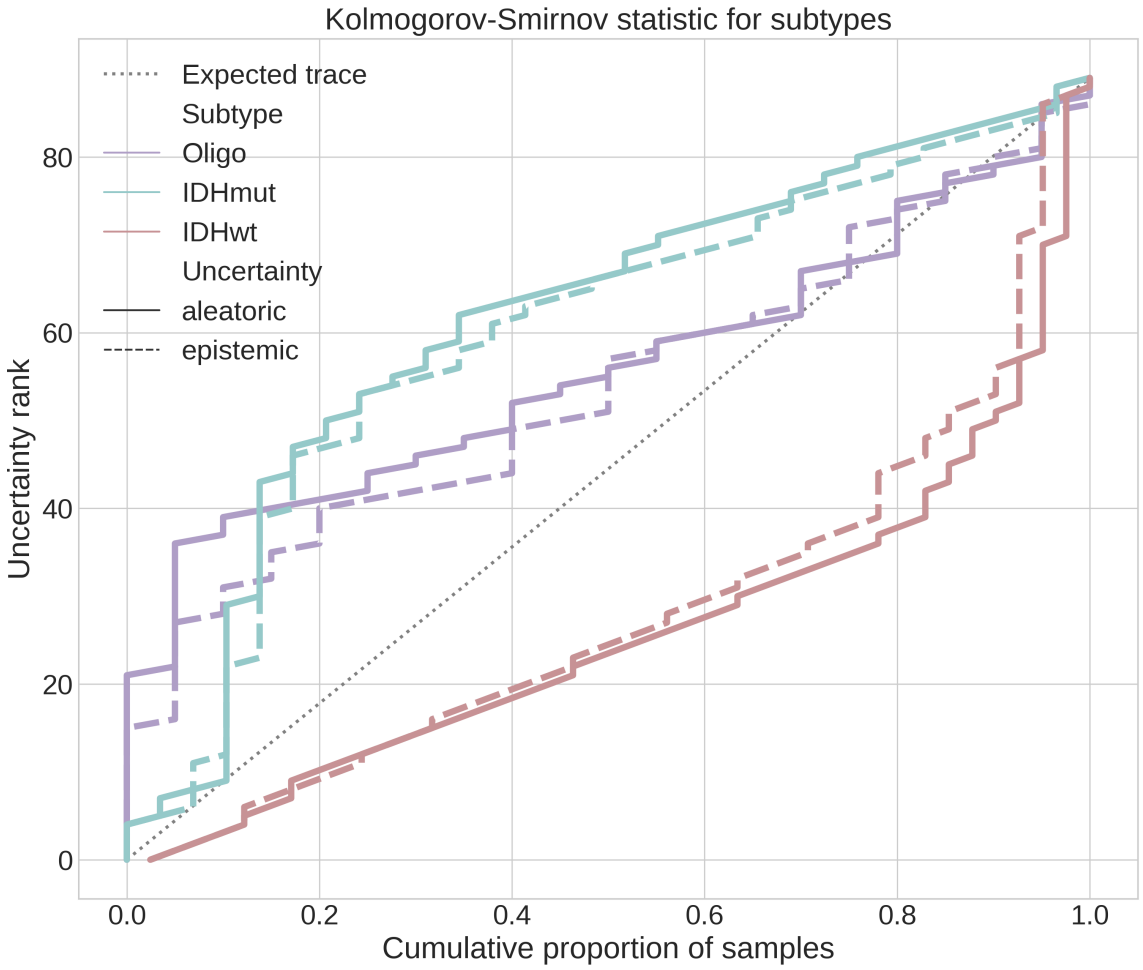

Figure 5.12: Kolmogorov-Smirnov statistic across three different subtypes for test set. Oligodendroglioma (Oligo), IDH-mutant astrocytoma (IDHmut), and IDH-wildtype astrocytoma (IDHwt). Solid lines: ranked samples by aleatoric uncertainty. Dotted lines: ranked samples by epistemic uncertainty.

results illustrates that our model captures the underlying aleatoric and epistemic uncertainties while predicting the survival. This is important when building models to predict survival for patients, as depending on the underlying uncertainty it has the potential to guide us through additional appropriate measurements or sample incorporation to improve the model performance and prediction results. Bayesian survival neural network points toward the future of survival models, where the model reliably measures the uncertainty while predicting the survival times, recognizes the source for uncertainty and asks for intervention from human specialist when needed. In future, we are planning to use this model to detect out of distribution samples, and to leverage this model to decide on incorporating more measurements or more samples to improve the model performance.

Acknowledgements

The results on glioma patinets' data are based upon glioma data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

Bibliography

[1] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[2] Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

[3] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[4] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 2424–2433, 2016.

[5] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[6] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[7] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6:26286, 2016.

[8] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7:46450, 2017.

[9] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[10] Riku Turkki, Nina Linder, Panu E Kovanen, Teijo Pellinen, and Johan Lundin. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *Journal of pathology informatics*, 7, 2016.

[11] Dmitrii Bychkov, Riku Turkki, Caj Haglund, Nina Linder, and Johan Lundin. Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer. In *Medical Imaging 2016: Digital Pathology*, volume 9791, page 979115. International Society for Optics and Photonics, 2016.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[20] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[21] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[22] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8(1):1–11, 2018.

[23] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[24] Chaoyang Yan, Kazuaki Nakane, Xiangxue Wang, Yao Fu, Haoda Lu, Xiangshan Fan, Michael D Feldman, Anant Madabhushi, and Jun Xu. Automated gleason grading on prostate biopsy slides by statistical representations of homology profile. *Computer Methods and Programs in Biomedicine*, 194:105528, 2020.

[25] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Rus-

lan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[26] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[27] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *arXiv preprint arXiv:1806.03335*, 2018.

[28] Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS Workshop on Bayesian Deep Learning*, volume 192, 2016.

[29] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[30] Riku Turkki, Dmitrii Byckhov, Mikael Lundin, Jorma Isola, Stig Nordling, Panu E Kovanen, Clare Verrill, Karl von Smitten, Heikki Joensuu, Johan Lundin, et al. Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast cancer research and treatment*, 177(1):41–52, 2019.

[31] Dai Feng and Lili Zhao. Bdnnsurv: Bayesian deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:2101.03170*, 2021.

[32] Hrushikesh Loya, Pranav Poduval, Deepak Anand, Neeraj Kumar, and Amit Sethi. Uncertainty estimation in cancer survival prediction. *arXiv preprint arXiv:2003.08573*, 2020.

[33] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce, and Vincent L Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

[34] Waloddi Weibull et al. A statistical distribution function of wide applicability. *Journal of applied mechanics*, 18(3):293–297, 1951.

[35] Peter R Fisk. The graduation of income distributions. *Econometrica: journal of the Econometric Society*, pages 171–185, 1961.

[36] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[37] Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

[38] Bernard Altshuler. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11, 1970.

[39] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[40] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[41] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

[42] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[43] Philippe Lambert, Dave Collett, Alan Kimber, and Rachel Johnson. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in medicine*, 23(20):3177–3192, 2004.

[44] Niels Keiding, Per Kragh Andersen, and John P Klein. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine*, 16(2):215–224, 1997.

[45] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[46] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[47] Ting Chen and Christophe Chefd'Hotel. Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer, 2014.

[48] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[49] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[50] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[51] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[52] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[53] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[54] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *stat*, 1050(2), 2016.

[55] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

[56] Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *International Conference on Machine Learning*, pages 1049–1057. PMLR, 2013.

[57] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[58] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[59] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[60] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[61] Michelle Ntampaka, Hy Trac, Dougal J Sutherland, Sebastian Fromenteau, Barnabás Póczos, and Jeff Schneider. Dynamical mass measurements of contaminated galaxy clusters using machine learning. *The Astrophysical Journal*, 831(2):135, 2016.

[62] Siamak Ravanbakhsh, Junier Oliva, Sebastian Fromenteau, Layne Price, Shirley Ho, Jeff Schneider, and Barnabás Póczos. Estimating cosmological parameters from the dark matter distribution. In *International Conference on Machine Learning*, pages 2407–2416. PMLR, 2016.

[63] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR, 2013.

[64] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[65] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[67] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *arXiv preprint arXiv:1703.06114*, 2017.

[68] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Deep learning with sets and point clouds. *arXiv preprint arXiv:1611.04500*, 2016.

[69] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

[70] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 1681–1691, 2015.

[71] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

[72] Andrew Gardner, Neshat Elhami, and Rastko R Selmic. Classifying unordered feature sets with convolutional deep averaging networks. In *2019 IEEE Interna-*

tional Conference on Systems, Man and Cybernetics (SMC), pages 3447–3453. IEEE, 2019.

[73] Alexander Richard and Juergen Gall. A bag-of-words equivalent recurrent neural network for action recognition. *Computer Vision and Image Understanding*, 156:79–91, 2017.

[74] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[75] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

[76] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[77] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[78] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[79] Jun Kong, Olcay Sertel, Kim L Boyer, Joel H Saltz, Metin N Gurcan, and Hiroyuki Shimada. Computer-assisted grading of neuroblastic differentiation. *Archives of pathology & laboratory medicine*, 132(6):903–904, 2008.

[80] M Khalid Khan Niazi, Keluo Yao, Debra L Zynger, Steven K Clinton, James Chen, Mehmet Koyutürk, Thomas LaFramboise, and Metin Gurcan. Visually meaningful histopathological features for automatic grading of prostate cancer. *IEEE journal of biomedical and health informatics*, 21(4):1027–1038, 2016.

[81] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287. IEEE, 2008.

[82] Jian Ren, Evita T Sadimin, Daihou Wang, Jonathan I Epstein, David J Foran, and Xin Qi. Computer aided analysis of prostate histopathology images gleason grading especially for gleason score 7. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3013–3016. IEEE, 2015.

[83] Sonal Kothari, John H Phan, Andrew N Young, and May D Wang. Histological image classification using biologically interpretable shape-based features. *BMC medical imaging*, 13(1):1–17, 2013.

[84] Olcay Sertel, Jun Kong, Hiroyuki Shimada, Umit V Catalyurek, Joel H Saltz, and Metin N Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern recognition*, 42(6):1093–1103, 2009.

[85] Mohammad Faizal Ahmad Fauzi, Michael Pennell, Berkman Sahiner, Weijie Chen, Arwa Shana'ah, Jessica Hemminger, Alejandro Gru, Habibe Kurt, Michael Losos, Amy Joehlin-Price, et al. Classification of follicular lymphoma:

the effect of computer aid on pathologists grading. *BMC medical informatics and decision making*, 15(1):1–10, 2015.

[86] M Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011.

[87] Jun Kong, Lee AD Cooper, Fusheng Wang, Jingjing Gao, George Teodoro, Lisa Scarpace, Tom Mikkelsen, Matthew J Schniederjan, Carlos S Moreno, Joel H Saltz, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PloS one*, 8(11):e81049, 2013.

[88] David A Gutman, Jake Cobb, Dhananjaya Somanna, Yuna Park, Fusheng Wang, Tahsin Kurc, Joel H Saltz, Daniel J Brat, Lee AD Cooper, and Jun Kong. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data. *Journal of the American Medical Informatics Association*, 20(6):1091–1098, 2013.

[89] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[91] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classifica-

tion with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[92] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[93] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[94] David E Reuss, Yasin Mamatjan, Daniel Schrimpf, David Capper, Volker Hovestadt, Annekathrin Kratz, Felix Sahm, Christian Koelsche, Andrey Korshunov, Adriana Olar, et al. Idh mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for who. *Acta neuropathologica*, 129(6):867–873, 2015.

[95] Whitney B Pope, James Sayre, Alla Perlina, J Pablo Villablanca, Paul S Mischel, and Timothy F Cloughesy. Mr imaging correlates of survival in patients with high-grade gliomas. *American Journal of Neuroradiology*, 26(10):2466–2474, 2005.

[96] Anthony Michael Carter. Placental oxygen consumption. part i: in vivo studies—a review. *Placenta*, 21:S31–S37, 2000.

[97] RW Redline. Placental pathology: a systematic approach with clinical correlations. *Placenta*, 29:86–91, 2008.

[98] T Yee Khong, Eoghan E Mooney, Ilana Ariel, Nathalie CM Balmus, Theonia K Boyd, Marie-Anne Brundler, Hayley Derricott, Margaret J Evans, Ona M Faye-Petersen, John E Gillan, et al. Sampling and definitions of placental lesions: Amsterdam placental workshop group consensus statement. *Archives of pathology & laboratory medicine*, 140(7):698–713, 2016.

[99] Janet M Catov, Matthew F Muldoon, Steven E Reis, Roberta B Ness, Lananh N Nguyen, J-M Yamal, Hyunsoo Hwang, and W Tony Parks. Preterm birth with placental evidence of malperfusion is associated with cardiovascular risk factors after pregnancy: a prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(8):1009–1017, 2018.

[100] Karen K Mestan, Jennifer Check, Lucy Minturn, Sushmita Yallapragada, Kathryn N Farrow, Xin Liu, Emily Su, Nicolas Porta, Nina Gotteiner, and Linda M Ernst. Placental pathologic changes of maternal vascular underperfusion in bronchopulmonary dysplasia and pulmonary hypertension. *Placenta*, 35(8):570–574, 2014.

[101] Eve Blair, Jan de Groot, and Karin B Nelson. Placental infarction identified by macroscopic examination and risk of cerebral palsy in infants at 35 weeks of gestational age and over. *American journal of obstetrics and gynecology*, 205(2):124–e1, 2011.

[102] David JP Barker, Johan G Eriksson, Tom Forsén, and Clive Osmond. Fetal origins of adult disease: strength of effects and biological basis. *International journal of epidemiology*, 31(6):1235–1239, 2002.

[103] Rajesh Kumar, Yunxian Yu, Rachel E Story, Jacqueline A Pongracic, Ruchi Gupta, Colleen Pearson, Kathryn Ortiz, Howard C Bauchner, and Xiaobin Wang. Prematurity, chorioamnionitis, and the development of recurrent wheezing: a prospective birth cohort study. *Journal of Allergy and Clinical Immunology*, 121(4):878–884, 2008.

[104] Drucilla J Roberts. Placental pathology, a survival guide. *Archives of pathology & laboratory medicine*, 132(4):641–651, 2008.

[105] Chen-Chih J Sun, Vania O Revell, Anthony J Belli, and Rose M Viscardi.

Discrepancy in pathologic diagnosis of placental lesions. *Archives of pathology & laboratory medicine*, 126(6):706–709, 2002.

[106] Lee AD Cooper, Alexis B Carter, Alton B Farris, Fusheng Wang, Jun Kong, David A Gutman, Patrick Widener, Tony C Pan, Sharath R Cholleti, Ashish Sharma, et al. Digital pathology: Data-intensive frontier in medical imaging. *Proceedings of the IEEE*, 100(4):991–1003, 2012.

[107] MR Jackson, TM Mayhew, and PA Boyd. Quantitative description of the elaboration and maturation of villi from 10 weeks of gestation to term. *Placenta*, 13(4):357–370, 1992.

[108] Eric Jauniaux and Graham J Burton. Pathophysiology of placenta accreta spectrum disorders: a review of current findings. *Clinical obstetrics and gynecology*, 61(4):743–754, 2018.

[109] Rashmi Mukherjee. Morphometric evaluation of preeclamptic placenta using light microscopic images. *BioMed research international*, 2014, 2014.

[110] Mudher Al-Adnani, Andreas Marnerides, Simi George, Alia Nasir, and Martin A Weber. "delayed villous maturation" in placental reporting: concordance among consultant pediatric pathologists at a single specialist center. *Pediatric and Developmental Pathology*, 18(5):375–379, 2015.

[111] Gitta Turowski and Martin Vogel. Re-view and view on maturation disorders in the placenta. *Apmis*, 126(7):602–612, 2018.

[112] JK Grether, A Eaton, R Redline, R Bendon, K Benirschke, and K Nelson. Reliability of placental histology using archived specimens. *Paediatric and perinatal epidemiology*, 13(4):489–495, 1999.

[113] Tracy A Manuck, Madeline Murguia Rice, Jennifer L Bailit, William A Grobman, Uma M Reddy, Ronald J Wapner, John M Thorp, Steve N Caritis, Mona Prasad, Alan TN Tita, et al. Preterm neonatal morbidity and mortality by gestational age: a contemporary cohort. *American journal of obstetrics and gynecology*, 215(1):103–e1, 2016.

[114] Robin B Kalish and Frank A Chervenak. Sonographic determination of gestational age. *The ultrasound review of obstetrics and Gynecology*, 5(4):254–258, 2005.

[115] Robin B Kalish, Howard T Thaler, Stephen T Chasen, Meruka Gupta, Seth J Berman, Zev Rosenwaks, and Frank A Chervenak. First-and second-trimester ultrasound assessment of gestational age. *American journal of obstetrics and gynecology*, 191(3):975–978, 2004.

[116] Pekka Taipale and Vilho Hiilesmaa. Predicting delivery date by ultrasound and last menstrual period in early gestation. *Obstetrics & Gynecology*, 97(2):189–194, 2001.

[117] Alexander Maly, Gal Goshen, Jona Sela, Alexander Pinelis, Michael Stark, and Bella Maly. Histomorphometric study of placental villi vascular volume in toxemia and diabetes. *Human pathology*, 36(10):1074–1079, 2005.

[118] Katherine Leavey, Samantha J Benton, David Grynspan, Shannon A Bainbridge, Eric K Morgen, and Brian J Cox. Gene markers of normal villous maturation and their expression in placentas with maturational pathology. *Placenta*, 58:52–59, 2017.

[119] Sina Salsabili, Anika Mukherjee, Eranga Ukwatta, Adrian DC Chan, Shannon Bainbridge, and David Grynspan. Automated segmentation of villi

in histopathology images of placenta. *Computers in biology and medicine*, 113:103420, 2019.

[120] Zaneta Swiderska-Chadaj, Tomasz Markiewicz, Robert Koktysz, and Szczepan Cierniak. Image processing methods for the structural detection and gradation of placental villi. *Computers in biology and medicine*, 100:259–269, 2018.

[121] Michael Ferlaino, Craig A Glastonbury, Carolina Motta-Mejia, Manu Vatish, Ingrid Granne, Stephen Kennedy, Cecilia M Lindgren, and Christoffer Nellåker. Towards deep cellular phenotyping in placental histology. *arXiv preprint arXiv:1804.03270*, 2018.

[122] Elisheva D Shanes, Leena B Mithal, Sebastian Otero, Hooman A Azad, Emily S Miller, and Jeffery A Goldstein. Placental pathology in covid-19. *American journal of clinical pathology*, 154(1):23–32, 2020.

[123] Alexa A Freedman, Jeffery A Goldstein, Gregory E Miller, Ann Borders, Lauren Keenan-Devlin, and Linda M Ernst. Seasonal variation of chronic villitis of unknown etiology. *Pediatric and Developmental Pathology*, 23(4):253–259, 2020.

[124] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[125] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3. Citeseer, 2003.

[126] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet:

A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[127] Julian K Christians and David Grynspan. Placental villous hypermaturation is associated with improved neonatal outcomes. *Placenta*, 76:1–5, 2019.

[128] H Pinar, CJ Sung, CE Oyer, and DB Singer. Reference values for singleton and twin placental weights. *Pediatric Pathology & Laboratory Medicine*, 16(6):901–907, 1996.

[129] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[130] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[131] Brandon T Larsen, Maxwell L Smith, Brett M Elicker, Jessica M Fernandez, Guillermo A Arbo-Oze de Morvil, Carlos AC Pereira, and Kevin O Leslie. Diagnostic approach to advanced fibrotic interstitial lung disease: bringing together clinical, radiologic, and histologic clues. *Archives of pathology & laboratory medicine*, 141(7):901–915, 2017.

[132] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[133] Daniel Clymer, Stefan Kostadinov, Janet Catov, Lauren Skvarca, Liron Pantanowitz, Jonathan Cagan, and Philip LeDuc. Decidual vasculopathy identification in whole slide images using multiresolution hierarchical convolutional neural networks. *The American Journal of Pathology*, 190(10):2111–2122, 2020.

[134] Somak Roy, Liron Pantanowitz, Milon Amin, Raja R Seethala, Ahmed Ishtiaque, Samuel A Yousem, Anil V Parwani, Ioan Cucoranu, and Douglas J Hartman. Smartphone adapters for digital photomicrography. *Journal of pathology informatics*, 5, 2014.

[135] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[136] Yinyin Yuan. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor perspectives in medicine*, 6(8):a026583, 2016.

[137] Dovile Zilenaite, Allan Rasmusson, Renaldas Augulis, Justinas Besusparis, Aida Laurinaviciene, Benoit Plancoulaine, Valerijus Ostapenko, and Arvydas Laurinavicius. Independent prognostic value of intratumoral heterogeneity and immune response features by automated digital immunohistochemistry analysis in early hormone receptor-positive breast carcinoma. *Frontiers in oncology*, 10:950, 2020.

[138] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):1–8, 2017.

[139] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[140] Alexis Bellot and Mihaela Van der Schaar. A hierarchical bayesian model for

personalized survival predictions. *IEEE journal of biomedical and health informatics*, 23(1):72–80, 2018.

[141] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[142] RM Neal. Bayesian training of backpropagation networks by the hybrid monte carlo method (tech. rep. crg-tr-92-1). *Toronto, Canada: Department of Computer Science, University of Toronto*, 1992.

[143] Christopher M Bishop. Bayesian methods for neural networks. 1995.

[144] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[145] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[146] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[147] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[148] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[149] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[150] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3):4, 2016.

[151] John S Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, pages 853–859, 1990.

[152] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[153] Dan Jackson, Ian R White, Shaun Seaman, Hannah Evans, Kathy Baisley, and James Carpenter. Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in medicine*, 33(27):4681–4694, 2014.

[154] Jeffrey J Harden and Jonathan Kropko. Simulating duration data for the cox model. *Political Science Research and Methods*, 7(4):921–928, 2019.