UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SURVIVAL PREDICTION OF PEDIATRIC LEUKEMIA UNDER MODEL UNCERTAINTY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

YUNING CUI
Norman, Oklahoma
2024

SURVIVAL PREDICTION OF PEDIATRIC LEUKEMIA UNDER MODEL UNCERTAINTY

A DISSERTATION APPROVED FOR SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Rui Zhu, Chair

Dr. Yifu Li, Co-chair

Dr. Chongle Pan

Dr. Talayeh Razzaghi

Dr. Shivakumar Raman

TABLE OF CONTENT

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Rui Zhu. Whenever I encountered challenges, Dr. Rui Zhu offered me timely clarity and directions. I also appreciate your encouragement and patience when I was overwhelmed by my research and studies during my doctoral journey. It is my great honor to be one of your students. Thanks for your invaluable guidance, support, and patience throughout my doctoral research.

To my co-advisor, Dr. Yifu Li, thanks for providing professional suggestions for my research work. I appreciate each meeting with you. It has deepened my understanding of machine learning and deep learning, which has greatly enhanced my research work.

To my committee members, Dr. Shivakumar Raman, Dr. Chongle Pan, and Dr. Talayeh Razzaghi, thanks for all the advice and support during my general exam, research proposal, and doctoral defense. Your guidance on leukemia understanding, mathematical modeling, and data information has provided me with valuable insights, significantly enhancing my research during my doctoral journey.

To my research collaborators, Dr. Stephanie R. Brown, Dr. Rachel E. Gallant, Dr. Amanda E. Janitz, Dr. Hanumantha R. Pokala, and Weixuan Dong, thank you for providing professional suggestions on leukemia-related understanding. Your advice has become an essential part of my research work.

Additionally, I would like to thank Dr. Kash Barker for the guidance on course selection and for reminding me of each significant milestone throughout my Ph.D. journey. I also appreciate your reminder email regarding seminars, which offer me opportunities to learn from other academic fields each week.

To Melodi Franklin and Jennifer Ille, thank you for helping me address complicated issues, including business traveling, room reservation, printing, etc.

Furthermore, I would like to thank my family for their unconditional support and companionship throughout my doctoral studies. They encouraged me when I faced difficulties and were there to share in my joy during happy times.

Finally, I am grateful to the National Cancer Institute (NCI) for providing access to the Surveillance, Epidemiology, and End Results (SEER) database for the research work in section 2, which was published in Computers in Biology and Medicine in 2023. I appreciate the support from all the collaborators. Additionally, I thank the funding support from the American Cancer Society (ACS) through the Institutional Research Grant numbered 134128-IRG-19-142-01 and the support from the Oklahoma Shared Clinical and Translational Resources, funded by the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences (NIGMS) under grant number U54GM104938.

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

Abstract

This dissertation addresses critical challenges in survival prediction for pediatric leukemia, particularly Acute Lymphoblastic Leukemia (ALL), by introducing novel predictive models that incorporate Bayesian principles and advanced machine learning and deep learning techniques. Recognizing the complexity and heterogeneity of leukemia, our research emphasizes the need for precise and individualized predictions that factor in the recurrence and survival probability, two pivotal aspects that significantly influence treatment outcomes in children and adolescents. In the first segment, we introduce a Bayesian survival model that diverges from traditional survival analysis by integrating full Bayesian inference, providing more accurate patient-specific survival predictions that account for model uncertainty. This allows for more confident decision-making in clinical settings. The second part of our work proposes a Transformer-based deep survival model that not only predicts the time to event but also employs Shapley Additive explanations (SHAP) for model interpretability, shedding light on how clinical variables influence predictions. Further, we propose a Bayesian Transformer-based survival model that combines the feature extraction capabilities of the Transformer encoder with a Bayesian Neural Network (BNN) layer. This model outputs recurrence probabilities under model uncertainty. The outputs from the Transformer encoder are sued for K-means clustering to evaluate model performance. Our work demonstrates strong potential in survival and recurrence prediction for children with leukemia, providing robust predictive survival models for clinicians to offer efficient and effective medical care to pediatric leukemia patients.

CHAPTER 1

Introduction

Childhood leukemia is a type of cancer that originates in the white blood cells within the bone marrow's hematopoietic stem cells. As the malignant cells proliferate, they impede the bone marrow's ability to produce numbers of normal white blood cells, red blood cells, and platelets. It is recognized that childhood leukemia represents the most prevalent form of cancer in children, constituting 31% of all pediatric cancers in the United States. Beyond the severe morbidity it causes, childhood leukemia is responsible for almost 39% of all deaths from cancer in children. Treatment in the healthcare setting refers to the medical or surgical management of a health condition. There is also a focus on early intervention, which encompasses the services and support provided to children who are at risk of developing such health conditions. While recent advances in treatment have improved survival rates for childhood leukemia, the formulation of early intervention strategies is less advanced. These strategies are crucial for averting life-threatening incidents and minimizing long-term consequences after treatment. Acute lymphoblastic leukemia (ALL) is a fast-moving malignancy of the white blood cells that causes the bone marrow to overproduce immature cells by upsetting the usual environment. These cells disrupt the immune system, increasing the risk of infection and potentially catastrophic events that often go untreated. Recent cancer statistics also show that ALL is the most common type of leukemia among children and adolescents, accounting for 76% of children and adolescents with leukemia in the United States. Among all potential clinical attributes of ALL, recurrence remains a significant factor that affects ALL-induced survival. Children with recurrence of ALL mostly have a lower

survival probability compared to others. Therefore, predicting the recurrence of ALL becomes urgent as healthcare practitioners can use prediction. Additionally, the mortality rate associated with childhood leukemia is not uniform across different demographics, often due to disparities in access to oncology care resources. Implementing a predictive strategy for preventive care could potentially enhance the survival rates for leukemia more substantially and bridge these gaps. Therefore, it is imperative to develop an accurate, cost-effective predictive method that leverages commonly collected leukemia data within current monitoring systems. Such an approach would enable proactive healthcare for at-risk children and contribute to improving their chances of survival.

In the medical field, survival analysis is employed to estimate the time until an event of interest, such as death or disease recurrence. It's viewed as a statistical method that assesses the likelihood of patients with certain conditions surviving within a specific timeframe. Survival analysis can determine these survival probabilities at individual time points based on the clinical characteristics of the patients. Many clinical practices have integrated survival analysis to forecast outcomes on a patient-by-patient basis over the past years. Yet, conventional techniques like Cox regression models presuppose certain survival time distributions, assumptions that are frequently invalid due to the extended treatment duration for diseases like Acute Lymphoblastic Leukemia (ALL), leading to models that do not perform optimally in practical scenarios. Furthermore, other survival prediction models, such as Support Vector Machines, Random Survival Forests, and deep learning-based neural networks, often require intensive computation, which can hinder their speed and efficiency. These methods also face limitations in acknowledging long-term temporal relationships and in assigning significance to inputs during prediction, which may diminish the accuracy and

interpretability of the survival predictions they generate. Most importantly, model uncertainty is mostly neglected.

In this dissertation, we present three works to address the limitations of existing works. First, we propose a Bayesian survival model to make patient-specific survival predictions for children with leukemia. This paper proposes a Bayesian survival model that integrates a backbone survival model with Bayesian inference to take model uncertainty into account. Specifically, we estimate the posterior distribution of model parameters using a full Bayesian inference approach. By considering model uncertainty, the proposed Bayesian survival model can provide an accurate patient-specific survival prediction. Second, we develop a Transformer-based explainable deep survival model, aiming to predict the time to the occurrence of death for children and adolescents with ALL. Specifically, we first train the Transformer-based deep survival model by minimizing a loss function describing the discrepancy between predicted and actual survival status. Second, we use the well-trained model to predict the survival probability at a particular point in time. Lastly, we utilize Shapley Additive explanations (SHAP) to explain the proposed model and quantify the contributions of clinical variables to predictive results in global and local views. Third, we use the Transformer encoder to feature extracting and a BNN layer to output recurrence probabilities by taking model uncertainty into account. In addition, the outputs from the Transformer encoder are clustered into two groups through the K-means algorithm to assess model performance.

The outline of this dissertation is as follows. Chapter 2 introduces a published paper named “Bayesian Inference for Survival Prediction for Children with Leukemia.” Chapter 3 presents a work called “An Explainable Transformer-based Deep Survival Model for

Survival Prediction of Childhood Acute Lymphoblastic Leukemia." Chapter 5 presents the study called "A Bayesian Transformer-based Deep Survival Model for Recurrence Prediction of Childhood Acute Lymphoblastic Leukemia." Chapter 6 concludes the dissertation and provides future directions.

CHAPTER 2

Bayesian Inference for Survival Prediction of Childhood Leukemia

2.1 Introduction

Childhood Leukemia is a malignancy of white blood cells that begins in the hematopoietic stem cells in the bone marrow. As the cancer cells increase, the bone marrow can no longer make adequate numbers of normal white blood cells, red blood cells, and platelets [1]. It is noted that childhood Leukemia is the most common cancer among children, accounting for 31% of childhood cancers in the United States [2]. In addition to the painful morbidity, the high mortality rate of childhood Leukemia also contributes to nearly 39% of cancer-induced childhood deaths [3]. In the medical domain, treatment is the action or way of treating a health condition medically or surgically. Additionally, the process of providing services and support to children at risk for the health condition is called early intervention. Nowadays, advancements in treatment have increased the survivability of childhood Leukemia. However, the development of early intervention strategies remains sketchy, and these strategies are essential to prevent occurrences of life-threatening events and long-term post-treatment sequelae [4]. Despite numerous efforts to improve Leukemia survival, there remains a subgroup of children dying from Leukemia. Moreover, childhood Leukemia mortality is unevenly distributed among groups due to cancer care resource disparity [5]. A predictive approach to providing preventive care to children may improve Leukemia survival to a greater extent and mitigate these disparities. Hence, there is an urgent need to conduct an accurate and low-cost predictive approach, preferably utilizing the existing monitoring

system with commonly monitored attributes for Leukemia, to provide children preventive care and improve survival rate.

In the medical domain, survival analysis predicts the time to a certain event for downstream decision-making [6]. A broad of illnesses utilize survival analysis to make patient-specific survival predictions. For example, predicting the occurrence of Alzheimer's using a deep learning model [7]. In the well-known Framingham Heart Study, a Cox regression model is used to predict the 30-year risk of cardiovascular disease [8]. There is also research discussing the racial differences in the survival of breast cancer [9]. The study illustrates that the risk of dying from breast cancer is greater among Black patients than among White patients by controlling certain variables (e.g., geographic site and age). Moreover, the results show that Black patients continue to demonstrate a slightly increased risk of death after adjusting for other variables (e.g., stage and treatment). Additionally, survival analysis is applied to identify key factors affecting the survival rate of children with primary malignant brain tumors [10]. These studies have demonstrated the importance of survival analysis in the medical domain as it provides predictive results to improve outcomes. However, as described above, these patient-specific predictions rely on a single model and focus on improving model performance. Nonetheless, an ensemble of models with approximately identical performance can exhibit wide variability in predictions. This is called model uncertainty. A patient-specific prediction is less accurate if it counts on a single model with model uncertainty ignored. Hence, it calls for a survival model taking model uncertainty into account to predict patient-specific survival probabilities of childhood Leukemia, thereby providing well-informed decision-making on prognosis and treatments.

Bayesian inference is a powerful tool to capture model uncertainty in survival analysis.

Conventionally, most existing Bayesian survival models have prohibitively long run times, leading to significant computational costs. Moreover, inappropriate sampling methods result in highly correlated samples and are less efficient in sampling high-dimensional parameters. Most importantly, hand-actuated tunning is always required, which is also time-consuming. To overcome these hurdles, we develop a novel Bayesian survival model through full Bayesian inference. The implemented sampling method is relatively faster and more efficient in sampling high-dimensional distributions. Additionally, a specific sampler is induced, which is capable of tuning certain parameters automatically. Through the developed Bayesian survival model, we aim to predict the time to the occurrence of death for children with Leukemia. First, we develop a survival model to serve as the basis of Bayesian inference. Second, we place prior distributions over various model parameters and estimate their posterior distributions with full Bayesian inference. Third, we predict the survival probability at a certain time point by considering model uncertainty induced by posterior distribution. Experimental results show that the proposed Bayesian survival model is effective and efficient in predicting the survival probability of childhood Leukemia. This present investigation is conducive to providing preventive care and reducing Leukemia-related life-threatening events.

This paper is outlined as follows. Section 2.2 describes various existing survival models, model uncertainty applications, and Bayesian inference approaches. Section 2.3 presents the research methodology of Bayesian survival analysis. Section 2.4 evaluates and validates the proposed methodology with experimental results. Section 2.5 concludes the research and presents future research directions.

2.2 Research Background

Existing survival analysis based on traditional statistics either fails to make patient-specific predictions or make assumptions on the distribution of survival time. Moreover, these predictions rely on a single model while ignoring model uncertainty. To account for model uncertainty and make patient-specific survival predictions in survival modeling, a Bayesian inference approach is involved. By taking the idea from physical science, the proposed sampling method obtains the target distribution more accurately. Therefore, more accurate patient-specific survival prediction outcomes are obtained.

2.2.1 Survival Modeling in Clinical Applications

Survival analysis is a typical statistical methodology to model time-to-event data, aiming to predict the time to a certain event [11]. One primary interest in survival analysis is survival probability, which describes the probability that a patient survives beyond a certain time point [12]. By obtaining the survival probabilities of patients, healthcare practitioners could identify whether patients are at risk of a certain disease or not. Therefore, effective services and support could be provided to those patients who are at risk, thereby preventing occurrences of life-threatening events. In the medical domain, researchers estimate the survival probabilities of patients with certain diseases through various survival models. These models have been widely utilized and are of great importance in the medical domain, even though there remain challenges to address.

Specific models in survival analysis can be classified into three categories based on traditional statistics, such as the no-parametric model, the parametric model, and the semi-parametric model. First, a non-parametric model relies on mathematical inference without requiring assumptions on distributions of data. A representative is the Kaplan-Meier

estimator [13]. This method obtains survival probabilities at a certain time point solely through counting statistics. Although the Kaplan-Meier estimator provides a generalized view of survival probability, it is incapable of performing patient-specific predictions on survival probabilities. Moreover, it fails in multivariate analysis as it only identifies the effect of one clinical attribute. Second, a parametric model (e.g., accelerated failure time model) assumes that the survival time follows a particular distribution, such as Weibull and Gamma distributions [14]. Parametric models could provide suboptimal results as the distribution assumptions are violated. Third, another type of model is called the semi-parametric model, such as the cox proportional hazard (cox PH) regression model [15]. Cox PH regression model does not make assumptions on survival time distributions during the modeling period. Moreover, it could be utilized to make patient-specific survival predictions in multivariate analysis.

In addition to the above limitations of survival models based on traditional statistics, predictions of all these three kinds of models rely on a single best model. “A best model” is defined as a model providing the most accurate results. In predictive analysis, model parameters optimization plays an important role in obtaining the best model. Some methods apply a series of significant tests (e.g., Log Rank test in Kaplan-Meier method). Others use certain estimation methods (e.g., Maximum Likelihood Estimation in Cox PH regression model). However, by selecting a single model, predictions are conditioned on the selected model, and model uncertainty is ignored. As mentioned in Section 1, model uncertainty is imperative to patient-specific predictions in survival analysis. A patient-specific prediction could be less accurate if the analysis counts on a single model and neglects model uncertainty. Therefore, it is essential to involve model uncertainty in patient-specific survival prediction.

2.2.2 Model Uncertainty in Predictive Modeling

Model uncertainty can be seen as uncertainty about the true function that underlies the observed process [16]. It can be regarded as disagreement of the outcome predictive distributions conditioned on certain inputs [17]. In a certain model, given certain inputs and model parameters, the model function leads to a certain output. However, a distribution over model functions with certain inputs is induced by a distribution of model parameters. Different model functions contribute to different predictive distributions. Therefore, a distribution over model functions yields a distribution of predictive distributions.

Historically, extensive research has explored the importance of inducing model uncertainty in modeling. For example, a previous study has captured model uncertainty in weather and climate prediction [18]. This study discusses the importance of multi-model ensembles. It illustrates that an ensemble system is more reliable than a single model. In addition, model uncertainty in risk analysis [19] discussed distinctions between classical and Bayesian approaches to risk models. In machine learning and deep learning field, model uncertainty is also discussed, such as quantifying model uncertainty in groundwater storage change and language modeling [20,21]. These works explain model uncertainty explicitly, contributing much to increasing the model's accuracy.

In the medical domain, one objective is to predict the patient-specific survival probability. Prediction outcomes vary by training the model multiple times as the variability comes from different sets of optimal model parameters. Accordingly, it is problematic for practitioners to offer well-informed decision-making on prognosis and treatment. However, such a challenge could be addressed by regarding model parameters as a distribution, which also considers model uncertainty and produces more precise patient-specific predictions. In

this paper, we propose a Bayesian survival model to predict patient-specific survival probabilities of children with Leukemia, accounting for model uncertainty through a full Bayesian inference approach (detailed in Section 2.3).

2.2.3 Bayesian Inference and Posterior Sampling

Bayesian inference is an essential statistic tool applying Bayes' theorem to update the distribution of a hypothesis, where a hypothesis is also called a prior distribution [22]. In terms of predictive analysis, the intractable distribution of model parameters leads to model uncertainty. Bayesian inference aims to obtain a posterior distribution of model parameters by placing a prior distribution, accounting for model uncertainty. Accordingly, predictions are based on the posterior distribution of model parameters instead of a set of optimal model parameters, which captures model uncertainty and yields more precise outcomes.

Sampling posterior is an essential part of Bayesian inference. Broadly speaking, there are two main streams of basic sampling methods. One is called independent sampling (e.g., rejection sampling), which requires few samples to obtain target distribution but meets challenges in scaling high dimension parameters. The other is called dependent sampling. A representative method of dependent sampling is Markov Chain Monte Carlo (MCMC), which comprises a class of algorithms to sample a posterior distribution [23]. Specifically, through MCMC, one obtains samples at various successive states of a constructed time-discrete Markov Chain. The more states are included, the closer the distribution of the samples is to the posterior distribution. A posterior distribution is obtained when the MCMC converges to a stationary distribution. Here, a stationary distribution is defined as a distribution on the state space of the chain that is preserved by the transition function [24]. MCMC is a widely utilized sampling method as its ability to sample high-dimensional

probability distributions. For example, MCMC sampling methods have been applied to determine optimal models for earth science problems [25]. Additionally, existing research utilizes MCMC to address non-negative source separation [26]. These works have demonstrated the significance of MCMC sampling.

Conventionally, there is a broad of methods being used to construct a Markov Chain. For example, existing research discusses variable selection in building a multiple regression model [27]. This study claims that key predictor subsets are those with high posterior probability. By using MCMC, computational burden is alleviated. In addition, another research constructs a Markov Chain to sample paths according to a given distribution of a network, which contributes much to route choice and guidance [28]. However, these MCMC sampling methods could produce highly correlated samples and are less efficient in sampling a high-dimensional distribution. Moreover, the random walk algorithm does not provide specific information on how to propose samples at a new state. Therefore, it is of great importance to propose a more efficient approach to sampling model parameters to address these challenges. In this paper, we investigate an alternative approach to sampling a joint distribution of induced auxiliary parameters and model parameters taking the idea from physical science. By acknowledging the derivatives of the auxiliary variables and model parameters with respect to time, the proposed method is capable of informing specific directions to the target joint distribution. Once the values of these parameters are stable, auxiliary variables are marginalized from the joint distribution to obtain the target distribution of the model parameters. The proposed method takes advantage of scaling better to a high-dimensional distribution. Moreover, the high acceptance rate of new samples secures the efficiency of the sampling process. Most importantly, inducing auxiliary

variables and using the information in the gradients better control the sampling process, thereby directing toward the target distribution more accurately.

2.3 Research Methodology

As shown in Figure 1, this paper proposes a Bayesian survival model which integrates a backbone survival model with Bayesian inference to take model uncertainty into account. Specifically, we estimate the posterior distribution of model parameters using a full Bayesian inference approach. By considering model uncertainty, the proposed Bayesian survival model can provide an accurate patient-specific survival prediction.

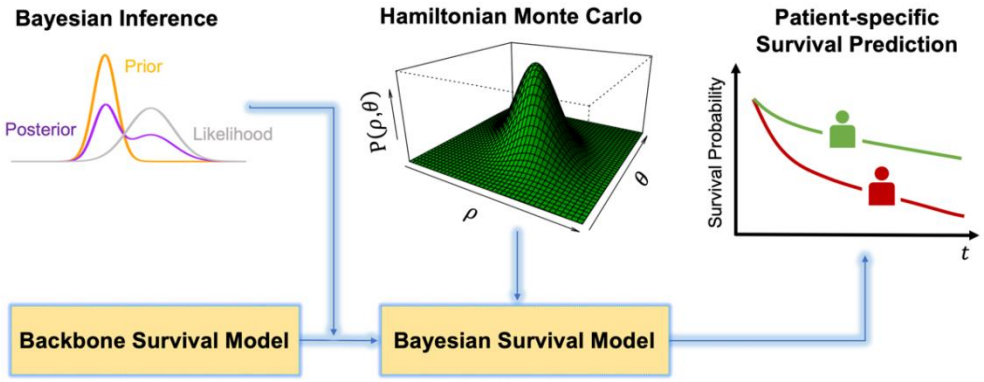

Figure 1. Flowchart of research methodology

2.3.1 Backbone Survival Model for Bayesian Inference

In survival analysis, a survival model provides a complete picture of longitudinal survival, which is generally formulated as

$$S(t) = P(T > t|\boldsymbol{z}) \tag{2.1}$$

where T is a random variable that represents survival time, t denotes a certain time point, $\boldsymbol{z} = (z_1, \dots, z_\tau)$ is a set of τ variables (e.g., race, gender, primary site, and chemotherapy), and $S(t)$ outputs the survival probability at a certain time point t.

A survival model can be both parametric (e.g., AFT model) and semiparametric (e.g.,

cox PH regression). Specifically, in cox PH regression, a hazard is defined as the instantaneous rate of occurrence of death at time t, and it is formulated as

$$h(t|\boldsymbol{z}) = h_0(t)e^{\boldsymbol{\beta z}} \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_\tau)$ denotes a set of model parameters and $h_0(t)$ is a baseline hazard function. Further, a cumulative hazard function is defined based on hazard

$$H(t) = \int_{u=0}^{t} h(t|\boldsymbol{z})du \tag{2.3}$$

In the survival analysis, the following equation holds

$$S(t) = e^{-H(t|\boldsymbol{z})} \tag{2.4}$$

In this paper, we use M-spline distribution as the baseline hazard function [29], because it does not presume a distribution of times that events occur [30]. Accordingly, the hazard is given by

$$h(t|\boldsymbol{z}) = \sum_{n=1}^{N} \epsilon_n M_n(t, \boldsymbol{k}, o)e^{\boldsymbol{\beta z}} \tag{2.5}$$

where n denotes the index of M-spline basis function, $n = 1, \dots, N$, M_n represents the nth basis function, $\boldsymbol{k}$ is a vector of knots, o denotes the order of polynomials in M-spline function, and ϵ_n is the parameter of nth basis function, $\sum_{n=1}^{N} \epsilon_n = 1$.

2.3.2 Accounting for Model Uncertainty with a Bayesian Framework

Model uncertainty is the uncertainty regarding the true underlying function. State-of-the-art survival predictions often count on a single best model or average across an ensemble of models. However, models in an ensemble with nearly equivalent performance accuracy can lead to disagreement on output distribution for a given patient, which further results in disagreement regarding decisions made based on model outputs [31]. Decision-making based on uncertain model outputs can cause adverse consequences. Therefore, it is critical

to account for model uncertainty in survival predictions.

Provided the backbone survival model, model uncertainty can be represented with a distribution of learnable model parameters, e.g., regression coefficients and auxiliary parameters [32], which are collectively denoted as $\boldsymbol{\theta}$. This is because the distribution of model parameters induces the distribution of models [33,34]. The proposed Bayesian framework captures the model uncertainty through three steps: (1) we place a prior distribution $p(\boldsymbol{\theta})$ over the model parameter set $\boldsymbol{\theta}$; (2) we estimate the posterior distribution of model parameters $p(\boldsymbol{\theta}|\boldsymbol{D})$ with a full Bayesian inference approach (detailed in Section 2.3.3); (3) and we predict the patient-specific survival probability accounting for model uncertainty as

$$p(S_i(t)|\boldsymbol{z}_i, \boldsymbol{D}) = \int p(S_i(t)|\boldsymbol{z}_i, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{D}) d\boldsymbol{\theta} \tag{2.6}$$

where $\boldsymbol{D}$ denotes the data, and $S_i(t)$ is the survival probability of ith patient.

2.3.3 Estimating Posterior Distribution using Full Bayesian Inference

HMC is a full Bayesian inference approach [35,36], which obtains the posterior distribution through a Hamiltonian energy function

$$H(\boldsymbol{\theta}, \boldsymbol{\rho}) = V(\boldsymbol{\theta}) + U(\boldsymbol{\rho}|\boldsymbol{\theta}) \tag{2.7}$$

where $V(\boldsymbol{\theta})$ represents a potential energy function of model parameters, $U(\boldsymbol{\rho}|\boldsymbol{\theta})$ is a kinetic function of model parameters $\boldsymbol{\theta}$ and auxiliary momentum variables $\boldsymbol{\rho}$, $\boldsymbol{\rho} \sim Multinormal(0, \boldsymbol{M})$, and $\boldsymbol{M}$ is the covariance matrix. Accordingly, equation (7) can be written in the form

$$H(\boldsymbol{\theta}, \boldsymbol{\rho}) = V(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\rho}^T \boldsymbol{M}^{-1} \boldsymbol{\rho} \tag{2.8}$$

To estimate the posterior distribution of $\boldsymbol{\theta}$, at each time step t, we draw samples of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$

from the joint system $H(\boldsymbol{\theta}, \boldsymbol{\rho})$. According to Hamiltonian equations, the following relationship between $\boldsymbol{\rho}$, $\boldsymbol{\theta}$, and t holds

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial H}{\partial \boldsymbol{\rho}} = \boldsymbol{M}^{-1}\boldsymbol{\rho} \tag{2.9}$$

$$\frac{d\boldsymbol{\rho}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = -\frac{\partial V}{\partial \boldsymbol{\theta}} \tag{2.10}$$

Therefore, at each time step, we can approximate values of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ as

$$\boldsymbol{\rho}_{t+\frac{\gamma}{2}} \leftarrow \boldsymbol{\rho}_t - \frac{\gamma}{2}\frac{\partial V}{\partial \boldsymbol{\theta}} \tag{2.11}$$

$$\boldsymbol{\theta}_{t+\gamma} \leftarrow \boldsymbol{\theta}_t + \gamma \boldsymbol{M}^{-1}\boldsymbol{\rho}_{t+\frac{\gamma}{2}} \tag{2.12}$$

$$\boldsymbol{\rho}_{t+\gamma} \leftarrow \boldsymbol{\rho}_{t+\frac{\gamma}{2}} - \frac{\gamma}{2}\frac{\partial V}{\partial \boldsymbol{\theta}} \tag{2.13}$$

where γ is a small increment in time towards the next time step [37]. A No-U-Turn Sampler is introduced here to optimize the incremental value of γ and the number of increments at each time step [38]. At the end of each time step, we use Metropolis acceptance to determine whether values of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ are acceptable. If so, we proceed to the next time step with updated values of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$. If not, we restart the approximation at the current time step [39]. Once values of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ stabilize, we obtain the posterior distribution of $\boldsymbol{\theta}$ by marginalizing out auxiliary momentum variables $\boldsymbol{\rho}$.

2.4 Experimental Results

In the experiment, we predict the survival probability of Leukemia-related death using the childhood Leukemia dataset in SEER database. To evaluate the predictive performance, C-index is introduced and computes the ratio of the number of concordant pairs to the number of comparable pairs. Also, standardized survival probabilities of the censored group and the deceased group are compared to further validate the proposed model. With the proposed

model, we also investigate the effects of clinical attributes on the mortality of childhood Leukemia. These significant experimental results indicate the importance in prognosis and decision-making on treatment of childhood Leukemia.

2.4.1 Data Description

Table 1. Explanation of variables in childhood Leukemia dataset.

Variable	Description
Age	Identifies the age group of the patient at diagnosis
Gender	Identifies the gender of the patient at diagnosis
Race	Identifies the race information of the patient
Site	Identifies the subtype of Leukemia
Grade	Indicates cell types and morphology
Radiation therapy	Identifies the different types of radiation therapy, including those who did not receive radiation therapy
Chemotherapy	Identifies whether the patient received chemotherapy
COD	Identifies the cause of death
MFDT	Represents months period from diagnosis to treatment
Tumor	Represents the number of tumors that occurred for a patient
Time	Represents the survival time
Status	Identifies whether the patient is deceased or censored

The database used in this research is Incidence-SEER Research Plus Data, 12 Registries, Nov 2021 Sub (2000-2019) [40], supported by the National Cancer Institute. In this research, we are particularly interested in the childhood Leukemia dataset (ages 0-19 years). In the dataset, children below nine years old comprise 61% of all childhood Leukemia patients. In addition, lymphocytic Leukemia is the most common type of Leukemia among children who are less than 19 years old, accounting for 76% of all Leukemia patients. The childhood Leukemia dataset consists of 17,539 children with Leukemia. Demographic and clinical variables in the dataset are used for Bayesian survival modeling, including age,

gender, race, primary site, grade, radiation therapy, chemotherapy, months from diagnosis to treatment (MFDT), number of tumors, cause of death (COD), survival time, and status. Explanations of these variables are summarized in Table 1.

2.4.2 Predictive Performance Evaluation of Bayesian Survival Model

In the experiment, we use C-index as the performance metric to evaluate the predictive performance of the proposed model [41,42]. C-index is defined as the number of concordant pairs divided by the number of comparable pairs which is formulated as

$$C = \frac{\sum_{i,j} I(T_i > T_j) * I(\eta_j > \eta_i) * d_j}{\sum_{i,j} I(T_i > T_j) * d_j} \tag{2.14}$$

where (i, j) refers to indices of a pair of patients, η_i and η_j are risk scores for patients i and j, d_j is a binary variable indicating whether a patient is deceased or not, $d_j = 0$ if the patient is censored, otherwise $d_j = 1$, and T_i and T_j are survival times of patient i and j given in the real data. However, there are some cases that are not comparable. These include cases (1) when patients i and j are both censored; (2) and when the observed survival time of censored patient j is shorter than the deceased patient i (as we cannot determine who actually survives longer in this case). Here, we introduce d_j to eliminate these incomparable pairs as when patient j is censored, $d_j = 0$, and thus equation (14) does not take these cases into account. Values of C-index range from zero to one. For model evaluation, a higher value of C-index indicates stronger model performance and higher predictive accuracy. A value of 0.5 means that the model is no better than random chance while values over 0.8 indicate the model is of high precision [43]. C-index of the proposed Bayesian survival model is 0.93, which indicates the effectiveness and accuracy of the proposed model. Further, the model is significant for healthcare practitioners to better inform prognosis and treatment of childhood

Leukemia.

2.4.3 Comparison of Standardized Survival Probabilities – Censored Group vs Deceased Group

To further validate the proposed model, we compare the standardized survival probabilities of different groups which is formulated as

$$p(S^*(t)|\boldsymbol{D}) = \frac{1}{L}\sum_{l=1}^{L} p(S_i(t)|\boldsymbol{z_i}, \boldsymbol{D}) \quad (2.15)$$

where L is the number of patients in a particular group (censored or deceased) and $S^*(t)$ is the standardized survival probability of the corresponding group. The standardized survival probability assists us to investigate the overall survival probabilities of different groups of patients.

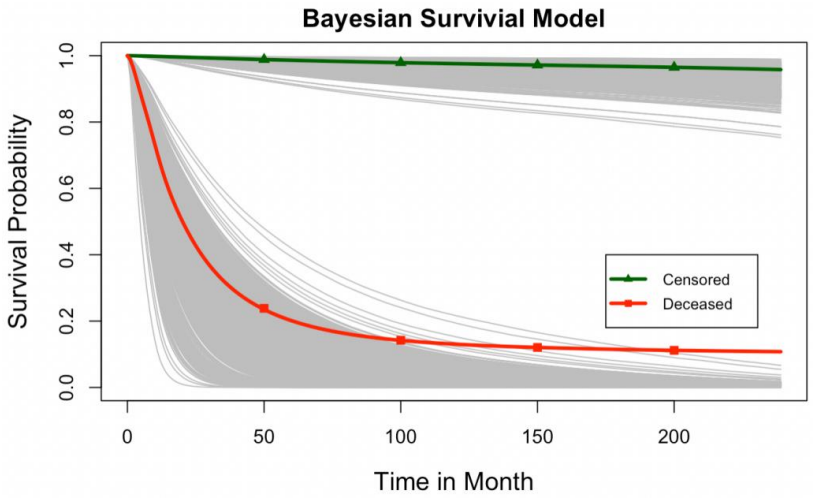

Figure 2. Patient-specific survival probability predicted by Bayesian survival model. Red line shows the standardized survival probability of the censored group, green line shows the standardized survival probability of deceased group, and gray line gives the individual survival probability of all patients changing with respect to time.

As shown in Figure 2, the green line represents the standardized survival probability of censored group, while the red line is the standardized survival probability of the deceased group. Each gray line gives the patient-specific survival prediction changing with respect to

time. However, the standardized survival probability of the censored group is consistently higher than the probability of the deceased group, suggesting that the proposed model is robust to predict survival probability accurately.

2.4.4 Investigating Effects of Clinical Attributes on Mortality of Childhood Leukemia

In the experiment, we also investigate the effects of certain clinical attributes, e.g., primary site, chemotherapy, and radiation therapy, on standardized survival probabilities of children with Leukemia. Each of these variables is selected due to its high hazard ratio referred to as

$$Hazard\ Ratio = e^{\beta_\tau} \tag{2.16}$$

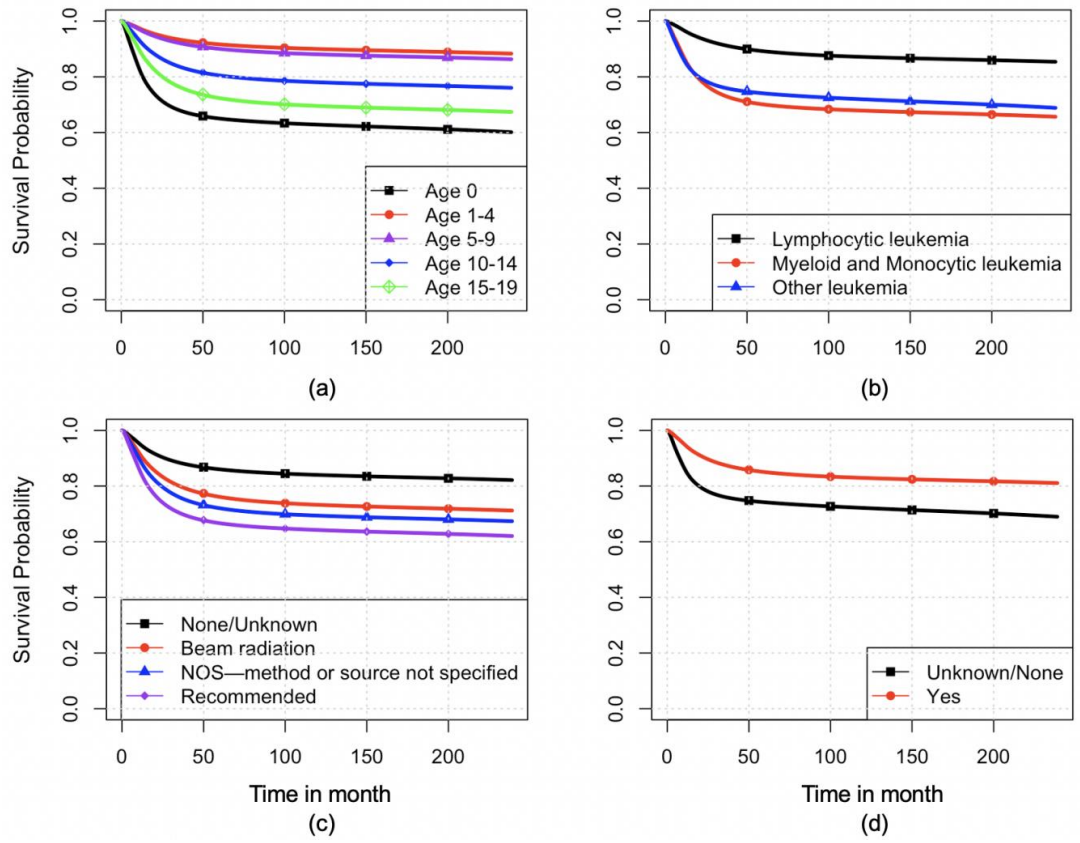

Figure 3. Effects of clinical parameters on survival probability of children with leukemia: (a) Effects of different age groups; (b) Effects of different types of leukemia; (c) Effects of radiation therapy; (d) Effects of chemotherapy.

The hazard ratio of variable z_τ indicates the effect of its unit increase on the survival probability, assuming that all the other variables hold constant. Among patients with Leukemia, those who are less than one year old and are greater than ten years old are regarded as high-risk patients. High-risk group patients are more likely to die and have lower standardized survival probabilities. As shown in Figure 3 (a), standardized survival probabilities of patients from different age groups are presented. Specifically, newborn children (children who are less than one year old) with Leukemia have the lowest survival probabilities among all age groups. Moreover, children over ten years old have lower survival probabilities than those who are less than ten years old and greater than one year old. This result verifies the empirical Leukemia-related facts. Figure 3 (b) shows that children with lymphocytic Leukemia have the highest standardized survival probability. On the other hand, children with myeloid, monocytic, and other types of Leukemia have lower standardized survival probabilities than the lymphocytic type. Furthermore, different types of therapy methods (e.g., chemotherapy and radiation therapy) also influence the occurrence of decease. As shown in Figure 3 (c), the group without radiation therapy has the highest standardized survival probability. This is probably because treatments in pediatric Leukemia are tailored to the risk stratification at diagnosis. Patients in the low-risk group are less severe than the other groups, which has led to a less intensity of radiation therapy. Furthermore, another interesting finding is that those patients who received beam radiation therapy possess the second highest standardized survival probability. Figure 3 (d) demonstrates the effect of chemotherapy on mortality of childhood Leukemia. Children who received chemotherapy possess a higher standardized survival probability than those without chemotherapy. Accordingly, chemotherapy does increase the standardized survival probability of children

with Leukemia. In conclusion, the proposed model can help identify the effects of various attributes on survivals of childhood Leukemia patients. This is conducive to reducing Leukemia-related death and providing appropriate interventions to life-threatening events in a timely manner, thereby positively influencing cancer care.

2.5 Conclusions

Childhood Leukemia is the most common cancer among children. Its serious morbidity and mortality rate motivate researchers to investigate the occurrence of Leukemia-related death, assisting healthcare practitioners to make a better decision upon timely prognosis and treatment. In the medical domain, survival analysis is widely applied in a broad of illnesses. However, most previous work on survival analysis relies more on a single model with high predictive precision, which fails to consider model uncertainty. To fill this gap, we develop a Bayesian survival model to predict the occurrence of death for children with Leukemia by taking model uncertainty into account. Specifically, the proposed model integrates a backbone survival model with Bayesian inference. Then, we estimate the posterior distribution of model parameters using a full Bayesian inference approach. By considering model uncertainty, the proposed Bayesian survival model can provide patient-specific predictions efficiently and effectively.

Experimental results show that C-index of the Bayesian survival model is 0.93, suggesting the proposed model is of high predictive precision. Moreover, the discrepancy in standardized survival probabilities between the censored group and the deceased group further reveals the predictive accuracy of the proposed model. The Bayesian survival model can also identify the effects of clinical attributes (e.g., age, chemotherapy, radiation therapy, and types of Leukemia) on Leukemia-related death, which is conducive to reducing the

Leukemia-related death and providing appropriate interventions to life-threatening events in a timely manner. This research is of great importance because: (1) Clinicians can better inform prognosis and treatments of childhood Leukemia based on accurate prediction. (2) The effects of clinical attributes can be tracked with the proposed model, thereby helping with decision-making on interventions. Further research could study multistate outcomes of childhood Leukemia rather than binary outcomes, i.e., censored and deceased. As more patients are added to the SEER database, deep learning models (e.g., Long short-term memory network) can be used for the model development due to their great power to deal with nonlinear dynamics that are inherent in the electronic health record, thereby better assisting medical decision-making.

CHAPTER 3

Explainable Transformer-based Deep Survival Model for Survival Prediction of Childhood Acute Lymphoblastic Leukemia

3.1 Introduction

Acute lymphoblastic leukemia (ALL) is a type of cancer that affects the white blood cells, particularly the lymphocytes. The disease is characterized by the overproduction of abnormal and immature lymphocytes in the bone marrow [44]. These abnormal lymphocytes are unable to function properly or mature into healthy blood cells, resulting in a decrease in normal white blood cells, red blood cells, and platelets [45]. The decrease in these blood cells can lead to disruptions in the immune system, which significantly raises the risks of infection and life-threatening events, often fatal without timely intervention. Most recent statistics in 2024 reveal that leukemia is the most common cancer among children (ages 0 to 14 years) and adolescents (ages 15 to 19 years), accounting for 41% of all pediatric and adolescent cancers in the United States. Moreover, ALL is the most prevalent type of leukemia in this demographic, representing 76% of children and adolescents diagnosed with leukemia in the United States [3]. Despite advancements in diagnosis and treatment, ALL remains a significant health concern, as evidenced by a growing trend in the rate of new cases, which has escalated from 1.5% in 1992 to 1.7% per 100,000 individuals in 2020 [46]. One critical challenge lies in the lack of an accurate and efficient predictive approach to forecasting ALL survival for pediatric patients, leading to delays in monitoring and intervention. Addressing

this challenge is of paramount importance to improve survival rates among ALL patients in childhood.

In healthcare, survival analysis aims to predict the time until a particular health-related event occurs. It is a statistical approach that can generate survival probabilities over a given time frame for patients diagnosed with a specific disease [47]. In recent decades, various survival models have been widely adopted to make survival predictions [48]. For example, a semi-parametric survival model, the Cox proportional hazard (PH) regression model, has been widely used to predict survival probabilities [49]. This model assumes that the hazard function, which represents the instantaneous event occurrence, is proportional across different levels of covariates. This also indicates that the hazard of each covariate remains constant over time. However, this assumption is often violated in real-world situations [50]. For example, a treatment may have a different impact on survival probabilities in the short term compared to the long term, indicating time-dependent effects of covariates on survival. Other survival models, such as Support Vector Machines (SVM), Random Survival Forests, and deep neural networks (DNNs), demonstrate accurate predictive performance but often face time-consuming computational processes, primarily due to limitations in parallel computing capabilities [51,52]. Moreover, these models are limited in capturing the inherent long-term dependencies within the time-to-event data and prioritizing specific informative input segments in the predictive process, resulting in less accurate and interpretable predictions.

To address this challenge, the Transformer has emerged as a solution. Vaswani et al. proposed the Transformer architecture in 2017, initially used as a foundation for various natural language processing models [53]. Transformers rely on self-attention and multi-head

attention mechanisms to process sequential data. The self-attention mechanism calculates attention weights, which helps the transformer effectively process sequential data and make contextually informed decisions. This mechanism prioritizes important tokens by assigning higher attention weights while assigning lower weights to less important ones [54]. On the other hand, the multi-head attention mechanism enables the model to focus on various parts of input data simultaneously. This mechanism splits the input embedding into multiple parallel heads, with each head independently calculating attention weights. The use of multiple attention heads allows the transformer to capture diverse representations within the input data, which in turn improves the model performance [55]. In recent years, Transformers have been applied in multiple domains due to their flexibility and robust performance, including speech recognition [56], language modeling [57], and text classification [58]. Compared with other deep learning models, such as recurrent neural networks (RNNs) [59], which process sequences step by step, Transformers leverage a parallel computational process, which significantly enhances computational efficiency and allows Transformers to digest long-term dependencies more effectively. Therefore, in this paper, we develop an explainable Transformer-based deep survival model to predict the time to death for children and adolescents diagnosed with ALL. Specifically, we first train the Transformer-based deep survival model by minimizing a loss function that measures the discrepancy between predicted and actual survival status. Second, we predict the survival probability at a particular point in time with the well-trained model. Third, we explain the survival model and evaluate the contributions of clinical attributes to predictions with Shapley Additive explanations (SHAP) from both global and local perspectives [60].

The outline of this paper is as follows. Section 3.2 discusses existing survival models and deep learning applications in the field of survival analysis. Section 3.3 presents the research methodology of the proposed explainable Transformer-based deep survival model. Section 3.4 evaluates the proposed methodology. Section 3.5 concludes the research.

3.2 Research Background

Conventional survival modeling relies on assumptions of event time distributions and the estimation of model parameters. With the advancements in machine learning and deep learning, survival analysis has become more flexible and capable of capturing nonlinear dynamics, thus addressing complex real-world problems effectively. Further, by enhancing the transparency and explainability of the survival models, SHAP improves the quality of decisions made in clinical practice.

3.2.1 Statistical Survival Modeling

Survival models based on traditional statistical methods have been commonly applied in the medical domain. Conventionally, statistical survival models are typically categorized into non-parametric, parametric, and semi-parametric models. These models were developed primarily to deal with the time-to-event data for the prediction of survival probability, which refers to the probability of a patient surviving beyond a particular time threshold. By understanding a patient's survival probability over time, clinicians can provide timely interventions, thereby reducing the occurrences of disease-induced deaths.

Non-parametric models provide an overview of the disease without explicitly identifying relationships between response and predictor variables. Non-parametric methods include the Nelson-Aalen estimator [61], the Kaplan-Meier estimator [13], and the life-table method [62]. These methods directly estimate the overall hazard or survival function for a

group of patients. On the other hand, parametric models estimate model parameters by assuming prior information about event time distributions, such as exponential, gamma, Weibull, lognormal, and log-logistic distributions [63]. Parametric survival models, including the accelerated failure time (AFT) model [14], penalized regression [64], and the Buckley-James model [65], can provide patient-specific survival predictions and mathematically track relationships among variables. Nowadays, most researchers prefer semi-parametric methods because of their effectiveness and interpretability, such as the Cox PH regression model [66]. The Cox PH does not make prior assumptions about event time distributions. Instead, it estimates a specific hazard function using a pre-assumed baseline survival function.

Despite the longstanding presence of statistical survival models, they often face challenges related to counting statistics and pre-assumptions, which are often violated in real-world clinical settings. Moreover, existing statistical survival models overlook nonlinear relationships among variables. As a result, these survival models fail to generalize effectively in complex scenarios, leading to unrealistic predictive outcomes.

3.2.2 Machine Learning and Deep Learning in Survival Analysis

Machine learning-based survival models have been broadly applied in the medical domain due to their ability to digest complex variable relationships within medical records [67]. Prior to the rapid advancement of deep learning in recent years, ensemble approaches were frequently utilized for nonlinear machine learning [68-70]. However, their ability to handle nonlinear data usually falls short in addressing complex variable structures compared to deep learning-based survival models.

In contrast, deep learning models can handle complex data structures by quickly expanding with additional layers [71]. Specifically, DNNs exhibit a strong capability to scale with high-dimensional and large-sized datasets thanks to their unique network structure and training approaches. As a result, DNNs have become a popular choice in survival modeling within the medical domain. For example, The DeepSurv model [72], which employs a feedforward neural network, has demonstrated promising outcomes in enhancing early intervention and personalized treatment planning through improved survival prediction. However, the problem of generalization remains under-addressed. Subsequently, the DeepHit survival model was proposed to model time-to-event data without making distribution assumptions [73]. DeepHit combines recurrent neural networks and multi-layer perceptron (MLP) to capture the temporal dynamics of clinical data without imposing assumptions. However, DeepHit's learning capability can be limited as it relies solely on a simple MLP. To this end, in recent years, researchers have been exploring variations of deep learning models to address these challenges.

3.2.3 Enhancing Model Interpretability in Real-world Clinical Applications

Explaining the predictive outcomes of black-box models like machine learning and deep learning models is becoming increasingly important. SHAP, a model-agnostic technique grounded in game theory, can enhance model interpretability by revealing how each feature affects a model's predictive results. Specifically, SHAP calculates Shapley values for each instance and each feature, illustrating the contributions of different variables to the predictions. Positive Shapley values indicate that the inclusion of the variable increases the prediction beyond the average prediction of all instances, while negative values suggest that the variable decreases the prediction below the average prediction. Unlike other

interpretability-enhancing measures (e.g., Feature importance), Shapley values identify the most influential features for a prediction and the direction of their influence (e.g., raising or lowering the forecast).

SHAP has been extensively used in various domains to explain black-box models. For example, researchers have combined deep neural networks with SHAP to forecast sales [74]. In addition, researchers use RSF to analyze failure mode and effects in the manufacturing domain and further investigate feature importance with SHAP [75]. Moreover, machine learning and deep learning methods, such as SVM, Long-short term memory (LSTM), and XGBoost, are combined with SHAP to offer more insightful perspectives in survival modeling explanation [76]. These works have demonstrated the prevalence of using SHAP for enhancing model interpretability, providing practitioners with additional information for informed decision-making.

3.3 Research Methodology

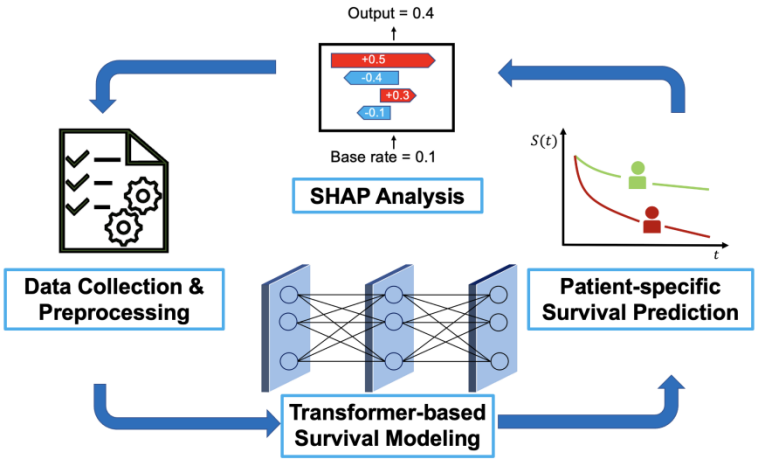

Figure 4. Flowchart of Research Methodology

This section presents the proposed explainable Transformer-based deep survival model to predict survival for children and adolescents diagnosed with ALL. As shown in

Figure 4, we first preprocess the raw ALL dataset, which is obtained from the University of Oklahoma (OU) Health Cancer Registry, OU Health electronic health records (EHR), and the Oklahoma Central Cancer Registry. Second, we develop a Transformer-based deep survival model to predict patient-specific survival probabilities. Third, we explain the black-box model and quantify the contribution of each feature to predictive outcomes with SHAP.

3.3.1 Data Preprocessing and Problem Formulation in Survival Prediction

The dataset used in this paper is sourced from the OU Health Cancer Registry, OU Health electronic health records (EHR), and the Oklahoma Central Cancer Registry. This dataset involves 275 children and adolescents diagnosed with ALL from 2005 to 2019 [77]. The clinical variables selected for survival modeling include gender, histology (following the International Classification of Childhood Cancer 3rd Edition [78]), age at diagnosis (ranging from 0 to 19 years of age), recurrence status (defined as relapse or recurrence after complete remission following initial induction therapy), race/ethnicity (American Indian, Hispanic, Non-Hispanic (NH) Black, NH White, and NH Asian/Pacific Islander), prognosis status (based on post-induction prognosis), diagnosis year, insurance/payer type (Indian Health Service, Insured, Medicaid, Uninsured, Unknown), distance to care (in miles) from address at diagnosis (DCAD) to the Oklahoma Children's Hospital OU Health, area deprivation index using national rankings based on address at diagnosis (ADAD) [79], distance to care from the current address (DCCA) in the EHR to the Oklahoma Children's Hospital OU Health, area deprivation index based on current address (ADCA), overall survival time (OST) calculated from date of diagnosis to death or end of follow-up (October 20, 2020) (whichever occurred first), and event status (deceased or censored). Detailed descriptions of these variables are provided in Table 2.

Table 2. Descriptions of Variables

Variable	Description
Age at Diagnosis	Identifies the age of the patient at diagnosis
Gender	Identifies the gender of the patient at diagnosis
Race/Ethnicity	Identifies the race and ethnicity information of the patient
Recurrence Status	Indicates relapse or recurrence that occurred after complete remission following initial induction therapy
Histology	Indicates ALL subtypes
Prognosis Status	Indicates the prognosis status based on the post-induction assessment
Diagnosis Year	Identifies the year of diagnosis
Insurance Type/Payer	Indicates the insurance type or primary payer of a patient
DCAD	Represents the distance to care (in miles) from address at diagnosis to the Oklahoma Children's Hospital OU Health
ADAD	Represents the area deprivation index using national rankings based on address at diagnosis
DCCA	Represents the distance to care from the current address in the EHR to the Oklahoma Children's Hospital OU Health
ADCA	Represents the area deprivation index based on the current address
OST	Represents the overall survival time calculated from the date of diagnosis to death or end of follow-up
Event Status	Identifies whether the patient is deceased or censored

We denote the entire dataset as $\boldsymbol{z} = \{\boldsymbol{x}, y, T\}_{p=1}^{N}$, where p represents the patient index, N is the total number of patients recorded in the dataset, $\boldsymbol{x} = (x_1, \dots, x_l)$ denotes a set of l predictor variables, $y \in (0,1)$ is a binary response variable indicating a patient's survival status (i.e., zero indicates censored status, while one represents deceased status), and T is a random variable that represents the OST. Further, we perform one-hot encoding for categorical variables and regularize continuous variables to enhance the learning ability of the model. We divide the dataset into three segments: 60% of the data serves as a training

set for model development, 20% as a testing set, and the remaining 20% as a validation set for performance evaluation. The training process stops when there are no further improvements in the performance of the validation set.

In this paper, the survival function, defined as the probability that a patient survives beyond a particular time point, is formulated as

$$S(t_k|\boldsymbol{x},\boldsymbol{\theta}) = P(T > t_k|\boldsymbol{x},\boldsymbol{\theta}) \tag{3.1}$$

where $t_k \in \{t_0, \dots, t_K\}$ denotes a particular time point, t_K is the maximum survival time found in OST, $\boldsymbol{\theta}$ represents the model parameters of the Transformer-based deep survival model, $\boldsymbol{x}$ is a set of variables, and $S(t_k|\boldsymbol{x},\boldsymbol{\theta})$ represents the survival probability at t_k.

In addition, a hazard function refers to the conditional probability that the patient experiences an event at t_k given they remain event-free before t_k, is formulated as

$$h(t_k|\boldsymbol{x},\boldsymbol{\theta}) = P(T = t_k|T > t_{k-1},\boldsymbol{x},\boldsymbol{\theta}) \tag{3.2}$$

By applying Bayes' Theorem, we can express the survival function as

$$S(t_k|\boldsymbol{x},\boldsymbol{\theta}) = P(T > t_k|T > t_{k-1},\boldsymbol{x},\boldsymbol{\theta})\, P(T > t_{k-1}|\boldsymbol{x},\boldsymbol{\theta}) \tag{3.3}$$

where the first term on the right-hand side is equivalent to $1 - P(T = t_k|T > t_{k-1},\boldsymbol{x},\boldsymbol{\theta})$. Therefore, combining Equations (3.2) and (3.3), the survival function can be defined as

$$S(t_k|\boldsymbol{x},\boldsymbol{\theta}) = \big(1 - h(t_k|\boldsymbol{x},\boldsymbol{\theta})\big)S(t_{k-1}|\boldsymbol{x},\boldsymbol{\theta}) \tag{3.4}$$

By recursively expanding Equation (3.4), we can derive the formula for the survival function as

$$S(t_k|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{t_0}^{t_k}\big(1 - h(t_k|\boldsymbol{x},\boldsymbol{\theta})\big) \tag{3.5}$$

$$h(t_0|\boldsymbol{x},\boldsymbol{\theta}) = P(T = t_0) \tag{3.6}$$

According to Equation (3.5), we can calculate the survival function by cumulatively

multiplying the hazard function. Therefore, in our proposed model (detailed in 3.3.2), we first output the hazard function at each time point and then use the cumulative multiplication described in Equation (3.5) to derive the desired survival function. This process ensures the survival probability decreases monotonically over time, reflecting real-world scenarios where the absence of effective treatments leads to a non-increasing survival function.

3.3.2 Transformer-based Deep Survival Model for Childhood ALL

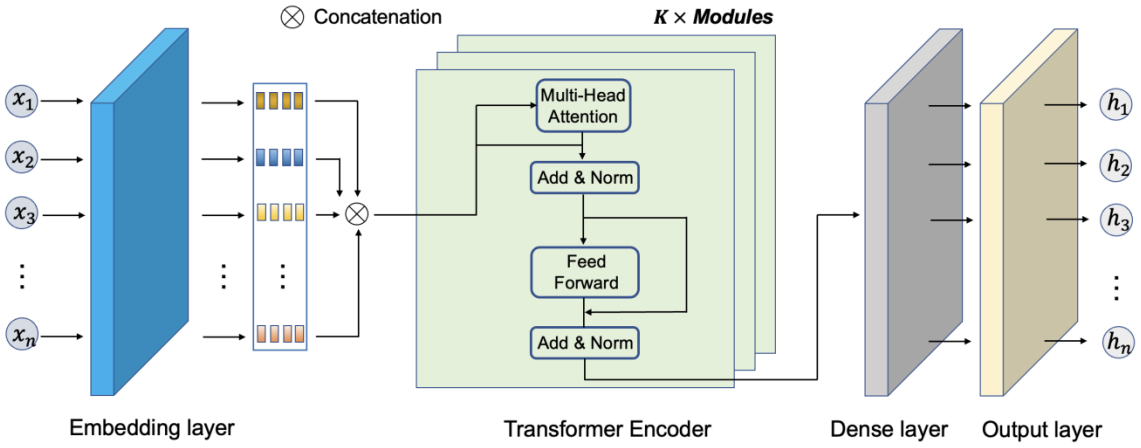

Figure 5. The architecture of the Transformer-based Deep Survival Model

As shown in Figure 5, the architecture of the proposed Transformer-based deep survival model consists of four parts: (1) An embedding layer for feature extraction, mapping the sparse feature space into continuous and fixed vectors; (2) A Transformer encoder with several encoder blocks to perform layer normalization, multi-head attention mechanism, dropout application, and residual connection; (3) A dense layer to scale the output from the Transformer encoder to a fixed dimension, consistent with the maximum event time in the data set; (4) An output layer with sigmoid function to estimate the hazard functions at different time points. The input features are first fed into an embedding layer individually. Then, the resulting outputs from the embedding layer are concatenated and used as input for

the Transformer encoder. The output obtained from the Transformer encoder is then channeled into a dense layer and reshaped to match the maximum event time dimension. Finally, an output layer employing a sigmoid function generates the target hazard functions.

Table 3. Patient Examples with Survival Status

Patient Index	**(Time, Status)**	**Survival Status over Time**			
p	(T_p, y_p)	$e_{t_0}^p$	$e_{t_1}^p$	$e_{t_2}^p$	$e_{t_3}^p$
1	(1,0)	1	0	0	0
2	(2,0)	1	1	0	0
3	(170,0)	1	1	1	1
4	(174,1)	1	1	1	1

In this paper, we propose a loss function tailored for predicting survival outcomes related to death events based on the cross-entropy theorem. Table 3 outlines two types of events we considered in the loss function, i.e., censored and deceased. The deceased patient (e.g., Patients 1, 2, and 3) is assigned a survival status of 0 following the occurrence of a death event, whereas the censored patient (e.g., Patient 4) has a survival status of 1 until the end of follow-up. In the table, we denote p as the patient index, T_p as the event time for a deceased patient or the end of follow-up for a censored patient, $y_p \in (0,1)$ as the event status for patient p at time T_p, and $e_{t_k}^p$ as the survival status for Patient p at a particular time point t_k. Accordingly, we formulated the loss function as

$$L_c = -\sum_{t_k=t_0}^{t_K} \sum_{p \in R_{t_k}} \left\{ e_{t_k}^p \, ln\left(S_p(t_k|\boldsymbol{x}, \boldsymbol{\theta})\right) \; + \; \left(1 - e_{t_k}^p\right) ln\left(1 - S_p(t_k|\boldsymbol{x}, \boldsymbol{\theta})\right) \right\} \quad (3.7)$$

where R_{t_k} is a subset of patients. $R_{t_k} = \{p \mid \{y_p = 0\}, or \, \{y_p = 1 \; and \; T_p > t_k\}\}$ consists of those patients who have either experienced the death event before the maximum survival

time t_K in OST, represented by $y_p = 0$, or those who have not reached the end of follow-up T_p at time t_k, indicated by $y_p = 1$ and $T_p > t_k$. We train the model by minimizing the loss function L_c, ensuring that the predicted survival probabilities align closely with the actual survival status at each time point.

3.3.3 Improving Model Interpretability using SHAP Analysis

SHAP is a model-agnostic method that explains the outputs of machine learning and deep learning models [80]. SHAP values highlight the positive and negative impacts of individual features on predictions, the relative importance of features, and the interrelationships among these features in a specific model's decision-making process. Specifically, SHAP interprets the predictions through an explanation model formulated as

$$\eta(\boldsymbol{\beta}) = \phi_0 + \sum_{j=1}^{l} \phi_j \beta_j \tag{3.8}$$

where η is the explanation model, $\boldsymbol{\beta} \in \{0,1\}^l$ represents whether a feature can be observed, β_j gives the j^{th} value in $\boldsymbol{\beta}$, l is the number of input features, ϕ_j denotes the Shapley value for feature j, and ϕ_0 is a constant value when all inputs are missing. The cumulative sum of Shapley values for all features reflects the deviation between the predicted outcome generated by the explanation model and the average prediction across all samples [81].

In this paper, we formulate the Shapley value ϕ_j as

$$\phi_j = \sum_{M \subseteq \{(x_1,\ldots,x_l)\} \setminus \{x_j\}} \frac{|M|!(l-|M|-1)!}{l!} \left(f_x\left(M \cup \{x_j\}\right) - f_x(M) \right) \tag{3.9}$$

where M is a subset of features used in the original model f_x, $f_x(M)$ indicates the model prediction with M, and $f_x\left(M \cup \{x_j\}\right)$ presents the model prediction with M and an additional feature x_j. According to Equation (3.8) and Equation (3.9), the Shapley value indicates the impact of individual features on the final output of the Transformer-based deep survival

model. Note that Shapley values can be either positive or negative. Positive Shapley values push the model's prediction values higher, while negative Shapley values drive the model's prediction values lower.

3.4 Experimental Results

In the experiment, we first compare the average survival probabilities of censored and deceased patients to evaluate the model performance. Second, we use C-index as a performance metric to benchmark the predictive accuracy of our proposed model against two commonly used deep survival models across the 25%, 50%, and 75% quartiles of the event time. Third, we use the SHAP method to further explain predictive outcomes, considering both global and local perspectives.

3.4.1 Evaluation of Model Discriminability between Censored Group and Deceased Group

The average survival probability offers a comprehensive view of survival trends among children and adolescents diagnosed with ALL. A reliable and accurate predictive model is expected to consistently yield a higher average survival probability for censored patients compared to deceased patients. Hence, we compare the average survival probabilities to validate and evaluate the proposed model's discriminatory ability between the censor and deceased groups.

In the experiment, we use the Adam as an optimizer with a fixed learning rate of {0.001, 0.0001} and weight decay of {0.01, 0.001}. We determine the number of Transformer layers and heads to be {2, 4}, and the embedding and hidden sizes are selected from {32, 64}. Furthermore, we maintain a batch size of 8. Figure 6 illustrates the patient-specific survival predictions for children and adolescents with ALL. Each gray line represents the predicted survival probability for an individual patient. The green line presents

the average survival probability for censored patients in the testing set, while the red line represents the average survival probability for deceased patients. The average survival probability of censored patients is consistently higher than that of deceased patients, which indicates that the proposed model can effectively discriminate between the two groups. This demonstrates the model's potential in predicting personalized survival outcomes for children and adolescents with ALL, thereby facilitating early intervention.

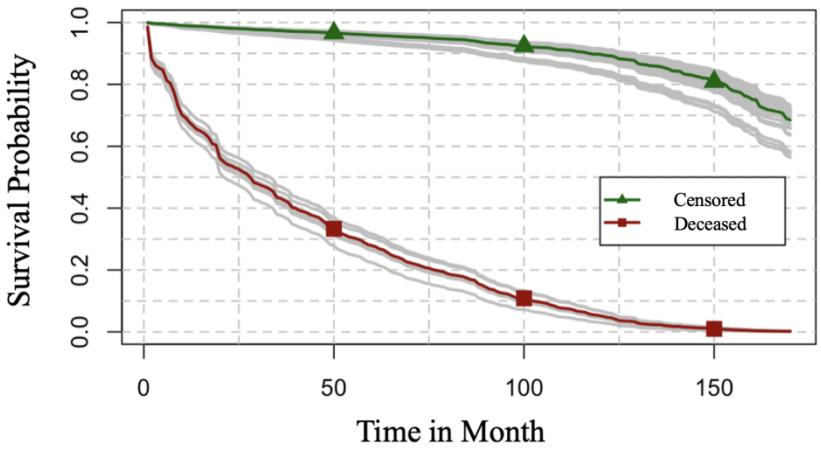

Figure 6. Patient-specific survival probabilities over time for children and adolescents diagnosed with ALL. Green line represents the average survival probability for the censored patients, while red line shows the average survival probability for the deceased patients.

3.4.2 Quantification of Model Performance with Concordance-index (C-index)

We use the C-index to evaluate the predictive accuracy of the proposed survival model, as it can effectively assess the model's ability to rank the survival times of patients [42]. It is formulated as

$$C = \frac{\sum_{u \neq v} I(T_u > T_v) * I(r_v > r_u) * g_v}{\sum_{u \neq v} I(T_u > T_v) * g_v} \tag{3.10}$$

where the indices u and v represent a pair of patients, T_u and T_v denote the survival times of patients u and v, I is an indicator function, $g_v \in (0,1)$ is a binary variable, and r_v and r_u

are their respective risk scores, refering to numerical values calculated to estimate the probability of an individual experiencing a particular event (e.g., death). A patient v is considered censored if $g_v = 0$, while patient v is deceased if $g_v = 1$. According to Equation (10), the C-index can be interpreted as the ratio of the number of concordant pairs to the number of all comparable pairs. The dataset consists of three types of cases. First, when both patients u or v are deceased, the pair (u, v) is considered concordant if $T_u > T_v$ and $r_v > r_u$, while it is discordant if $T_u > T_v$ and $r_u > r_v$. Second, in cases where one patient is censored and the other is deceased (e.g., patient u is censored and patient v is deceased), two scenarios are considered: (1) if $T_u < T_v$, we cannot determine which patient dies first if patient u dies after the end of follow-up. Therefore, these pairs are excluded from the C-index computation; (2) if $T_u > T_v$, the pair (u, v) is concordant if $r_v > r_u$ and discordant if $r_u > r_v$. Third, when both patients u and v are censored, and the occurrence of an event after the follow-up period and which event occurs first are unknown, these pairs are also discarded in the computation process. Values of the C-index range from zero to one, where higher values indicate greater model accuracy. A C-index near 0.5 implies the model performs no better than random guessing. A C-index greater than 0.8 indicates strong model performance.

In the experiment, we compare the C-index of our proposed with two commonly used deep survival models, namely DeepSurv and DeepHit. The comparison is conducted at the 25%, 50%, and 75% quartiles of the event time. As shown in Table 4, the proposed model achieves C-indices of 0.815, 0.828, and 0.831 at these quartiles. The experimental results demonstrate that the proposed model outperforms both DeepSurv and DeepHit across all quartiles. These results highlight the robustness and efficacy of our research methodology,

further validating the potential application of the proposed model in patient-specific survival predictions for ALL.

Table 4. C-index Values of Models at Different Quantiles of Event Time

Survival Model	Quartiles of Event Time		
	25%	50%	75%
DeepSurv	0.774	0.782	0.803
DeepHit	0.792	0.791	0.821
Transformer-based Survival Model	0.815	0.828	0.831

3.4.3 Model Interpretation with SHAP from Global to Local Perspectives

The Transformer-based deep survival model we propose exhibits both accuracy and robustness in predicting survival outcomes. However, the foundation of this model is a neural network, which is typically regarded as a black box. Hence, we address this opacity by offering a comprehensive global perspective and a detailed local interpretation of the model's outputs with SHAP analysis. Our goal is to better assist clinicians to make informed decisions in real-world scenarios.

Figure 7 presents a global perspective on model interpretation through SHAP values corresponding to each feature. Blue and red markers denote individual data points, with red indicating higher feature values and blue indicating lower feature values. The clinical features are ranked in descending order of impact on the model output, displayed on the left side. Among them, prognosis status, recurrence status, and histology rank as the top three impactful features. Notably, recurrence status and prognosis status determined by the post-induction assessments negatively impact model output, specifically survival probability.

These findings can be crucial for healthcare professionals in decision-making processes, especially when using the proposed model as a supporting tool. They enable clinicians to offer informed prognoses and implement early interventions and treatments for ALL.

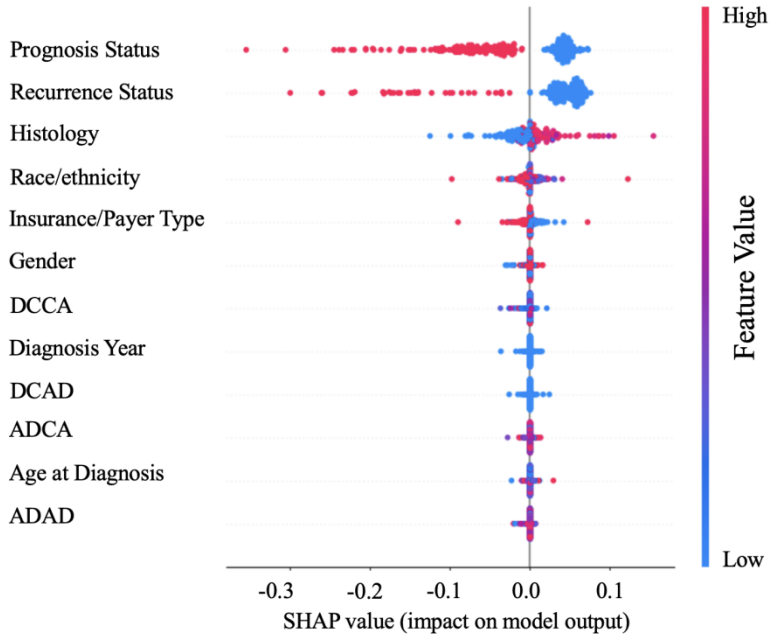

Figure 7. Global perspective of SHAP interpretation across various features.

We also include two waterfall plots in Figure 8, illustrating the impact of each feature on survival predictions for two different patients. As shown in each subfigure, $f(x)$ represents the predicted survival probability at the median event time, whereas $E[f(X)]$ corresponds to the average predicted survival probability across all patients within the testing set. Figure 8 (a) illustrates how SHAP values impact the survival prediction for the first patient. Specifically, a recurrence status of one reduces the predicted survival probability by 0.1, a prognosis status of one decreases the predicted survival probability by 0.09, and an ADCA value of 0.76 lowers the predicted survival probability by 0.01. Conversely, a histology of six pushes the predicted survival probability higher by 0.1. For the second

patient, as shown in Figure 8 (b), a prognosis status of one and a histology of one negatively affect the survival predictions, decreasing the survival probability by 0.07 and 0.02, respectively. On the other hand, a recurrence status of zero and an insurance or payer type of 1 positively impact survival predictions, driving the predicted survival probability higher by 0.07 and 0.01, respectively. According to local perspectives from SHAP, detailed explanations of each feature's impact can be specified, facilitating personalized medical care for children and adolescents with ALL.

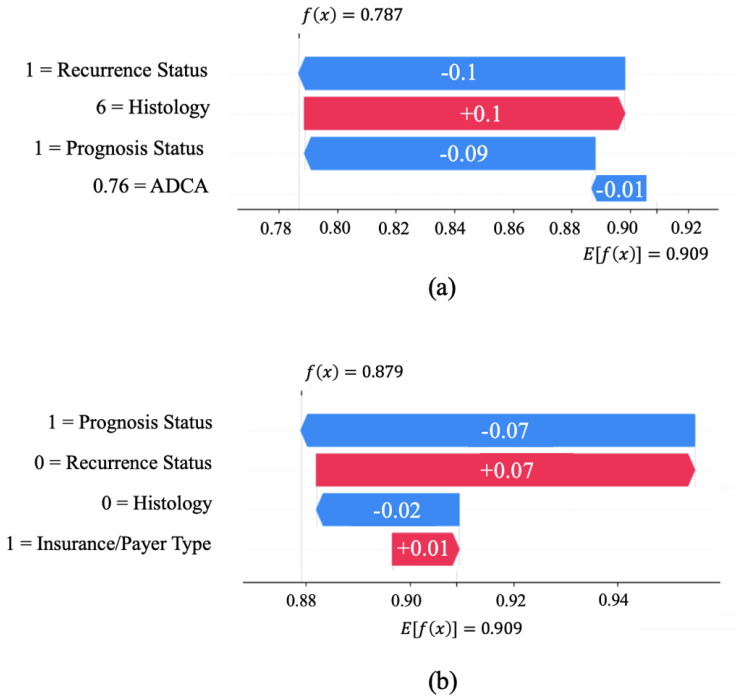

Figure 8. Local perspective of SHAP values for two patients with waterfall plots: (a) SHAP values for the first patient; (b) SHAP values for the second patient.

3.5 Conclusions

ALL is the most common type of leukemia among children and adolescents. It produces abnormal and immature white blood cells in the bone marrow, mainly lymphocytes, posing risks to various health complications and life-threatening events. Enhancing survival rates of

ALL relies on a reliable predictive strategy that facilitates timely monitoring and early interventions. Survival analysis, as a statistical approach, can predict the time until a health-related event occurs. However, most existing survival models rely on unrealistic assumptions and are limited in their learning capacity with simple model architectures. This paper aims to address these limitations by developing a Transformer-based deep survival model to predict patient-specific survival probabilities. Moreover, to enhance the model interpretability for healthcare professionals, we also explain the impacts of clinical attributes on the model output from both global and local perspectives with SHAP analysis.

Experimental results show that the censored group's average survival probability is consistently higher than that of the deceased group, suggesting the proposed survival model is accurate and robust in survival prediction. Moreover, we use the C-index as a performance metric to further evaluate the model. The C-index of the proposed model outperforms two commonly used deep survival models at three quartiles of the event time and achieves the highest value of 0.831 at the 75% quartile. In conclusion, these experimental results indicate that the proposed model can precisely predict patient-specific survival probabilities for children and adolescents diagnosed with ALL. There are several directions for future work. One is to study sequential modeling for time-series survival data that exhibit time-varying clinical attributes. Another one is to investigate the methods for incorporating model uncertainty into deep learning-based survival models, which aims to account for variability in prediction outcomes, ultimately improving the accuracy and robustness of patient-specific survival predictions.

CHAPTER 4

A Bayesian Transformer-based Survival Model for Recurrence Prediction of Pediatric Acute Lymphoblastic Leukemia

4.1 Introduction

Acute lymphoblastic leukemia (ALL) is an acute malignancy of the white blood cells that causes the bone marrow to overproduce immature cells by upsetting the usual environment [82]. These cells disrupt the immune system, increasing the risk of infection and death if not treated appropriately in a timely fashion. Recent cancer statistics also show that ALL is the most common type of leukemia among children and adolescents, accounting for 76% of children and adolescents with leukemia in the United States [46]. Among all potential clinical attributes of ALL, recurrence remains a significant factor that affects survival from ALL. Children with recurrence of ALL have a lower survival probability than those that do not experience a recurrence [83]. Therefore, predicting the recurrence of ALL becomes urgent as healthcare practitioners can use prediction results to provide timely and targeted medical care to children with ALL, thereby improving corresponding survival.

Conventionally, predictive approaches in healthcare range from machine learning methods to deep learning algorithms. Machine learning techniques like decision trees [84], logistic regression [85], and support vector machines [86] are grounded in solid mathematical foundations and have the advantage of interpretability. However, they may lack the capacity to process large volumes of data or capture complex, nonlinear relationships as effectively as deep learning models [87]. Researchers mainly utilize deep learning methods to make predictions, especially deep neural networks (DNNs). Mimicking the complexity of the

human brain, DNNs consist of an input layer to receive the data, multiple hidden layers that compute and transform the input into higher levels of abstraction, and an output layer that delivers the final decision or prediction [88]. However, DNNs often require substantial computational resources, leading to a significant computational burden. Most importantly, most existing machine learning and deep learning models rely on a single best model, lacking in considering and quantifying model uncertainty. Prediction results can be less accurate with model uncertainty neglected [89]. Therefore, in this paper, we addressed the issues and proposed a deep learning model by integrating the Transformer encoder and Bayesian neural networks (BNNs) to predict the recurrence of ALL-diagnosed children by quantifying model uncertainty.

The transformer model [53], first presented by Vaswani et al., has completely changed the area of deep learning by providing a unique architecture that draws global dependencies between input and output by relying on a mechanism called attention. Because of the design of this model, training is parallelized, and computing performance is significantly increased, allowing sequences to be processed in their entirety [90]. Because of their adaptability and firm performance, Transformers have recently been used and demonstrated success in various fields, including text categorization [91] and speech recognition [92]. BNNs infuse the principles of Bayesian probability within the structure of neural networks to estimate the uncertainty in predictions, providing a mathematical framework to account for model uncertainty [93]. Specifically, BNNs use probability distributions to represent the network weights, enabling the model to express confidence in its outputs and output an optimal distribution of model parameters, accounting for model uncertainty. In this paper, we use the Transformer encoder to feature extracting along with a BNN layer to output recurrence

probabilities and quantify model uncertainty. In addition, the output vectors from the Transformer encoder are clustered into two groups through the K-means algorithm to assess model performance.

The outline of this paper is as follows. Section 4.2 introduces various machine learning, deep learning, and model uncertainty applications in the medical domain. Section 4.3 presents the research methodology. Section 4.4 presents experimental results to evaluate the proposed method. Section 4.5 concludes the research.

4.2 Research Background

This section reviews existing machine learning, DNNs, and model uncertainty applications in the medical domain. Most machine learning methods have demonstrated strong potential in predictive tasks but struggle to capture complicated nonlinear relationships among variables. In addition, DNNs can produce more accurate results but suffer from substantial computational costs. Most importantly, model uncertainty is less considered in predictive approaches, leading to less precise prediction results. Therefore, it calls for an accurate, low-cost deep model that accounts for model uncertainty.

4.2.1 Predictive Modeling using Machine Learning Approaches

Machine learning (ML) is a subset of artificial intelligence that focuses on developing algorithms capable of learning from and making predictions or decisions based on data [94]. ML empowers computers to detect patterns and connections within data collections, fostering predictions grounded in empirical evidence, aiming to produce more accurate predictions. The fundamental principle involves training algorithms on a specific dataset, allowing data-driven predictions and decision-making rather than following static program instructions. ML is broadly divided into supervised and unsupervised learning, each with

unique approaches widely used in healthcare.

Supervised learning models are trained on a labeled dataset, pairing each training example with a corresponding output label, allowing the model to learn a mapping relationship between input variables and consequent outputs. Common supervised learning algorithms include linear regression for predicting continuous outcomes, logistic regression for binary classification, support vector machines (SVMs) for margin-based classification, and decision trees for hierarchical decision-making. For example, previous research applied logistic regression to identify critical factors' contribution to cancer detection [95]. In addition, researchers utilized a decision tree for early detection of breast cancer [96]. Moreover, SVMs are applied to classify brain tumors, assisting in post-processing the extracted region, like tumor segmentation [97]. Machine learning has increasingly become a powerful tool for predictive analytics across numerous fields, demonstrating significant promise in its ability to analyze large datasets and uncover patterns that can predict future outcomes accurately. ML has increasingly become a powerful tool for predictive analytics in the healthcare domain, demonstrating significant promise in its ability to analyze medical datasets and uncover patterns that can predict future outcomes accurately.

In contrast, unsupervised learning algorithms are used when the information about the output labels is unknown. The primary objective is to explore the underlying structure or distribution in the data, discovering patterns without the guidance of a known outcome. Unsupervised learning is practical for exploratory data analysis, anomaly detection, and complex data generation. There are many unsupervised learning methods applied to historical research. For example, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering approach that defines clusters as areas of high density

separated by areas of low density. K-means is another popular unsupervised learning algorithm for clustering the data into K groups by minimizing the variance within each cluster. In healthcare, there are plenty of applications using K-means. For example, researchers used K-means to reduce the dimension of predicting survival outcomes in patients with breast cancer [98]. Additionally, K-means clustering was utilized to classify acute leukemia [99]. Moreover, researchers developed an enhanced K-means clustering method for cancer subtype classification from gene expression data [100]. K-means remains a prevalent clustering technique due to its ease of implementation and effectiveness in clinical applications. These characteristics ensure its ongoing relevance in clinical data analysis and beyond.

4.2.2 Applications of DNNs in the Medical Domain

Machine learning methods have been widely adopted in healthcare because of their profound capability to process and analyze medical records that contain intricate interdependencies among variables. Nonetheless, its capacity to assimilate high-dimensional datasets often falls short. Compared to traditional machine learning methods, DNNs have a superior capacity for managing and interpreting data across extensive datasets, owing to their layered architecture and sophisticated training algorithms, enabling them to identify subtle patterns and dependencies in the data more effectively than conventional machine learning approaches.

DNNs comprise multiple interconnected layers of artificial neurons, extracting features from raw data for a more informative representation. Each layer in a DNN transforms its input data into more abstract representations, effectively building a hierarchy of learned features. The ability of DNNs to learn directly from raw data minimizes the need

for feature engineering, which is a significant advantage over traditional machine learning models. However, this depth comes with increased computational costs and the need for substantial training data to achieve remarkable performance. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are two specialized types of DNN architectures designed for handling different types of data and tasks [101,102]. RNNs are generally designed to process sequential data, such as time series data, while CNNs usually work on image data. Both of the approaches have shown strong model performance. However, RNNs and CNNs bring huge computational costs because of lacking parallel learning capability. Therefore, the Transformer model was developed and widely applied in various fields.

Introduced by Vaswani and colleagues in 2017, the transformer model has radically transformed deep learning with its novel structure that captures comprehensive dependencies between inputs and outputs solely through an attention mechanism. This configuration facilitates parallelized training, markedly enhancing computational efficiency, and permits the processing of entire sequence data simultaneously. Due to its versatility and exceptional performance, the transformer architecture has seen rapid adoption and achieved remarkable success in clinical applications. In 2021, a novel approach to survival analysis using Transformer-based models was introduced, emphasizing time-variant modeling of survival data [103]. In this methodology, patients are analogized to sentences, with individual "words" representing the combined feature embeddings and positional encodings of specific time points, discretely measured. Following this, in 2022, the SurvTRACE model was created to address multiple competing events by employing a multi-task learning approach that utilizes a common backbone structure [104]. These advancements have shown that the

application of Transformer technology can lead to more precise survival prediction models. However, neither of these models considers model uncertainty. Prediction results can be brittle with model uncertainty ignored.

4.2.3 Predictive Modeling with Model Uncertainty

Model uncertainty refers to the lack of definite knowledge of the observed process's actual function [70]. This type of uncertainty can result in varying predicted outcomes for the same input data. Specific inputs and model parameters typically generate a singular predictive outcome distribution. However, altering the parameters can lead to various models, each producing a unique predictive outcome distribution [105]. Hence, a range of possible model functions implies a corresponding range of potential predictive distributions, leading to significant uncertainty in predictive modeling.

Conventionally, extensive studies have explored the significance of quantifying model uncertainty in predictive modeling. For example, researchers provide an example of the application of Bayesian model averaging in predicting the spatial distribution of an arboreal marsupial in the Eden region of southeastern Australia [106]. In addition, studies obtained improved credit risk estimates and associated uncertainties using the Bayesian model [107]. In these studies, a straightforward elucidation of model uncertainty is pivotal in refining the models' accuracy. By rigorously quantifying the uncertainty inherent in the model predictions, researchers can better assess the reliability of their results and the potential variability in the outcomes, allowing more informed decision-making on predictions.

A popular deep learning model that considers model uncertainty is Bayesian Neural Networks (BNNs). BNNs integrate the principles of Bayesian inference with neural network structures, providing a statistical approach to interpreting deep learning outcomes. By

considering the weights and biases as variable distributions rather than fixed values, BNNs can evaluate the reliability of predictive outcomes. Such an ability to assess reliability is especially valuable in scenarios that demand critical decision-making (e.g., in the healthcare domain), where knowing the degree of certainty in predictions is imperative. A broad range of studies has demonstrated the significance of BNNs. For example, an existing work uses a Bayesian neural network for modeling and predicting the probability of gastric cancer patient death, indicating that the accuracy of BNNs is higher than that of artificial neural networks [108]. This work also illustrates that including model uncertainty aids in enhancing the model performance, especially in the clinical domain, to inform critical decisions.

4.3 Research Methodology

This section presents the methodology of the proposed survival model to make recurrence predictions for children with ALL. As shown in Figure 9, we use a Transformer encoder for feature extraction along with BNNs to form the Bayesian Transformer-based survival model. Then, the outputs of the proposed model are used for K-means clustering, relapse prediction, and model uncertainty quantification.

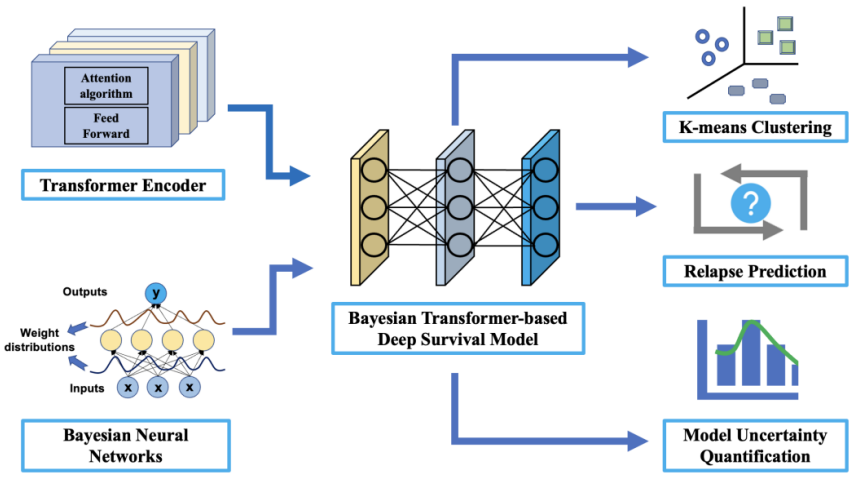

Figure 9. Flowchart of Research Methodology

4.3.1 Data Preprocessing and Problem Formulation of Survival Prediction

Table 5. Descriptions of Variables

Variable	Description
Age at Diagnosis	Identifies the age of the patient at diagnosis
Diagnosis Year	Identifies the year of diagnosis
DCAD	Represents the distance to care (in miles) from address at diagnosis to the Oklahoma Children's Hospital OU Health
ADAD	Represents the area deprivation index using national rankings based on address at diagnosis
DCCA	Represents the distance to care (in miles) from the current address in the EHR to the Oklahoma Children's Hospital OU Health
ADCA	Represents the area deprivation index based on the current address
Gender	Identifies the gender of the patient at diagnosis
Race/Ethnicity	Identifies the race and ethnicity information of the patient
Histology	Indicates ALL subtypes
Prognosis Status	Indicates the prognosis status based on the post-induction assessment
Insurance Type/Payer	Indicates the insurance type or primary payer of a patient
Recurrence Status	Indicates relapse or recurrence that occurred after complete remission following initial induction therapy

The data set used in this paper is from the OU Health Cancer Registry, OU Health electronic health records (EHR), and the Oklahoma Central Cancer Registry. We included 275 children and adolescents diagnosed with ALL from 2005 to 2019 [77]. In addition, we used 12 clinical variables for modeling in total. Among all variables, there are six numerical predictive variables, including age at diagnosis (0 to 19 years of age), diagnosis year, distance to care (in miles) from address at diagnosis (DCAD) to the Oklahoma Children's Hospital OU Health, area deprivation index using national rankings based on address at diagnosis (ADAD) [78], distance to care (in miles) from the current address (DCCA) in the EHR to the Oklahoma Children's Hospital OU Health, and area deprivation index based on

current address (ADCA). Moreover, there are five categorical predictive variables, including gender, histology (following the International Classification of Childhood Cancer 3rd Edition [79]), race/ethnicity (American Indian, Hispanic, Non-Hispanic (NH) Black, NH White, and NH Asian/Pacific Islander), prognosis status (based on post-induction prognosis), and insurance/payer type (Indian Health Service, Insured, Medicaid, Uninsured, Unknown). Further, the outcome variable is recurrence status, defined as relapse or recurrence after complete remission following initial induction therapy. We regularized continuous variables and performed one-hot encoding for categorical variables to enhance the learning ability of the model. To split the dataset, we divide the dataset into three segments: 70% of the data serves as the training set for model development, 10% as the validation set, and the remaining 20% as the test set. The training process stops when there are no further improvements in the performance of the validation set. Detailed descriptions of these variables are provided in Table 5.

4.3.2 Bayesian Transformer-based Deep Survival Model

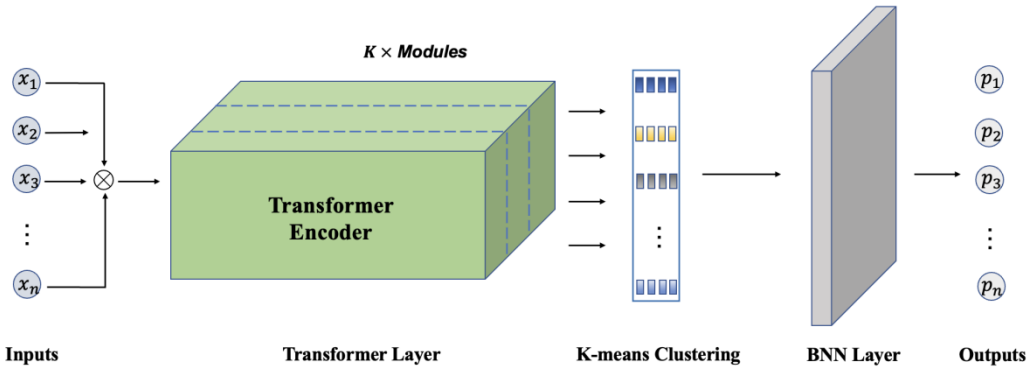

Figure 10. The Architecture of the Proposed Model

As shown in Figure 10, the architecture of the proposed model contains three components: (1) A Transformer encoder with several encoder blocks to perform layer

normalization, multi-head attention mechanism, dropout application, and residual connection; (2) A BNN layer to output the recurrence probabilities and to account for model uncertainty; (3) The output from the Transformer encoder is used for K-means clustering. Specifically, the process begins by feeding preprocessed input features into the Transformer encoder. Next, output vectors from the Transformer encoder are used for K-means clustering (detailed in 4.3.3). Finally, a BNN layer makes relapse predictions for pediatric ALL patients.

The loss function in BNNs fundamentally differs from the loss functions used in traditional neural networks due to the probabilistic nature of BNNs [31]. Specifically, we denote $\boldsymbol{w}$ as the model parameters in the entire network, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ is a normal distribution with learnable mean μ and diagonal covariance σ. Therefore, each w^i is obtained by drawing samples from the normal distribution (μ^i, σ^i). If we denote $q(\boldsymbol{w}|\boldsymbol{\theta})$ as the factorized weight posteriors, we train our Bayesian Transformer model by minimizing the Kullback-Leibler (KL) divergence [109] between the approximate weight posterior $q(\boldsymbol{w}|\boldsymbol{\theta})$ and the true but unknown posterior $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$, where $\boldsymbol{y}$ is the output, $\boldsymbol{X}$ represents the inputs. Accordingly, the loss function is defined as

$$L(\boldsymbol{\theta}) = \min KL[q(\boldsymbol{w}|\boldsymbol{\theta})||p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})]. \tag{4.1}$$

According to Bayesian theory and the variance inference approach, Equation (4.1) can be written as

$$L(\boldsymbol{\theta}) \propto \min KL[q(\boldsymbol{w}|\boldsymbol{\theta})||p(\boldsymbol{w})] - E_{q(\boldsymbol{w}|\boldsymbol{\theta})}[p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})], \tag{4.2}$$

where $p(\boldsymbol{w})$ represents the prior distribution of weight $\boldsymbol{w}$. Specifically, Equation (4.2) is equivalent to minimizing a KL divergence regularization term plus an expectation over the

negative log-likelihood term, $E_{q(\boldsymbol{w}|\boldsymbol{\theta})}[p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})]$.

In the experiment, $p(\boldsymbol{w})$ is a normal distribution with zero mean and 0.1 standard deviation for simplicity. In addition, a sigmoid function is applied on the BNN layer to output recurrence probabilities. Moreover, we use a fixed learning rate of {0.001,0.0001}. Further, the number of Transformer layers and heads is chosen from {2,4}, the hidden sizes are selected from {16,32}, and the batch size is set to 8.

4.3.3 Recurrence Identification using K-means Clustering

K-means clustering is a widely used unsupervised machine learning algorithm that sorts data into predetermined K clusters based on feature similarity [110]. The primary aim is to partition the dataset into K distinct groups such that each data point belongs to the cluster with the nearest mean, minimizing the squared differences between the data points and the cluster centroid [111]. In this paper, we aim to use the outputs from the Transformer encoder to apply the K-means approach, aiming to group the vectors into two identical groups.

Initially, we chose two random points from the data as starting centroids. Following this, each data point is assigned to the closest centroid, which is then updated to be the average of the points in its cluster. Specifically, the objective function Z of the K-means approach is defined as

$$Z = \min \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \boldsymbol{\alpha}_i\|^2. \tag{4.3}$$

where K is the number of clusters, $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m})$ is a set of m vectors, $\boldsymbol{S} = (S_1, S_2, \dots, S_K)$ represents a set of K clusters, and $\boldsymbol{\alpha}_i$ is the centroid of S_i,

$$\boldsymbol{\alpha}_i = \frac{1}{|S_i|} \sum_{\boldsymbol{x} \in S_i} \boldsymbol{x}, \tag{4.4}$$

where $|S_i|$ is the size of S_i.

Utilizing the K-means clustering algorithm, we can segregate pediatric ALL patients into two distinct groups, which enables us to identify patterns correlating with recurrence risk. This stratification can be instrumental in prognostication and inform decisions regarding the intensity of follow-up and interventions. The K-means approach hinges on validation against clinical outcomes to ensure its predictive reliability and carefully selects input features most indicative of recurrence risk.

4.4 Experimental Results

First, we choose five representative patients from the recurrence and non-recurrence groups to compute the patient-specific mean recurrence probability and corresponding standard deviation to quantify model uncertainty. Second, we compare the range of the recurrence probabilities and non-recurrence groups to validate the model performance. Third, we use the confusion matrix to calculate the accuracy and precision of the Bayesian Transformer-based deep survival model.

4.4.1 Quantification of Model Uncertainty

After the model is trained, we can obtain an optimal distribution of model parameters by optimizing the Bayesian Transformer model. Therefore, in the experiment, we draw samples from the optimal distribution to get a set of optimal weight tensors, thereby obtaining optimal prediction results. As shown in Figure 11, (a) is the recurrence probabilities of five patients from the recurrence group, and (b) is the recurrence probabilities of five patients from the non-recurrence group. The vertical dashed lines represent the mean recurrence probabilities for different patients. In addition, the red and blue shaded areas represent corresponding standard deviations, which accounts for model uncertainty.

We also list the recurrence probabilities of each patient from different groups. As

shown in Table 6, both the figure and the table indicate that the recurrence probabilities of the recurrence group are mostly greater than 0.5, while the recurrence probabilities of the non-recurrence group are mostly smaller than 0.3. This result indicates that our Bayesian Transformer model can discriminate whether a patient has a recurrence.

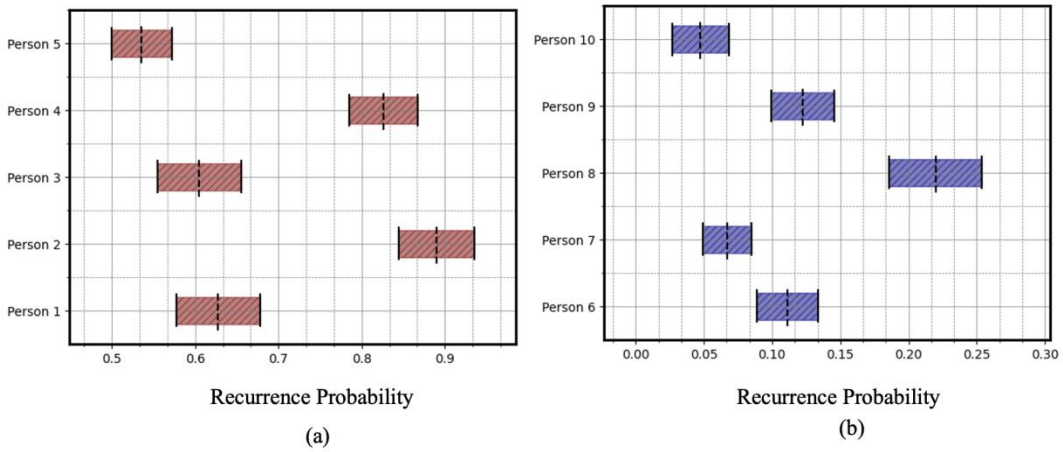

Figure 11. Plots of recurrence predictions under model uncertainty for: (a) recurrence group patients and (b) non-recurrence group patients. Vertical dashed lines represent the mean recurrence probabilities. Red and blue shaded areas indicate standard deviations that account for model uncertainty for recurrence and non-recurrence group patients.

Table 6. Recurrence Probabilities and standard deviations for ten patients from recurrence group and non-recurrence group

Recurrent Patient ID	Recurrence Probability	Standard Deviation	Non-recurrent Patient ID	Recurrence Probability	Standard Deviation
1	0.6278	0.0498	6	0.1112	0.0226
2	0.8903	0.0453	7	0.0671	0.0181
3	0.8254	0.0412	8	0.2199	0.0340
4	0.6047	0.0503	9	0.1225	0.0229
5	0.5357	0.0362	10	0.0475	0.0206

4.4.2 Performance Metric of the Bayesian Transformer Model

As we use K-means clustering to cluster the outputs from the Transformer encoder into two groups, we induce the confusion matrix as the performance metric of the proposed

model based on the test set [112]. There are fifty patients in the test set, 30 of whom are from the non-recurrence group, while 20 are from the recurrence group. As shown in Figure 12, 17 relapse patients are correctly predicted as recurrence group patients, while 24 non-relapse patients are correctly predicted as non-recurrence group patients.

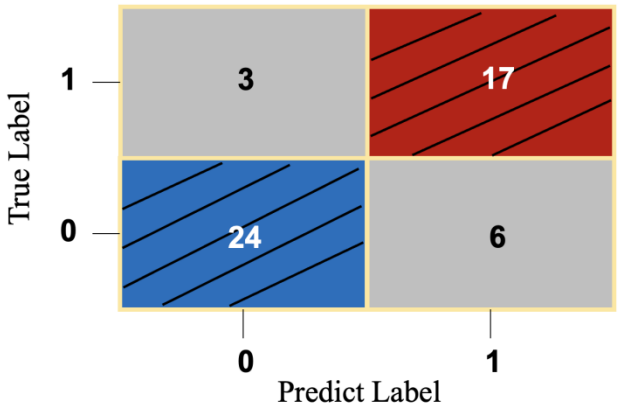

Figure 12. Confusion Matrix of the Bayesian Transformer-based Deep Survival Model

In addition, the precision measures the proportion of positive identifications that were correct, which is calculated by dividing the number of true positive predictions by the sum of true positive and false positive predictions:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4.5}$$

Another performance matrix is called accuracy, which is the overall correctness of the model and is calculated by dividing the sum of the true positive and true negative predictions by the total number of predictions:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions} \tag{4.6}$$

In this study, our model achieves a precision of 0.85 and an accuracy of 0.82, indicating the developed Bayesian Transformer model has strong model performance in predicting the recurrence probability.

4.5 Conclusions

ALL, being the most prevalent type of leukemia in children and young adults, presents significant challenges in treatment and prognosis. ALL is characterized by the production of abnormal white blood cells, which can cause various health issues and potentially fatal complications. Among all clinical attributes, recurrence status is essential in survival probabilities. Pediatric ALL patients who have recurrence tend to have a lower survival probability. Therefore, it calls for an effective and efficient predictive approach to predict the relapse of pediatric ALL patients, assisting clinicians in decision-making on prognosis and offering timely medical care.

Traditional machine learning and deep learning-based survival models struggle with huge computation costs and limited learning capacity due to their basic structural designs. Furthermore, most existing survival models seek to optimize the performance of a single best model, overlooking the inherent uncertainties in prediction outcomes as an ensemble of models with identical performance can exhibit variabilities in predictions. Prediction outcomes can be less precise without incorporating model uncertainty, leading to less reliable decisions on prognoses and treatments given to pediatric ALL patients. In response to these challenges, in this paper, we develop a Bayesian Transformer-based deep survival model to predict patient-specific recurrence probabilities. Through the Transformer encoder, we captured feature vectors for a more informative presentation. In addition, by leveraging Bayesian Neural Networks (BNNs), our model captures the uncertainties inherent in the prediction process, thus providing a more reliable measure of probability for recurrence. Furthermore, we use the K-means approach to cluster the output vectors from the Transformer encoder into two groups to identify whether a patient has a recurrence in the

lifetime.

Experimental results indicate that the range of recurrence probability of the recurrence group is significantly higher than that of the non-recurrence group, suggesting the model is accurate and robust in recurrence prediction. Moreover, we use the accuracy and precision obtained from the confusion matrix as the evaluation metric. The developed model achieves a precision of 0.85 and an accuracy 0.82, respectively. Consequently, these experimental results indicate that the proposed model precisely predicts patient-specific recurrence probabilities for ALL. Potential future research directions can include studying sequential modeling for time-series survival data that exhibit time-varying clinical features. With sequential clinical attributes involved, the model can obtain more timely information from the patient, thereby further improving the accuracy and robustness of patient-specific survival predictions.

CHAPTER 5

Conclusions and Future Works

In conclusion, the research presented in this dissertation contributes significantly to the field of pediatric oncology by addressing the challenge of predicting survival and recurrence probabilities in children with Acute Lymphoblastic Leukemia (ALL). First, we introduce a Bayesian survival model that integrates model uncertainty with a machine learning model. This approach enhances the accuracy of patient-specific survival predictions. Furthermore, the development of a Transformer-based explainable deep survival model represents a novel integration of advanced deep learning techniques with survival analysis. This model not only provides predictions for the time to occurrence of death but also offers explanations for the predictions, thereby increasing transparency and trust in the model's outputs. The third contribution, which combines the Transformer encoder with a Bayesian Neural Network (BNN) layer, further refines the prediction of recurrence probabilities. The inclusion of model uncertainty and the clustering of output features provide a robust method for assessing model performance and the significance of clinical variables. This work not only captured but also quantified model uncertainty through mean recurrence probabilities and corresponding standard deviations. Consequently, all three works form an integral, aiming to develop survival models of high precision. Therefore, healthcare practitioners can use these predictive models to produce accurate patient-specific survival and relapse probabilities, thereby informing early intervention strategies, ultimately improving survival rates, and addressing demographic disparities in healthcare access.

For future work, several directions can be explored. Firstly, the implementation of the

recurrence prediction can be improved by predicting the recurrence probability across the event-free time. Instead of predicting a particular probability for a single patient, predicting the function of relapse probability over time can help clinicians provide timely medical care to children with leukemia. Secondly, expanding the dataset and including a broader spectrum of demographic and genetic factors may enhance the models' predictive power. Thirdly, exploring time-varying deep survival models with model uncertainty is also a crucial part. In time-varying survival analysis, values of clinical attributes change over time. Therefore, potential models can collect more insightful information according to such changes.

BIBLIOGRAPHY

[1] J. Wiemels, "Perspectives on the Causes of Childhood Leukemia," *Chem. Biol. Interact.*, vol. 196, no. 3, pp. 59–67, 2012.

[2] T. P. Whitehead, C. Metayer, J. L. Wiemels, A. W. Singer, and M. D. Miller, "Childhood Leukemia and Primary Prevention," *Curr. Probl. Pediatr. Adolesc. Health Care*, vol. 46, no. 10, pp. 317–352, 2016.

[3] R. E. Goldsby *et al.*, "Long-term Sequelae in Survivors of Childhood Leukemia with Down Syndrome: A Childhood Cancer Survivor Study Report," *Cancer*, vol. 124, no. 3, pp. 617–625, 2018.

[4] S. M. Namayandeh, Z. Khazaei, M. Lari Najafi, E. Goodarzi, and A. Moslem, "Global Leukemia in Children 0-14 Statistics 2018, Incidence and Mortality and Human Development Index (HDI): GLOBOCAN Sources and Methods," *Asian Pacific Journal of Cancer Prevention*, vol. 21, no. 5, pp. 1487–1494, 2020.

[5] L. E. Winestone and R. Aplenc, "Disparities in Survival and Health Outcomes in Childhood Leukemia," *Curr. Hematol. Malig. Rep.*, vol. 14, no. 3, pp. 179–186, 2019.

[6] P. Wang, Y. Li, and C. K. Reddy, "Machine Learning for Survival Analysis," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.

[7] T. Nakagawa, M. Ishida, J. Naito, A. Nagai, S. Yamaguchi, and K. Onoda, "Prediction of Conversion to Alzheimer's Disease using Deep Survival Analysis of MRI Images," *Brain Commun.*, vol. 2, no. 1, 2020.

[8] M. J. Pencina, R. B. D'Agostino, M. G. Larson, J. M. Massaro, and R. S. Vasan, "Predicting the 30-Year Risk of Cardiovascular Disease," *Circulation*, vol. 119, no. 24, pp. 3078–3084, 2009.

[9] J. W. Eley, "Racial Differences in Survival from Breast Cancer," *JAMA*, vol. 272, no. 12, p. 947, 1994.

[10] J.-H. Tseng and M.-Y. Tseng, "Survival Analysis of Children with Primary Malignant Brain Tumors in England and Wales: A Population-Based Study," *Pediatr. Neurosurg*, vol. 42, no. 2, pp. 67–73, 2006.

[11] K. Ren *et al. Deep Recurrent Survival Analysis.* (2018). Accessed: July 2022. Online. Available: https://arxiv.org/abs/1809.02403.

[12] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival Analysis Part I: Basic Concepts and First Analyses," *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[13] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958.

[14] L. J. Wei, "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis," *Stat. Med.*, vol. 11, no. 14–15, pp. 1871–1879, 1992.

[15] J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, and T. H. Scheike, Eds., *Handbook of Survival Analysis*. Chapman and Hall/CRC, 2016.

[16] E. Wit, E. van den Heuvel, and J.-W. Romeijn, "'All Models are Wrong...': An Introduction to Model Uncertainty," *Stat. Neerl.*, vol. 66, no. 3, pp. 217–236, 2012.

[17] M. Clyde and E. I. George, "Model Uncertainty," *Statistical Science*, vol. 19, no. 1, 2004.

[18] T. N. Palmer, G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher, "Representing Model Uncertainty in Weather and Climate Prediction," *Annu. Rev. Earth Planet Sci.*, vol. 33, no. 1, pp. 163–193, 2005.

[19] T. Nilsen and T. Aven, "Models and Model Uncertainty in the Context of Risk Analysis," *Reliab. Eng. Syst. Saf.*, vol. 79, no. 3, pp. 309–317, 2003.

[20] J. Yin, J. Medellín-Azuara, A. Escriva-Bou, and Z. Liu, "Bayesian Machine Learning Ensemble Approach to Quantify Model Uncertainty in Predicting Groundwater Storage Change," *Science of The Total Environment*, vol. 769, pp. 144715, 2021.

[21] J.-T. Chien and Y.-C. Ku, "Bayesian Recurrent Neural Network for Language Modeling," *IEEE Trans. Neural Netw. Learn Syst.*, vol. 27, no. 2, pp. 361–374, 2016.

[22] M. Ellison, "Bayesian inference in ecology," *Ecol. Lett.*, vol. 7, no. 6, pp. 509–520, 2004.

[23] S. Brooks, "Markov chain Monte Carlo method and its application," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 1, pp. 69–100, 1998.

[24] M. Y. Wang and T. Park, “A Brief Tour of Bayesian Sampling Methods,” in *Bayesian Inference on Complicated Data*, Nisheng Tang, Ed., London, United of Kingdom: IntechOpen, 2020.

[25] K. Gallagher, K. Charvin, S. Nielsen, M. Sambridge, and J. Stephenson, “Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems,” *Mar. Pet. Geol.*, vol. 26, no. 4, pp. 525–535, 2009.

[26] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, “Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4133–4145, 2006.

[27] E. I. George and R. E. McCulloch, “Variable Selection via Gibbs Sampling,” *J. Am. Stat. Assoc.*, vol. 88, no. 423, pp. 881–889, 1993.

[28] G. Flötteröd and M. Bierlaire, “Metropolis–Hastings sampling of paths,” *Transportation Research Part B: Methodological*, vol. 48, pp. 53–66, 2013.

[29] W. Wang and J. Yan, “Shape-Restricted Regression Splines with R Package splines2,” *Journal of Data Science*, pp. 498–517, 2021.

[30] J. O. Ramsay, “Monotone Regression Splines in Action,” *Statistical Science*, vol. 3, no. 4, 1988.

[31] M. W. Dusenberry *et al.*, “Analyzing the Role of Model Uncertainty for Electronic Health Records,” in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 204–213.

[32] S. L. Brilleman, E. M. Elci, J. B. Novik, and R. Wolfe. *Bayesian Survival Analysis Using the rstanarm R Package.* (2020*)*. Accessed: July 2022. Online. Available: http://arxiv.org/abs/2002.09633.

[33] M. Girolami and B. Calderhead, “Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods,” *J. R. Stat. Soc. Series B. Stat. Methodol.*, vol. 73, no. 2, pp. 123–214, 2011.

[34] R. Kelter, “Analysis of Bayesian Posterior Significance and Effect Size Indices for the Two-sample T-test to Support Reproducible Medical Research,” *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 88, 2020.

[35] C. C. Monnahan, J. T. Thorson, and T. A. Branch, “Faster Estimation of Bayesian Models in Ecology using Hamiltonian Monte Carlo,” *Methods Ecol. Evol.*, vol. 8, no. 3, pp. 339–348, 2017.

[36] M. Betancourt, “The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo,” *Ann. Phys.*, vol. 531, no. 3, pp. 1700214, 2019.

[37] P. A. M. Dirac, “Generalized Hamiltonian Dynamics,” *Canadian Journal of Mathematics*, vol. 2, pp. 129–148, 1950.

[38] M. A. da Silva, E. S. B. de Oliveira, A. A. von Davier, and J. L. Bazán, “Estimating the DINA Model Parameters using the No-U-Turn Sampler,” *Biometrical Journal*, vol. 60, no. 2, pp. 352–368, 2018.

[39] A. K. Mazur, “Common Molecular Dynamics Algorithms Revisited: Accuracy and Optimal Time Steps of Störmer–Leapfrog Integrators,” *J. Comput. Phys*., vol. 136, no. 2, pp. 354–365, 1997.

[40] SEER*Stat Database: Incidence - SEER Research Plus Data, 17 Registries, Nov 2021 Sub (2000-2019) - Linked to County Attributes - Time Dependent (1990-2019) Income/Rurality, 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance, Epidemiology, and End Results (SEER) Program, 2022. Online. Available: https://seer.cancer.gov.

[41] R. B. D’Agostino and B.-H. Nam, “Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures,” *Handbook of Statistics*, vol. 23, pp. 1–25, 2003.

[42] F. E. Harrell, “Evaluating the Yield of Medical Tests,” *JAMA: The Journal of the American Medical Association*, vol. 247, no. 18, pp. 2543, 1982.

[43] S. Glen. "C-Statistic: Definition, Examples, Weighting and Significance." statisticshowto.com. https://www.statisticshowto.com/c-statistic/. (accessed July 2022).

[44] C.-H. Pui, M. V. Relling, and J. R. Downing, “Acute Lymphoblastic Leukemia,” *New England Journal of Medicine*, vol. 350, no. 15, pp. 1535–1548, Apr. 2004, doi: https://doi.org/10.1056/nejmra023001.

[45] C.-H. Pui and W. E. Evans, “Treatment of Acute Lymphoblastic Leukemia,” *New England Journal of Medicine*, vol. 354, no. 2, pp. 166–178, Jan. 2006, doi: https://doi.org/10.1056/nejmra052603.

[46] R. L. Siegel, A. N. Giaquinto, and A. Jemal, “Cancer statistics, 2024,” *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, Jan. 2024, doi: https://doi.org/10.3322/caac.21820.

[47] National Cancer Institute, “Acute Lymphocytic Leukemia - Cancer Stat Facts,” SEER, 2018. https://seer.cancer.gov/statfacts/html/alyl.html

[48] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, “Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods,” *British Journal of Cancer*, vol. 89, no. 3, pp. 431–436, Aug. 2003, doi: https://doi.org/10.1038/sj.bjc.6601119.

[49] N. Benítez-Parejo, M. M. Rodríguez del Águila, and S. Pérez-Vicente, “Survival analysis and Cox regression,” *Allergologia et Immunopathologia*, vol. 39, no. 6, pp. 362–373, Nov. 2011, doi: https://doi.org/10.1016/j.aller.2011.07.007.

[50] J.-H. Tseng and M.-Y. Tseng, “Survival Analysis of Children with Primary Malignant Brain Tumors in England and Wales: A Population-Based Study,” *Pediatric Neurosurgery*, vol. 42, no. 2, pp. 67–73, 2006, doi: https://doi.org/10.1159/000090458.

[51] S. V. Archontoulis and F. E. Miguez, “Nonlinear Regression Models and Applications in Agricultural Research,” *Agronomy Journal*, vol. 107, no. 2, pp. 786–798, Mar. 2015, doi: https://doi.org/10.2134/agronj2012.0506.

[52] L. Yang and K. Pelckmans, “Machine Learning Approaches to Survival Analysis: Case Studies in Microarray for Breast Cancer,” *International Journal of Machine Learning and Computing*, vol. 4, no. 6, pp. 483–490, 2014, doi: https://doi.org/10.7763/ijmlc.2014.v6.459.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[54] P.-F. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, “Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 182–192, Jan. 2018, doi: https://doi.org/10.18653/v1/d18-1017.

[55] J. Li, X. Wang, Z. Tu, and M. R. Lyu, “On the diversity of multi-head attention,” *Neurocomputing*, vol. 454, pp. 14–24, Sep. 2021, doi: https://doi.org/10.1016/j.neucom.2021.04.038.

[56] N. Moritz, T. Hori, and J. Le, “Streaming Automatic Speech Recognition with the Transformer Model,” *IEEE Xplore*, May 01, 2020. https://ieeexplore.ieee.org/abstract/document/9054476.

[57] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models," *ACM Computing Surveys*, Aug. 2023, doi: https://doi.org/10.1145/3617680.

[58] Murat Tezgider, B. Yildiz, and G. Aydin, "Text classification using improved bidirectional transformer," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 9, Jul. 2021, doi: https://doi.org/10.1002/cpe.6486.

[59] M. Hermans and B. Schrauwen, "Training and Analysing Deep Recurrent Neural Networks," *Advances in Neural Information Processing Systems*, vol. 26, pp. 190–198, Jan. 2013.

[60] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of Machine Learning Models Using Improved Shapley Additive Explanation," *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Sep. 2019, doi: https://doi.org/10.1145/3307339.3343255.

[61] O. Aalen, "Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models," *The Annals of Statistics*, vol. 6, no. 3, pp. 534–545, May 1978, doi: https://doi.org/10.1214/aos/1176344198.

[62] S. J. Cutler and F. Ederer, "Maximum utilization of the life table method in analyzing survival," *Journal of Chronic Diseases,* vol. 8, no. 6, pp. 699–712, Dec. 1958, doi: https://doi.org/10.1016/0021-9681(58)90126-7.

[63] M. J. Crowther and P. C. Lambert, "A general framework for parametric survival analysis," *Statistics in Medicine*, vol. 33, no. 30, pp. 5280–5297, Dec. 2014, doi: https://doi.org/10.1002/sim.6300.

[64] R. TIBSHIRANI, "THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, Feb. 1997, doi: https://doi.org/10.1002/(sici)1097-0258(19970228)16:4%3C385::aid-sim380%3E3.0.co;2-3.

[65] J. BUCKLEY and I. JAMES, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979, doi: https://doi.org/10.1093/biomet/66.3.429.

[66] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, Jan. 1972, doi: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[67] C. W. J. GRANGER, “Strategies for Modelling Nonlinear Time-Series Relationships,” *Economic Record*, vol. 69, no. 3, pp. 233–238, Sep. 1993, doi: https://doi.org/10.1111/j.1475-4932.1993.tb02103.x.

[68] Hong Wang and Gang Li, “A Selective Review on Random Survival Forests for High Dimensional Data,” *Quantitative Bio-Science*, vol. 36, no. 2, pp. 85–96, Nov. 2017, doi: https://doi.org/10.22283/qbs.2017.36.2.85.

[69] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, no. 3, Sep. 2008, doi: https://doi.org/10.1214/08-aoas169.

[70] Y. Cui, Y. Li, C. Pan, S. R. Brown, R. E. Gallant, and R. Zhu, “Bayesian inference for survival prediction of childhood Leukemia,” *Computers in Biology and Medicine*, vol. 156, p. 106713, Apr. 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106713.

[71] N. Rusk, “Deep learning,” *Nature Methods*, vol. 13, no. 1, pp. 35–35, Jan. 2016, doi: https://doi.org/10.1038/nmeth.3707.

[72] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Medical Research Methodology*, vol. 18, no. 1, Feb. 2018, doi: https://doi.org/10.1186/s12874-018-0482-1.

[73] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar, “DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: https://doi.org/10.1609/aaai.v32i1.11842.

[74] J. Chen, W. Koju, S. Xu, and Z. Liu, “Sales Forecasting Using Deep Neural Network And SHAP techniques,” *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Mar. 2021, doi: https://doi.org/10.1109/icbaie52039.2021.9389930.

[75] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, “Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach,” *Engineering Structures*, vol. 219, p. 110927, Sep. 2020, doi: https://doi.org/10.1016/j.engstruct.2020.110927.

[76] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, “Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival,” *Scientific Reports*, vol. 11, no. 1, Mar. 2021, doi: https://doi.org/10.1038/s41598-021-86327-7.

[77] A. E. Janitz et al., “Measuring disparities in event-free survival among children with acute lymphoblastic leukemia in an academic institute in Oklahoma, 2005–2019,” *Cancer epidemiology (Print)*, vol. 81, pp. 102275–102275, Dec. 2022, doi: https://doi.org/10.1016/j.canep.2022.102275.

[78] E. Steliarova-Foucher et al., “International incidence of childhood cancer, 2001–10: a population-based registry study,” *The Lancet Oncology*, vol. 18, no. 6, pp. 719–731, Jun. 2017, doi: https://doi.org/10.1016/s1470-2045(17)30186-9.

[79] A. J. H. Kind and W. R. Buckingham, “Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas,” *New England Journal of Medicine*, vol. 378, no. 26, pp. 2456–2458, Jun. 2018, doi: https://doi.org/10.1056/nejmp1802313.

[80] S. Lundberg, P. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, 2017, Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[81] R. Rodríguez-Pérez and J. Bajorath, “Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions,” *Journal of Computer-Aided Molecular Design*, vol. 34, no. 10, pp. 1013–1026, May 2020, doi: https://doi.org/10.1007/s10822-020-00314-0.

[82] M. Onciu, “Acute Lymphoblastic Leukemia,” *Hematology/Oncology Clinics of North America*, vol. 23, no. 4, pp. 655–674, Aug. 2009, doi: https://doi.org/10.1016/j.hoc.2009.04.009.

[83] C. N. Boyd, R. C. Ramberg, and E. Donnall. Thomas, “The incidence of recurrence of leukemia in donor cells after allogeneic bone marrow transplantation,” *Leukemia Research*, vol. 6, no. 6, pp. 833–

837, Jan. 1982, doi: https://doi.org/10.1016/0145-2126(82)90067-4.

[84] J. R. Quinlan, "Decision trees and decision-making," *IEEE Transactions on Systems*, Man, and Cybernetics, vol. 20, no. 2, pp. 339–346, 1990, doi: https://doi.org/10.1109/21.52545.

[85] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: https://doi.org/10.1080/00220670209598786.

[86] H. Bhavsar and M. Panchal, "A Review on Support Vector Machine for Data Classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. 2278–1323, 2012, Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d683a971524a0d76382ce335321b4b8189bc8299.

[87] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications,* vol. 09, no. 01, pp. 1–16, 2017, doi: https://doi.org/10.4236/jilsa.2017.91001.

[88] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017, doi: https://doi.org/10.1109/jproc.2017.2761740.

[89] I. Park, H. K. Amarchinta, and R. V. Grandhi, "A Bayesian approach for quantification of model uncertainty," *Reliability engineering & systems safety*, vol. 95, no. 7, pp. 777–785, Jul. 2010, doi: https://doi.org/10.1016/j.ress.2010.02.015.

[90] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," *IEEE Xplore*, Apr. 01, 2018. https://ieeexplore.ieee.org/document/8462506/.

[91] W. Cunha, F. Viegas, C. França, T. Rosa, L. Rocha, and M. A. Gonçalves, "A Comparative Survey of Instance Selection Methods applied to NonNeural and Transformer-Based Text Classification," *ACM Computing Surveys*, Jan. 2023, doi: https://doi.org/10.1145/3582000.

[92] Q. Zhang et al., "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," *IEEE Xplore*, May 01, 2020.

https://ieeexplore.ieee.org/abstract/document/9053896.

[93] J. Lampinen and A. Vehtari, “Bayesian approach for neural networks—review and case studies,” *Neural Networks*, vol. 14, no. 3, pp. 257–274, Apr. 2001, doi: https://doi.org/10.1016/s0893-6080(00)00098-8.

[94] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2020, doi: https://doi.org/10.1126/science.aaa8415.

[95] E. D. Pisano et al., “Cancer Cases from ACRIN Digital Mammographic Imaging Screening Trial: Radiologist Analysis with Use of a Logistic Regression Model,” *Radiology*, vol. 252, no. 2, pp. 348–357, Aug. 2009, doi: https://doi.org/10.1148/radiol.2522081457.

[96] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” *International Journal of Computer Applications*, vol. 98, no. 10, pp. 16–24, Jul. 2014, doi: https://doi.org/10.5120/17219-7456.

[97] T. S. Kumar, K. Rashmi, S. Ramadoss, L. K. Sandhya, and T. J. Sangeetha, “Brain tumor detection using SVM classifier,” *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, May 2017, doi: https://doi.org/10.1109/ssps.2017.8071613.

[98] M. Zhao, Y. Tang, H. Kim, and K. Hasegawa, “Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer,” *Cancer Informatics*, vol. 17, p. 117693511881021, Jan. 2018, doi: https://doi.org/10.1177/1176935118810215.

[99] Jakkrich Laosai and Kosin Chamnongthai, “Acute leukemia classification by using SVM and K-Means clustering,” *International Electrical Engineering Congress*, Mar. 2014, doi: https://doi.org/10.1109/ieecon.2014.6925840.

[100] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, “An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data,” *Computers in Biology and Medicine*, vol. 91, pp. 213–221, Dec. 2017, doi: https://doi.org/10.1016/j.compbiomed.2017.10.014.

[101] J. T. Connor, R. D. Martin, and L. E. Atlas, “Recurrent neural networks and robust time series prediction,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 240–254, Mar. 1994, doi: https://doi.org/10.1109/72.279188.

[102] S. Bhatnagar, Y. Afshar, S. Pan, K. Duraisamy, and S. Kaushik, “Prediction of aerodynamic flow fields using convolutional neural networks,” *Computational Mechanics*, vol. 64, no. 2, pp. 525–545, Jun. 2019, doi: https://doi.org/10.1007/s00466-019-01740-0.

[103] S. Hu, Egill Fridgeirsson, Guido van Wingen, and M. Welling, “Transformer-Based Deep Survival Analysis,” *Proceedings of Machine Learning Research*, vol. 146, pp. 132–148, May 2021.

[104] Z. Wang and J. Sun, “SurvTRACE: Transformers for Survival Analysis with Competing Events,” *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2022*, Aug. 2022, doi: https://doi.org/10.1145/3535508.3545521.

[105] J. B. Tenenbaum and T. L. Griffiths, “Generalization, similarity, and Bayesian inference,” *Behavioral and Brain Sciences*, vol. 24, no. 4, pp. 629–640, Aug. 2001, doi: https://doi.org/10.1017/s0140525x01000061.

[106] B. A. WINTLE, M. A. McCARTHY, C. T. VOLINSKY, and R. P. KAVANAGH, “The Use of Bayesian Model Averaging to Better Represent Uncertainty in Ecological Models,” *Conservation Biology*, vol. 17, no. 6, pp. 1579–1590, Dec. 2003, doi: https://doi.org/10.1111/j.1523-1739.2003.00614.x.

[107] R. Kazemi and A. Mosleh, “Improving Default Risk Prediction Using Bayesian Model Uncertainty Techniques,” *Risk Analysis*, vol. 32, no. 11, pp. 1888–1900, Nov. 2012, doi: https://doi.org/10.1111/j.1539-6924.2012.01915.x.

[108] M. H. Osman, “278PPredicting CA-125 status and ovarian cancer survival using artificial neural networks,” *Annals of Oncology*, vol. 29, no. suppl_9, Nov. 2018, doi: https://doi.org/10.1093/annonc/mdy436.025.

[109] J. R. Hershey and P. A. Olsen, “Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models,” *IEEE Xplore*, Apr. 01, 2007. https://ieeexplore.ieee.org/abstract/document/4218101

[110] S. Na, L. Xumin, and G. Yong, “Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm,” *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, Apr. 2010, doi: https://doi.org/10.1109/iitsi.2010.74.

[111] K. P. Sinaga and M.-S. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 8,

pp. 80716–80727, 2020, doi: https://doi.org/10.1109/access.2020.2988796.

[112] N. D. Marom, L. Rokach, and A. Shmilovici, “Using the confusion matrix for improving ensemble classifiers,” *IEEE Xplore*, Nov. 01, 2010. https://ieeexplore.ieee.org/abstract/document/5662159.