UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A TIME-AWARE LSTM APPROACH TO PREDICT TUMOR SIZE AND SURVIVAL MONTH IN NON-SMALL CELL LUNG CANCER

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

CHANGJAE KIM
Norman, Oklahoma
2022

A TIME-AWARE LSTM APPROACH TO PREDICT TUMOR SIZE AND SURVIVAL MONTH IN NON-SMALL CELL LUNG CANCER

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Talayeh Razzaghi, Chair

Dr. Dean Hougen

Dr. Charles D. Nicholson

I give thanks to God who loves and protects me in difficulties I have encountered through the entire journey in the United States. This thesis is also dedicated to my parents, Jin-young and Mi-ja Kim, my parents-in-law, Changseok and Kyungdeok Youn, and my family, Hye-lin Youn, Won-jae Kim and Min-jae Kim whose support has been unfaltering and enthusiastic, even when it was not easy to give it.

"In all this you greatly rejoice, though now for a little while you may have had to suffer grief in all kinds of trials. These have come so that the proven genuineness of your faith—of greater worth than gold, which perishes even though refined by fire—may result in praise, glory and honor when Jesus Christ is revealed."

1 Peter1:6-7

ACKNOWLEDGEMENTS

Most of all, I give thanks with all my heart to God whose grace allowed me to endure hardships. His love protects our family in any situations and lets us see the secret of facing plenty and hunger, abundance and need. Since pandemic situation, my family have still encountered some difficulties, but the hardships proved that he would not let us stray from your arms and he provided our need at the right time.

I am thankful for the unconditional support and love of my wife, Hyelin Youn, who has respected me and valued what I have done for my family. Without her, I cannot get to this point. She is the most precious gift which more than worth it in my liife. I am also blessed to have my son, Aaron Wonjae Kim, and my daugther, Annabeth Minjae Kim, during this pandemic situation. Their existence motivates me to strive for my family in any situations and gives me one of the happiest experiences in my life, being their father.

I give thanks to my parents, Jin-young and Mi-ja Kim. They have been a role model in my life. Their constant efforts and dedication to raise me and my younger sister have lighted up my life and their way of life suggested me to be a good person helping others but not a person seeking secular values. I am thankful for my younger sister, Ji-won Kim, who supported me to chase my dream to study abroad in financial hardship. I am also thankful to my parents-in-law, Changseok and Kyungduk Youn, who have cared for me in US, with all their love. Their Christianity gives me and my wife a way how to depend on God and how to be a Christian becoming a good influence on people. I also thank my sister-in law and her husband, Aelin Youn and Hamin Choi, and their son, Ethan Ahjun Choi, who encouraged me and stayed by my side.

I am really grateful for the support of Dr. Talayeh Razzaghi. She gave me a hand when I was in the most difficult situation in choosing career path and supported me to study a new field which

I have never experienced. She always encouraged me to do my best in any circumstances. I could not have reached this moment without her careful mentorship and support in my academic journey and life.

I would like to thank my committee members, Dr. Dean Hougen and Dr. Charles D. Nicholson. They provided informative comments and discussions for my research works. I am also thankful for the OU Supercomputing Center for Education and Research (OSCER). It enables me to run hundreds of hours of running machine learning models for this work. Finally, I am grateful to the University of Oklahoma's Data Science and Analytics Institute for their longstanding support. The family-like atmosphere and great care helpd me finalize my studies. I will not forget my time from the beginning to the end of my journey in Norman.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Recent advances in long short-term memory (LSTM) networks have enabled us to handle sequential and time-series data. However, some applications of LSTM networks in the healthcare domain have produced suboptimal performances, as the algorithm assumes constant elapsed times between consecutive elements of a patient health record. In reality, patient health records are heterogeneous information with irregular time intervals and different sequence lengths. The heterogeneity and temporal dynamics of the patients' data make it challenging to analyze long-timescale progression patterns of disease when we use traditional LSTM networks. This study proposes a novel LSTM architecture, called Time-Aware LSTM with power-law decay (T-pLSTM) networks, which can capture time irregularity and long-term dependency of patients' data. T-pLSTM can handle long-timescale patient records with irregular elapsed time by power-law forget gate and adjusted memory cell. The proposed model was tested to predict tumor size and survival month over time for non-small cell lung cancer (NSCLC) patients. The model was trained on patient records obtained from the Surveillance, Epidemiology, and End Results (SEER) Research Plus database, and its performance was evaluated by comparative analysis. The experiments using datasets with fixed and different sequence lengths showed that T-pLSTM outperformed the standard LSTM models. This result implies improvement of learning for long-term scale information with time irregularity in LSTM networks.

CHAPTER 1. INTRODUCTION

Technological advances have increased life expectancy more than before and the healthcare industry is investing huge amounts of money to solve challenging problems associated with the leading causes of death, like cancer. This chapter provides why my research is worthwhile in solving the problems in healthcare domain. In this section, subchapters describe what critical limitations motivated me to conduct my research related to the lung cancer patients and what the aim of this study is for the healthcare problem.

1.1. Research Motivation

Lung cancer is known as the leading cause of cancer-related deaths in the United States, ranked as the second most common cancer around the world (Siegel et al., 2013). The types of lung cancer are mainly classified as small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), the latter accounting for 80% to 85% of lung cancer incidents (Ries and Eisner, 2007). The American Lung Association has reported that the 5-year survival rate of lung cancer patients is approximately 18.6% which is lower than many other leading cancers such as breast, colorectal, and prostate cancers (American Lung Association, 2022). For instance, adenocarcinoma, one of the most common subtypes of NSCLC, has a high 5-year survival rate (approximately 63%) for early-stage patients, but surgery in cases with poor prognosis has a 35% risk of relapse (Hoffman et al., 2000). In hospitals, estimating survival time of lung cancer patients is highly affected by the clinician's knowledge and experience even if the estimates are imprecise as the decision is very subjective (Bartholomai et al., 2018). Another factor is that the patients with same stage of cancer might have different survival rates due to independent prognostic variables such as age, sex, histology, and so on, and thereby the estimation can be wide of the mark (Clément-Duchêne et al.,

2010; Bartholomai et al., 2018). Although several researchers have developed statistical methods to assist clinical decisions for NSCLC patients, the rapid growth of cancer and metastasis have complicated mechanisms to detect and diagnose early (Baeuerle and Gires, 2007; Barron et al., 2016; Chen et al., 2014a; Lai et al., 2020).

Machine learning has been increasingly popular in healthcare problems as the techniques can solve challenging problems from clinical research to hospital care to improve patient outcomes. Machine learning algorithms can provide prognostic analysis by learning from a larger volume and dimension of clinical data compared to existing clinical practice (Sun et al., 2018; Cutillo et al., 2020; Sumeet et al., 2022). Recent applications of machine learning in cancer research include a) early detection and survival prediction by capturing interdependent relationships between input features (or variables) and output response (or patient outcomes) and b) advancing cancer prognosis by recognizing interactions among biomarkers (Lai et al., 2020; Furey et al., 2000). Since the benefit of machine learning models is guaranteed by analyzing large volumes of patient records, researchers in cancer precision have extensively used the Surveillance, Epidemiology, and End Results (SEER) database, an authoritative repository of cancer statistics maintained by the National Cancer Institute in the United States (National Cancer Institute, 2008). This database encompasses patient information across several geographic regions in the United States, including patient demographics, survival month, and clinical information (e.g., cancer type, site, stage, and first course of treatment). Most previous studies have applied machine learning to predict survival rate and vital status of cancer patients based on comprehensive analysis of the SEER database for prognostic study of cancer disease (Agrawal et al., 2012; Bartholomai and Frieboes, 2018; Huang et al., 2019; Siah et al., 2019; Lu et al., 2020). However, existing models have been built based on static variables and ignored the effect of longitudinal variables. Capturing sequential information

between consecutive clinical events plays an important role in supporting decisions on clinical diagnosis and time-relevant knowledge can improve prediction performance of machine learning models for clinical prediction tasks (Choi et al., 2016; Chen et al., 2017; Baytas et al., 2017; Zhang et al., 2020).

Extracting longitudinal information from electronic health records (EHR) and population-based data increases the accuracy of predictive models for cancer prognosis, but large-scale heterogeneity of longitudinal patient records makes it difficult for clinicians to analyze and infer the high-level information embedded in the data. One well-known machine learning method to address this challenge is recurrent neural networks (RNNs) which enable capturing relationships and dependencies between consecutive elements of sequential data (Rumelhart et al., 1986; Che et al., 2018). However, as the time gap grows, the value of gradient becomes too small to learn information with long-term dependencies. Long short-term memory (LSTM) networks can handle the long-term dependencies by regulating a gated structure (Hochreiter and Schmidhuber, 1997). Traditional LSTM networks have been extended in many ways to improve prediction performance. One advanced model is bidirectional LSTM (Bi-LSTM) networks (SchusterKuldip and Paliwal, 1997) to preserve sequential information in two directions simultaneously (future to past and past to future). Training in two directions exploited in Bi-LSTM often leads to better performance than the traditional vanilla LSTM. With respect to time steps, standard LSTM networks use uniformly distributed time intervals between the elements of a sequence, but in reality, sequential events follow highly non-uniform distributions with different time gaps.

My work is motivated by Time-aware LSTM (T-LSTM) networks (Baytas et al., 2017) which take into account the time irregularity in learning from longitudinal healthcare data. T-LSTM networks demonstrate that adjusting forget gates impact training long time series data by

prolonging the memory of LSTM networks, and thus improve LSTM performance. Another modification of LSTM architecture is a slower forgetting mechanism developed by Chien et al. (2021). They introduced power-law forget gates to capture information for long-term dependencies. This advanced approach suggested that the power law coefficient should be smaller than one to capture long-range dependencies. As detailed in the literature review, several studies have used the time-aware LSTM method, but they have solely considered either time irregularity or long-timescale information so as to improve prediction outcomes. The aim of this study is to develop an advanced predictive method for capturing the long-timescale patient information with irregular time intervals. The proposed approach is applied to tumor size and survival month prediction by using longitudinal variables for NSCLC patients. I believe that these dynamic models will assist oncologists and clinicians for rapid and effective personalized treatment reassignment of NSCLC patients and open the door for building the patient's digital twin technology for precision medicine in the future.

1.2. Thesis Outline

In this study, I propose a novel LSTM approach to predict the tumor size and survival month using longitudinal patients' health information for NSCLC patients. The model captures time irregularity by adjusted memory cell and learns long-timescale dependency of the heterogeneous clinical data by power law forget gate. This work includes the following contributions in creating an accurate and efficient decision support tool for NSCLC patients.

- I extract longitudinal data for NSCLC patients whose primary cancer is lung cancer from SEER Research Plus data.
- I propose an advanced LSTM model to handle heterogeneous longitudinal data with time irregularities and long-term dependencies between consecutive records of the patient information. The proposed architecture improves the model performance for the longitudinal clinical data by adjusting the effect of previous memory and making a slower forgetting mechanism.
- I conduct a comparative analysis to show the efficacy of the proposed model and provide a guideline on how to use the developed method for prognostic research for NSCLC patients. The experimental investigation uses six types of LSTM architecture: a) vanilla and bidirectional LSTM networks (LSTM and Bi-LSTM); b) vanilla and bidirectional T-LSTM networks (T-LSTM and BiT-LSTM); and c) vanilla and bidirectional T-pLSTM networks (T-pLSTM and BiT-pLSTM). All models are trained by using various training datasets (i.e., patient records with the same and different sequence lengths) in order to investigate the effect of sequence lengths on prediction performance using SEER Research Plus data. The experiments show that my proposed model outperformed existing machine learning models in

predicting the tumor size and survival month for NSCLC patients but in some cases, lack of enough patient records degraded prediction performance of the T-pLSTM model.

This thesis is organized as follows: Chapter 2 explains technical details of how the advanced LSTM architecture is designed for the purpose of this study. Chapter 3 shows the initial data analysis for NSCLC patient records obtained from SEER Research Plus Data and correlations between selected features and output responses (tumor size and survival month). Chapter 4 presents experimental results based on comparative analysis and finally Chapter 5 summarizes how the proposed approach builds technical and clinical implications in healthcare.

CHAPTER 2. LITERATURE REVIEW

2.1. Predictive Analytics Models in Healthcare and Cancer Precision

Researchers in the healthcare industry have introduced various machine learning algorithms to promote better clinical decisions in prognostic analysis. Most previous studies have used machine learning methods to predict mortality and disease risk for patients and they recognize the importance of well-curated and well-organized clinical datasets to build accurate models (Fradkin, 2006; Chen et al., 2009; Krishnaiah et al., 2013; Weng et al., 2017; Sahni et al., 2018). Dimitoglu et al. (2012) use data mining techniques to extract valuable patterns from an extensive historical data set. They estimate applicability of the C4.5 algorithm and a Naive Bayes classifier in predicting survivability of lung cancer patients and confirm that the algorithms trained by valuable information extracted achieve prediction accuracy around 90%. Chen et al. (2014b) investigate how combining gene expression data with clinical data improves prediction outcomes in machine learning applications. They use various sets of artificial neural networks (ANNs) for the experiment to predict survival risks and show that models trained using gene expression data compute valid prediction outcomes for the survival classification. Lee et al. (2015) add patient similarity metric (PSM) from a cosine similarity-based calculation in predicting 30-day mortality prediction. They demonstrate that using the PSM values for training machine learning models outperforms existing intensive care unit (ICU) severity of illness score approaches by identifying similar patients from between patient records. Panahiazar et al. (2015) deploy machine learning models to predict heart failure survival and stratify its associated risk using electronic health records (EHRs) at Mayo Clinic. They use two types of variable sets, baseline dataset and the other one with additional variables, for their experiment and the results demonstrate the superiority of model using additional variables to determine heart failure survival risk. Lai et al. (2020) use a

systems biology approach to compute prognosis relevance values in order to identify novel prognostic gene biomarkers from the well-known biomarkers. Then, they train a deep neural network equipped with bimodal learning using both the selected biomarkers and clinical data to predict the 5-year survival status of NSCLC patients; their model leads into a high accuracy performance result (81.63% AUC).

In addition to data preparation, predictions in machine learning can be also improved by modifying the algorithms to extract valuable features from datasets. Lynch et al. (2017) compare prediction outcomes between traditional classification algorithms combined with their custom ensemble method assigning weight for prediction outputs of each model. The comparative study demonstrates a correlational approach coupled with supervised machine learning can be a valid way to provide a meaningful prediction in lung cancer patient survival prognosis. Huang et al. (2019) apply multivariable fractional polynomial (MFP) approach in developing machine learning models for mortality prediction of breast cancer patients. The technique determines the feature importance for each variable and their functional forms (nonlinear forms) to exclude some features for model improvement. The curated dataset is used to develop a multivariate Cox proportional hazards model for survivability prediction of breast cancer patients, and they also compare the outcomes between Asian and non-Asian patients. Huang et al. (2020) establish a predictive model using extreme gradient boosting (XGBoost) to predict 1-year survival status of NSCLC patients with bone metastases. They implement correlation analysis and feature selection to identify important variables for this study using the SEER database. The literature review shows significance of data preparation and advanced model development to help better decision-making using machine learning in healthcare but the fundamental issue behind the clinical applications is that none of these studies consider temporal information from clinical and epidemiological

datasets. In healthcare projects, it is crucial to deploy temporal information from longitudinal health records in risk stratification models as consecutive clinical events can accurately capture risk patterns and outcomes. The next subsection describes the application of longitudinal patients records and time-aware machine learning models to get improved prediction outcomes.

2.2. Predictive Analytics Using Longitudinal Clinical Data

Previous studies have extracted relevant information to identify meaningful patterns from EHR data by using RNN and LSTM (Lipton et al., 2016; Baytas et al., 2017; Che et al., 2018; Bai et al., 2018; Ruan et al., 2019; Yu et al., 2019). Despite the profound benefit of the algorithms to handle time series data, their use is limited to regular time series taking account for uniform time steps between consecutive elements of events, which is often not the case for many health datasets. For example, each patient has multiple encounters/visits with a healthcare provider and the timing of these encounters may vary for each patient. Particularly, EHR data has multivariate observations with irregular time intervals and thereby the existing methods fall short in handling patients' data with irregular sequences and interrelationships.

In order to address this problem, Lipton et al. (2016) train two machine learning models (linear regression model and RNN) by patient ICU records to investigate the effect of irregularly spaced missing data represented by a binary variable as well as other missing indicators. The comparison results demonstrate that RNN outperforms the linear regression model by recognizing binary indicators of missingness in the training dataset. Baytas et al. (2017) develop T-LSTM networks to understand longitudinal patient records with time irregularity and extract patterns for clinical prognosis. Using both synthetic and real datasets, their experiments show this advanced model capture temporal information with irregular time intervals. Furthermore, Bang et al. (2017)

propose an extended LSTM model, so-called as Phased-LSTM (P-LSTM), to handle longitudinal EHR data with missing values for disease diagnosis and prediction. They incorporate a decay rate as suggested by Che et al. (2018) in their model to deal with missing values. Bai et al. (2018) develop an interpretable deep learning model to capture the impact of both long-term chronic events and short-term acute events by learning time decay factors for every clinical code which is recorded in each visit indicative of patient's diagnosis and treatment during the visit. Their results demonstrate that the proposed method equipped with time decay factors outperforms RNN-based models. Xu et al. (2019) develop convolutional neural networks (CNNs) merged with RNNs using transfer learning to predict clinical outcomes based on time series CT images of lung cancer patients. The designed deep learning model demonstrate using time series scans significantly improves predicting survival and cancer-specific outcomes. Zhang et al. (2020) develop language models with time-aware layers to capture a multi-level sequential structure of clinical notes. Since time-aware layers use a flexible time decay function to reflect actual change of temporal importance on sequential patient records, adding them improves prediction performance of the proposed model compared to existing language models for clinical applications. Despite recent advances in the healthcare domain, existing LSTM networks use an exponential decay rate which restricts learning long-timescale patient's information. Some patients are treated by long-term medical care and thus we should consider the effect of these long-term successive clinical events in the predictive models.

CHAPTER 3. METHODOLOGY

This chapter describes details of how the proposed model learns long-timescale information with irregular time intervals from patient health records. As mentioned in the introduction, my work is motivated by neural networks capturing time-series information. RNNs are the first neural network algorithm to handle time series data. Although the algorithm has been applied to many research areas, its performance is degraded due to vanishing and exploding gradient problems (Hochreiter and Schmidhuber, 1997; Cho et al., 2014). Thus, many researchers have developed specialized versions of RNNs, such as LSTM and gated recurrent unit (GRU) networks to address the time-relevance problem (Sutskever et al., 2014; Cho et al., 2017; Che et al., 2018). Comparison between the two advanced algorithms demonstrated that GRUs with a smaller number of gates than LSTM showed faster training time, but LSTMs performed more accurately for a larger and high-level dataset (Yang et al., 2020). Thus, I used LSTMs in developing predictive models for tumor size and survival month prediction. In the next section, I will explain how each LSTM method learns time-relevant information and then what is the improvement of my proposed model. Finally, I will describe how to establish an experimental setup to demonstrate the novelty of my algorithm through comparative analysis.

3.1. Artificial Neural Networks (ANN) for Time Series Prediction

In the healthcare industry, researchers in major disease such as cancer or cardiology have used artificial neural networks (ANN) in their machine learning applications (Jiang et al., 2017). ANN inspired by McCulloch-Pitts Neuron (McCulloch and Pitts, 1990) mimics how a human brain processes information in solving non-linear problems. Figure 3.1 represents how a neuron in the ANN architecture processes information to get output signal. As shown in the figure, ANN is an interconnected system including several elements such as a) input neurons b) hidden neurons c) output neurons (Surguchev & Li, 2000; Xiong et al., 2020). In the ANN architecture, information gathered in the input neuron is transformed by the below function:

$$f(x; w, b) = w \cdot x + b \tag{1}$$

Where x is input vector, w and b are weight matrix and bias term in the layer, respectively. Then, the transformed input is used to get an output (e.g., prediction of the response variable) of the node by passing activation function u. The activation functions commonly used include the sigmoid function, hyperbolic tangent (tanh) function, and reflected linear unit (ReLU) functions (Xiong et al., 2020; Bennett, 2021). Advances in the algorithm have allowed us to perform time-series prediction. The well-known approach is recurrent neural networks (RNNs) capturing relationships and dependencies between consecutive elements of sequential data. The RNNs store previous information in the internal hidden states and combines the information with input vectors to transfer to next layers through recurrent connections (Rumelhart et al., 1986; Che et al., 2018; Chien et al., 2021). The algorithm is specialized in the sequence analysis process, but the method restricts to retain long period dependencies of information as the training process makes the value

representing training information go to zero exponentially fast, and researchers developed LSTM networks to overcome the problem (Baytas et al., 2017; Chien et al., 2021).

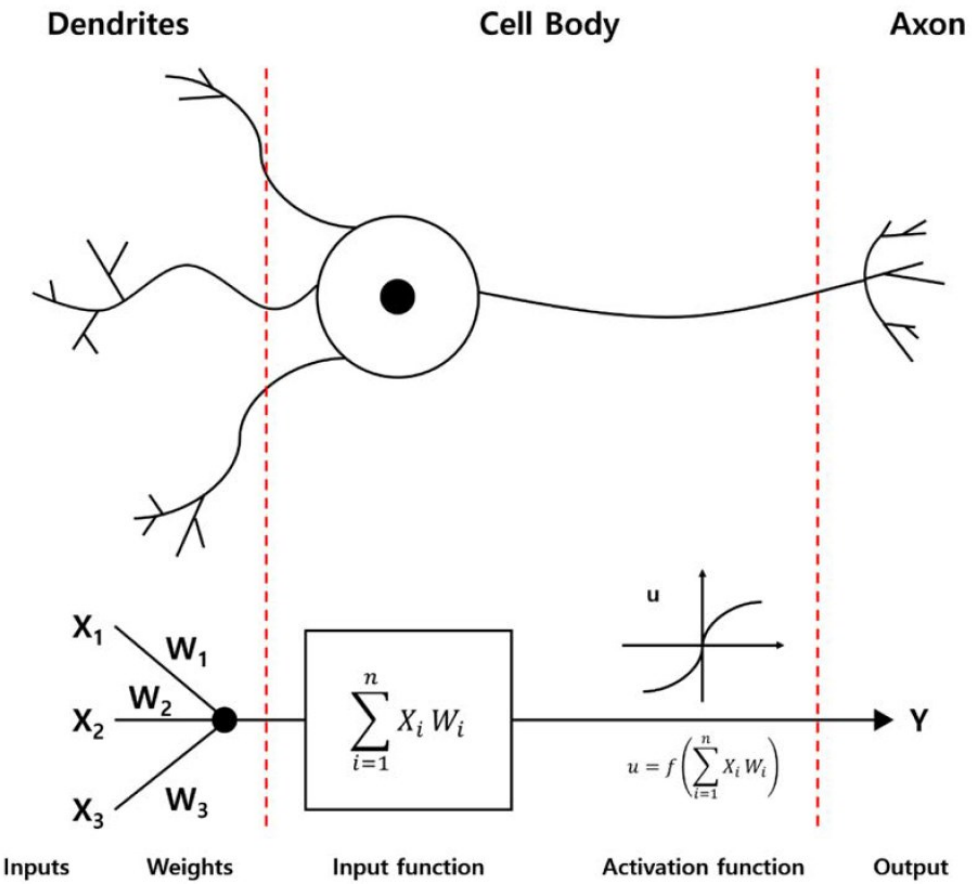


Figure 3.1. Representation of a neuron in the ANN architecture (Xiong et al., 2020).

3.2. Long Short-Term Memory (LSTM) Networks

In the healthcare domain, previous studies have used longitudinal patient health records to analyze interrelationships between elements of clinical events to derive more robust and real representations of data and enhance prediction performance (Baytas et al., 2017; Bang et al., 2017; Che et al., 2018; Bai et al., 2018; Zhang et al., 2020). Longitudinal data often has long timescale information, and LSTM networks allow capturing long-term dependencies of time series data. A standard LSTM cell consists of forget gates, input gates, output gates, and a memory cell. The mathematical formulation of LSTM is given below (Hochreiter and Schmidhuber, 1997).

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \quad \text{(Forget gate)} \quad (2)$$

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \quad \text{(Input gate)} \quad (3)$$

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \quad \text{(Output gate)} \quad (4)$$

$$\tilde{C}_t = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c) \quad \text{(Candidate memory)} \quad (5)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(Current memory)} \quad (6)$$

$$h_t = o_t \odot \tanh(C_t) \quad \text{(Current hidden state)} \quad (7)$$

Here x_t is the input at time t, h_{t-1} and h_t are hidden states at time $t-1$ and t, C_{t-1} and C_t are the cell memories at time $t-1$ and t, and $\tilde{C}_t$ is the candidate cell memory at time t. The network parameters of the forget, input, output gates and the candidate memory are represented by $[W_f, U_f, b_f]$, $[W_i, U_i, b_i]$, $[W_o, U_o, b_o]$, and $[W_c, U_c, b_c]$, respectively. Here W is weights for input value, U is weights for previous hidden states, and b is bias. Note that $\sigma(\cdot)$ and $tanh(\cdot)$ represent

the logistic sigmoid and hyperbolic tangent activation functions which are implemented elementwise in this formulation. Furthermore, $\odot$ denotes the pairwise multiplication of two vectors. The forget gate determines what information will be discarded from the cell state. When the cell state updates, the input gate generates a new memory vector with weights and indicates what new information can be preserved in the cell state for learning long-term dependencies. The cell memory updates long-term memory of the networks, and the output gate determines a new hidden state, which can be either an output of the model or input vector in the connected LSTM layer.

3.3. Time-Aware LSTM Networks (T-LSTM)

Although patient data has heterogeneous sequential records, the LSTM method assumes input data has regular time intervals between consecutive elements and this limitation can degrade LSTM performance. For this purpose, a new class of LSTM models, called Time-Aware LSTM (T-LSTM) (Baytas et al., 2017), has been proposed to overcome the issue by introducing an adjusted memory term in the network architecture to consider the time lapses between successive records. The detailed formulation of T-LSTM (Baytas et al., 2017) is given below.

$$C_{t-1}^{S} = tanh(W_d \cdot C_{t-1} + b_d) \quad \text{(Short−term memory)} \quad (8)$$

$$\hat{C}_{t-1}^{S} = C_{t-1}^{S} \cdot g(\Delta t) \quad \text{(Discounted short − term memory)} \quad (9)$$

$$C_{t-1}^{T} = C_{t-1} + C_{t-1}^{S} \quad \text{(Long−term memory)} \quad (10)$$

$$C_{t-1}^{*} = C_{t-1}^{T} + \hat{C}_{t-1}^{S} \quad \text{(Adjusted previous memory)} \quad (11)$$

Here C_{t-1} is previous memory cell, W_d and b_d are weight and bias of decomposition network, respectively. The elapsed time between x_{t-1} and x_t is represented by Δt. $g(\cdot)$ denotes a heuristic decaying function where larger values of Δt will lessen the effect of the short-term memory (Baytas et al., 2017). In the T-LSTM architecture, short-term memory is computed first and then the value is adjusted by multiplying with a non-increasing function of elapsed time. Then, integrating modified long-term and short-term memories produces an adjusted previous memory to update information with time irregularity.

3.4. Time-Aware LSTM with Power-Law Decay (T-pLSTM) Networks

In the healthcare domain, some patient data have long-timescale information. Thus, to handle this, it is necessary to use a power law function with slow information decay. However, the existing LSTM architecture has a forget gate with exponential decay. Thus, we need to modify the forget gate function as shown in the following equation so that the method can capture long timescale information via a slower forgetting mechanism. The advantage of the power law function relies on a recurrent coefficient (power law coefficient). It is suggested that a smaller value of the coefficient is better to retain long-range dependencies (Chien et al., 2021). The power law coefficient is,

$$f_t^* = \left(\frac{t-t_0+1}{t-t_0}\right)^{-p} \quad \text{(Power law forget gate)} \quad (12)$$

Where t_0 is the reference time point to represent the start of information decay, t is the elapsed time from t_0, f_t^* is the power law forget gate at time t, and p is a positive power. Using the recurrent coefficient, we develop a new T-LSTM model, called T-LSTM with power law forget

gate (T-pLSTM), which can deal with irregular and long timescale dependencies between the consecutive elements of sequential data. The architecture of the proposed T-pLSTM is shown in Figure 3.1. The data that we used to train the proposed network includes clinical records of k patients where the k-th patient has m clinical records. Thus, the clinical records of the k-th patient can be represented as $R^{(k)} = \left[R_1^{(k)}, R_2^{(k)}, R_3^{(k)}, \cdots R_{m-1}^{(k)}, R_m^{(k)}\right]$. Creation time of the patient cohort is in accordance with the order of the clinical data, denoted as $t(k) = t^{(k)} = \left[t_1^{(k)}, t_2^{(k)}, t_3^{(k)}, \cdots t_{m-1}^{(k)}, t_m^{(k)}\right]$. Each clinical record also consists of n input variables, such as patient health status and treatment information. The corresponding records can be denoted in terms of $R_m^{(k)} = \left[R_{1^{(m)}}^{(k)}, R_{2^{(m)}}^{(k)}, R_{3^{(m)}}^{(k)}, \cdots R_{n-1^{(m)}}^{(k)}, R_{n^{(m)}}^{(k)}\right]$. As observed in Figure 3.2, the proposed model uses two time-relevant terms which are elapsed time between two consecutive elements, $\Delta t_e = t_m - t_{m-1}$, and cumulative elapsed time after first diagnosis, $\Delta t_c = t_m - t_0$, respectively. Δt_e adjusts memory cell to capture time irregularity of the clinical data and Δt_c is used to compute power law forget gate to handle long timescale information on patient data. Including these two terms modifies the current memory term as $C_t = f_t^* \odot C_{t-1}^* + i_t \odot \tilde{C}_t$. Additionally, we have examined and applied the bidirectional approach in the advanced LSTM architecture to check if the bidirectional training would improve prediction performance. Adding directionality splits a LSTM cell into forward cell $\vec{h}$ and backward cells $\overleftarrow{h}$. Constructing time of patient cohort corresponds to forward direction $\vec{t} = [t_1, t_2, t_3, \cdots, t_{n-1}, t_n]$ and backward direction $\overleftarrow{t} = [t_n, t_{n-1}, t_{n-2}, \cdots, t_2, t_1]$. The directionality approach enables neural networks to preserve information in two directions from past periods to future periods and vice versa.

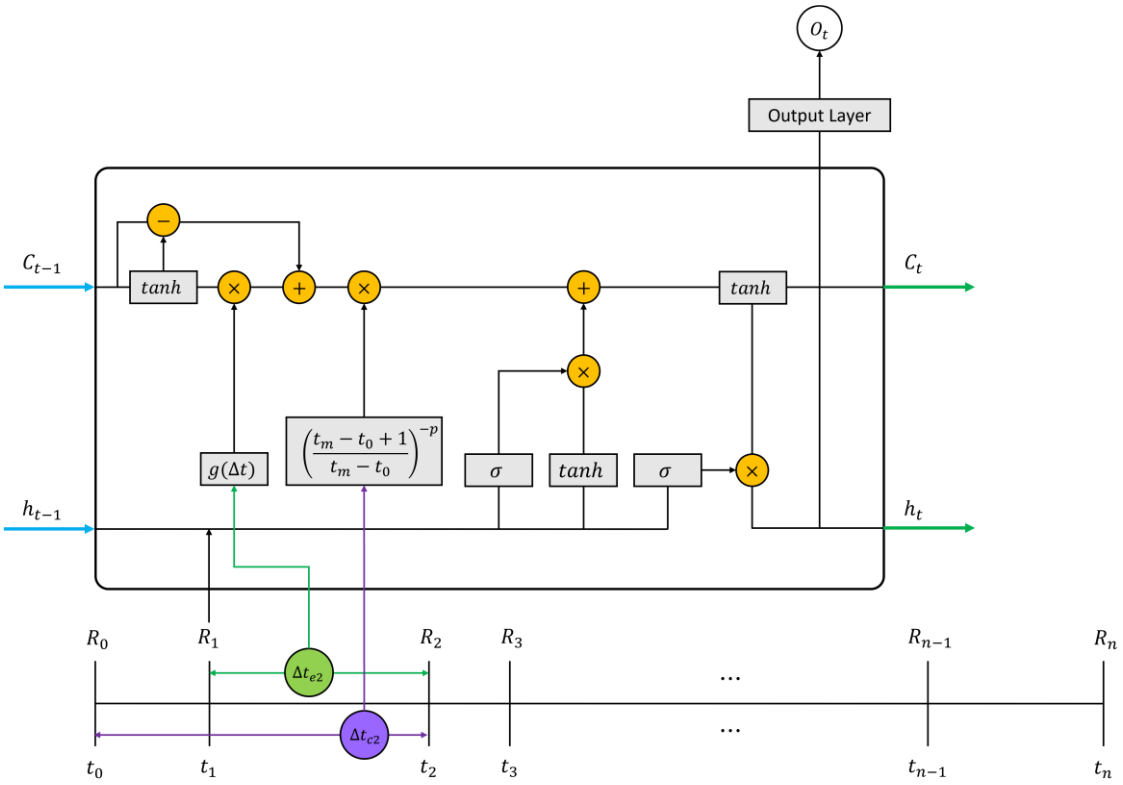


Figure 3.2. Illustration of the proposed T-pLSTM unit.

3.5. Experiment Setup

I designed an experiment to demonstrate the novelty of T-pLSTM by comparing its prediction performance with existing LSTM methods. First, I extract longitudinal data for NSCLC patients whose primary cancer is lung cancer from SEER Research Plus data. Then six types of LSTM models, a) vanilla and bidirectional LSTM networks (LSTM and Bi-LSTM); b) vanilla and bidirectional T-LSTM networks (T-LSTM and BiT-LSTM); and c) vanilla and bidirectional T-pLSTM networks (T-pLSTM and BiT-pLSTM), and an additional two supervised learning algorithms (random forest and ridge regression) for prediction of patient outcomes (tumor size and survival month) are trained by using the generated datasets.

In this stage, the hyperparameters for each model are optimized by a random search algorithm. Random search has been commonly used in hyperparameter searches of deep learning networks. Although the method does not take every hyperparameter combination in training machine learning models for consideration, it provides the combination to generate a relatively good performing model within a significantly short time (Bergstra and Bengio, 2012). The algorithm first explores a search space and then randomly picks sampling points within the space. Finally, it computes an optimized network architecture by testing out the neural network with a different architecture regardless of the results from previous iterations (Bergstra, Bengio, 2012). Table 3.1 to Table 3.4 list the optimized hyperparameters for the six LSTM models for tumor size and survival month predictions, determined by random search. I select two hyperparameters, the number of hidden neurons and the learning rate, for traditional LSTM models as well as an additional hyperparameter, the power law coefficient, for T-pLSTM models. Chien et al. (2021) suggested that a power law coefficient smaller than one appropriately handles long-timescale information. In evaluating the prediction performance of the T-pLSTM methods, I split 75% of the data for training and 25% of the data for validation and then conduct a comparative analysis.

In the model evaluation, I used three performance measures, root mean squared error (RMSE), mean absolute error (MAE), and Wilcoxon rank sum test. RMSE computes the average magnitude of the error between target variables and predictions based on the quadratic scoring rule as given in the equation below. The squared error is useful to detect undesirable errors by assigning a higher weight to larger errors. MAE estimates the errors in a set of forecasts, regardless of their direction. For both RMSE and MAE, y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the sample size, respectively. The Wilcoxon rank-sum test is an alternative to the two-sample t-test. In the equation, $U_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i > Y_i)$, where $I(X_i > Y_i) = 0$ (if $X_i \leq Y_i$) or 1 (if $X_i > Y_i$), for sample

X $(X_1, X_2, X_3, \cdots, X_m)$ and Y $(Y_1, Y_2, Y_3, \cdots, Y_n)$. The nonparametric alternative provides prediction measures by comparing two independent samples with non-normal distribution (Wilcoxon, 1945). In the next chapter, I will describe how I extract the longitudinal patient records from the SEER Research Plus database and provide initial data analysis for the dataset.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

$$P(X > Y) \approx \frac{U_1}{mn} \tag{13}$$

Table 3.1. Neural network architectures of each LSTM method for tumor size prediction using fixed length patient records (3 records, from time 1 to time 3).

Methods	Hyperparameters		
	Number of hidden neurons *(h)*	Learning rate (η)	Power law coefficient (p)
LSTM	64	0.001	-
BiLSTM	128	0.001	-
T-LSTM	64	0.005	-
BiT-LSTM	32	0.005	-
T-pLSTM	128	0.001	0.3
BiT-pLSTM	256	0.0025	0.9

Table 3.2. Neural network architectures of each LSTM method for tumor size prediction using variable length patient records (3 to 5 records).

Methods	Hyperparameters		
	Number of hidden neurons *(h)*	Learning rate (η)	Power law coefficient (p)
LSTM	32	0.001	-
BiLSTM	64	0.0025	-
T-LSTM	128	0.005	-
BiT-LSTM	256	0.001	-
T-pLSTM	64	0.0075	0.5
BiT-pLSTM	128	0.0025	0.5

Table 3.3. Neural network architectures of each LSTM method for survival month prediction using fixed length patient records (3 records, from time 1 to time 3).

Methods	Hyperparameters		
	Number of hidden neurons *(h)*	Learning rate (η)	Power law coefficient (p)
LSTM	128	0.0025	-
BiLSTM	128	0.0025	-
T-LSTM	32	0.0075	-
BiT-LSTM	256	0.001	-
T-pLSTM	64	0.0075	0.7
BiT-pLSTM	16	0.0075	0.5

Table 3.4. Neural network architectures of each LSTM method for survival month prediction using variable length patient records (3 to 5 records).

Methods	Hyperparameters		
	Number of hidden neurons *(h)*	Learning rate (η)	Power law coefficient (p)
LSTM	256	0.0025	-
BiLSTM	32	0.01	-
T-LSTM	32	0.0075	-
BiT-LSTM	128	0.0075	-
T-pLSTM	128	0.001	0.3
BiT-pLSTM	256	0.0025	0.9

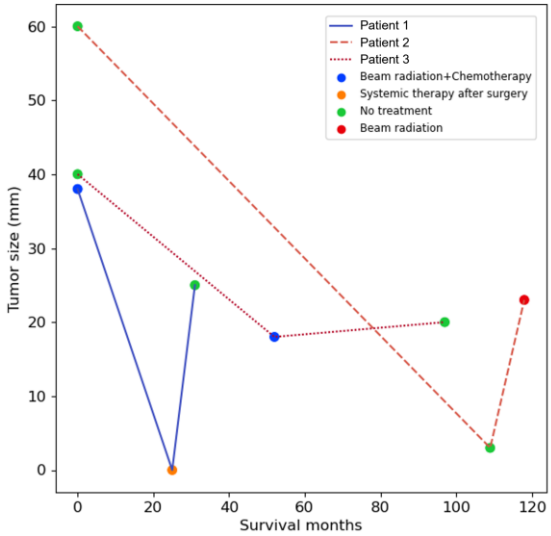
CHAPTER 4. GENERATION OF LONGITUDINAL PATIENT RECORDS FOR NSCLC

In this chapter, I will explain how longitudinal patient health records are extracted from the US population-based data, SEER Research Plus database. The subchapters describe data initialization of the longitudinal patient records with selected variables and what information I could recognize from the generated training dataset, respectively.

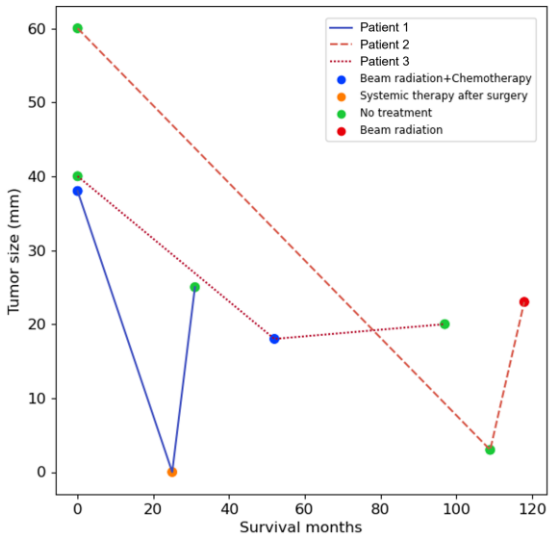
4.1. Data Initialization of Longitudinal Patient Health Records

I used the SEER Research Plus Data (National Cancer Institute, 2021) to extract longitudinal data for NSCLC patient records. The dataset includes variables related to the variation of tumor size, such as age, sex, marital status, site recode, histology ICD-O-2, treatment modality (surgery, chemotherapy, radiation, and systematic treatment), stage group, and survival months. In the input variables, site recode represents clinically relevant and histologically defined rare cancers grouped by cancer types. Histology ICD-O-2 means histologic type of tumor defined by International Classification of Diseases for Oncology (ICD-O). Stage group classifies amount and spread of cancer based on the staging system from the American Joint Committee on Cancer (AJCC). I extracted the patient cohort in my analysis using a few criteria. First, I identified and extracted the patients who had NSCLC as their primary diagnosis and multiple treatment records. The filtered data has patient records ranging from 1 to 9 and most cases are patients with 3 to 5 records. Thus, I only included the patients with 3 to 5 records. Furthermore, the tumor size information was limited to the period 2004-2015 in this database. In addition, there were 82% missing values in treatment modality variables and some patients lacked sequential treatment data and demographic information. These patient records were excluded from my analysis. Finally, I observed that there

were 879 complete patient records with treatment information for lung cancer. Specifically, there are 795, 84, and 30 patients with 3, 4, and 5 records respectively. Among these patients, 849 have the NSCLC. Of these patients, there are 762, 78, 9 patients with 3, 4, and 5 records respectively. Thus, I extracted longitudinal patients' observations from SEER Research Plus data with 3 to 5 timesteps to investigate the effect of sequence length on model performance. Furthermore, NSCLC patient records have missing values in output responses (tumor size and survival month). I excluded records with missing values and thus the final dataset includes 497, 63, and 7 patients with 3, 4, and 5 records respectively. In data preprocessing, continuous variables were normalized while categorical variables were encoded to convert nominal into dummy variables. I used Base-N encoding to indicate categorical data efficiently by reducing the number of features compared to binary encoding (McGinnis, 2016). Binary encoding is inappropriate to represent a large number of levels included in data and thus using Base-N encoding allows us to handle high dimensionality of the categorical variables. Figure 4.1 shows the trajectories of tumor size growth over time for three NSCLC patients. I used the longitudinal patient records to train the LSTM models.



(a)

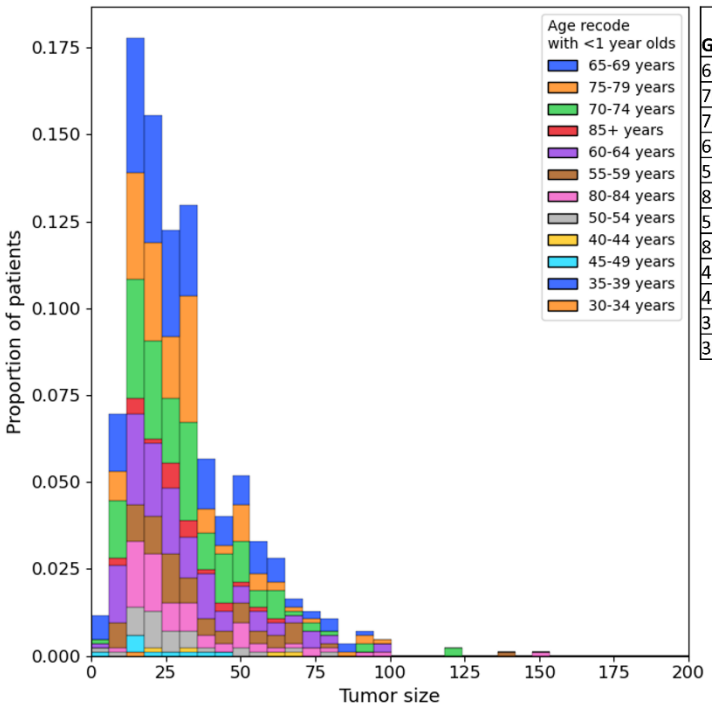


(b)

Figure 4.1. Longitudinal trajectories of NSCLC tumor size versus survival month and treatment modality for three patients. (a) Patient 1 has information from time 1 to time 3. Patient 2 and 3 have information from time 1 to time 4; (b) all three patients have three timesteps.

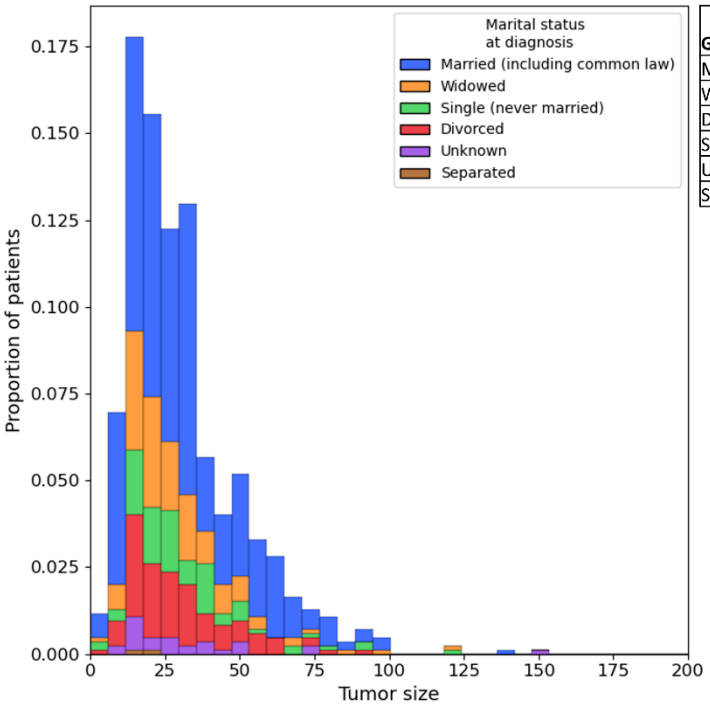
4.2. Data Exploration and Preparation

I set tumor size growth as the output response for the initial data analysis and conducted data exploration to see the relationship between the selected variables. Figure 4.2 shows tumor size distribution grouped by age, marital status, race, and sex. The plotted results show that the mid-60s to mid-70s account for a large portion of the distribution as elderly people have higher incidence for NSCLC. Marital status indicates that married people distribute across the whole range of tumor size. Their portion is much higher than others at larger tumor size because they are under care from their spouse. When it comes to race, the portion of white people is significantly higher than other races and the issue might be related to wealth and immigration status. Tumor size distribution by sex implies that males have larger tumor sizes than females. Additionally, I analyzed the pairwise relationship between tumor size and survival months to investigate survivability of NSCLC patients. As expected, NSCLC patients with higher tumor size have low survivability. In order to figure out how tumor size growth is correlated with other variables, I used random forest (RF) to rank feature importance for the selected input variables. Reif et al. (2006) verified that the RF is highly effective in identifying interrelationships between features with tiny effects in high dimensional data and thereby the algorithm fits to the integrated study of multiple types of datasets (categorical and continuous data). Figure 4.3 shows feature importance generated from the RF classifier and I could see variables representing patient health status had higher impact than treatment information. Generally, we should select features with high impact in predicting output variables, but I did not remove treatment information such as radiation and chemotherapy. In practice, treatment variation in cancer patients likely contributes to change of tumor size growth and they could play an important role in prognostic analysis.



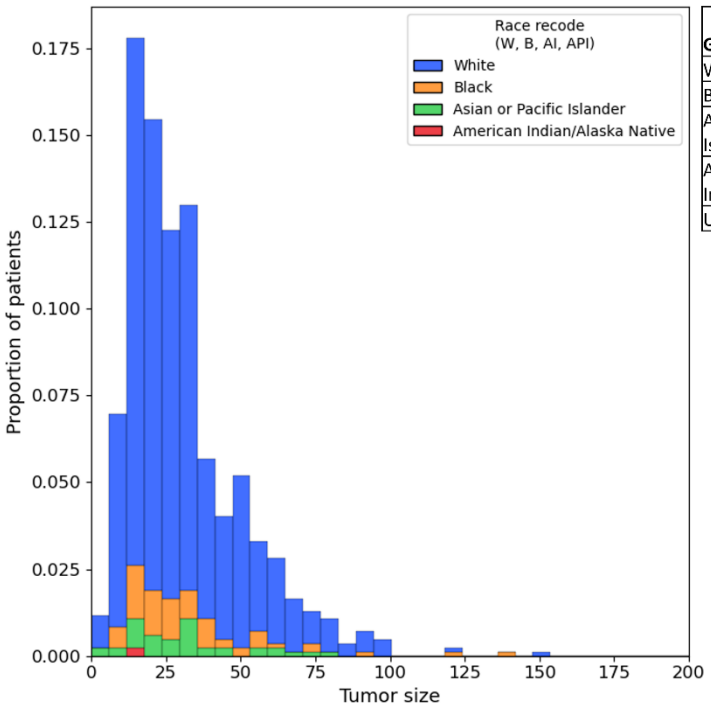
Group	Number of patients
65-69 years	195
70-74 years	162
75-79 years	137
60-64 years	126
55-59 years	76
80-84 years	76
50-54 years	34
85+ years	25
45-49 years	11
40-44 years	5
35-39 years	1
30-34 years	1

(a) age



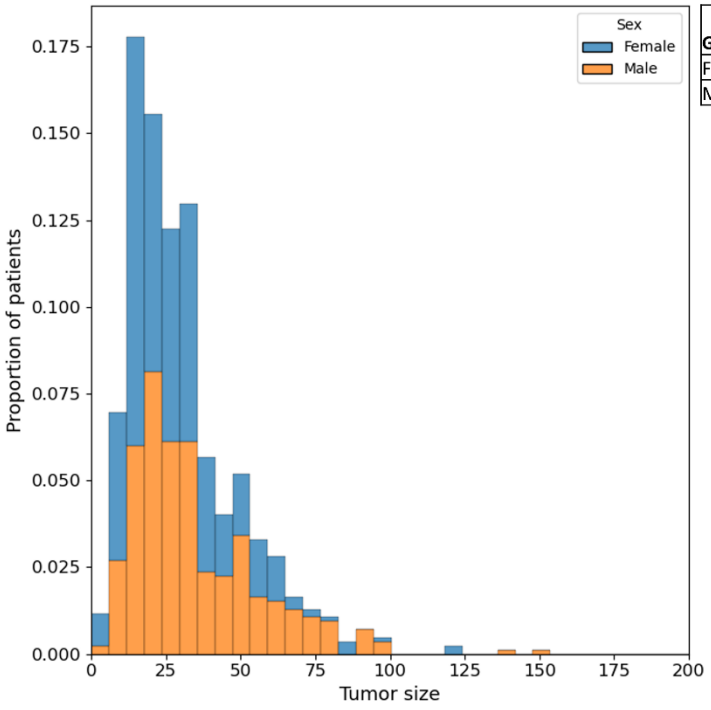
Group	Number of patient records
Married	476
Widowed	133
Divorced	121
Single	88
Unknown	29
Separated	2

(b) marital status



Group	Number of patients
White	728
Black	75
Asian or Pacific Islander	43
American Indian/Alaska Native	2
Unknown	1

(c) race



Group	Number of patients
Female	434
Male	415

(d) sex

Figure. 4.2. Tumor size distribution grouped my multiple categories.

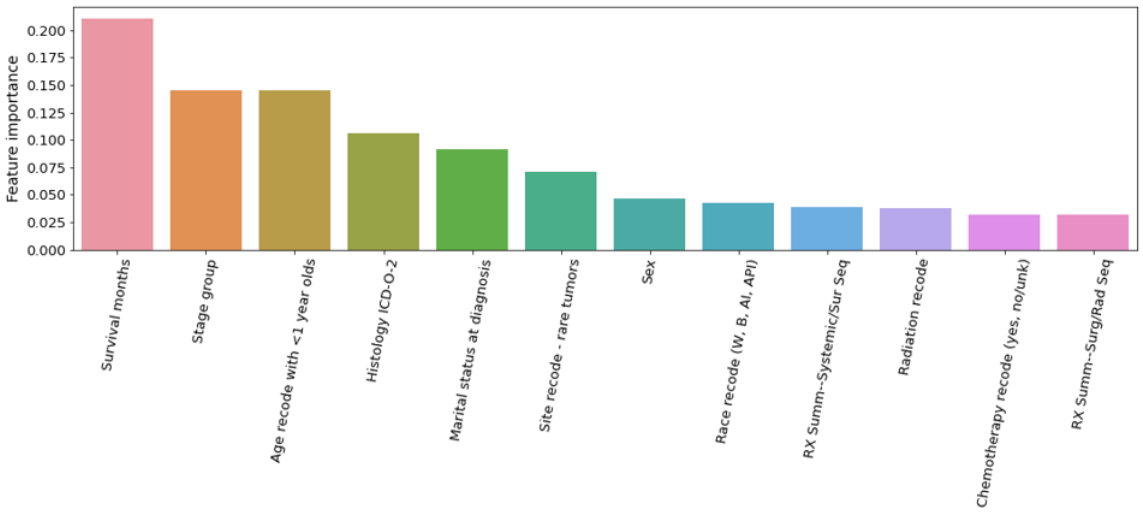


Figure 4.3. Feature importance analysis using random forest to explore interrelationship between tumor size and other variables.

Then, I conducted principal component analysis (PCA) to provide how the selected features are interrelated to tumor size only as I could see the output response is highly related to survival month. PCA is well-known as a popular feature extraction method. The algorithm creates new variables, called principal components, by projecting the original variables to the newly generated components (Wold et al., 1987). The linear transformation reduces data dimension based on their eigenvalues and thereby enables us to capture what is the most valuable information present in the original variables (Cateni et al., 2013; Awan et al., 2019). As a rule of thumb, we select first *m* components accounting for 95% of the total variance in the data. Figure 4.4 shows explained variance from PCA analysis, and we could see most of the information was attributed to three principal components. Figure 4.5 provides details of how each component is correlated to original input variables. The results indicate that survival months and histology are highly related to the first principal component, followed by stage group for second principal component and age for third principal component, respectively.

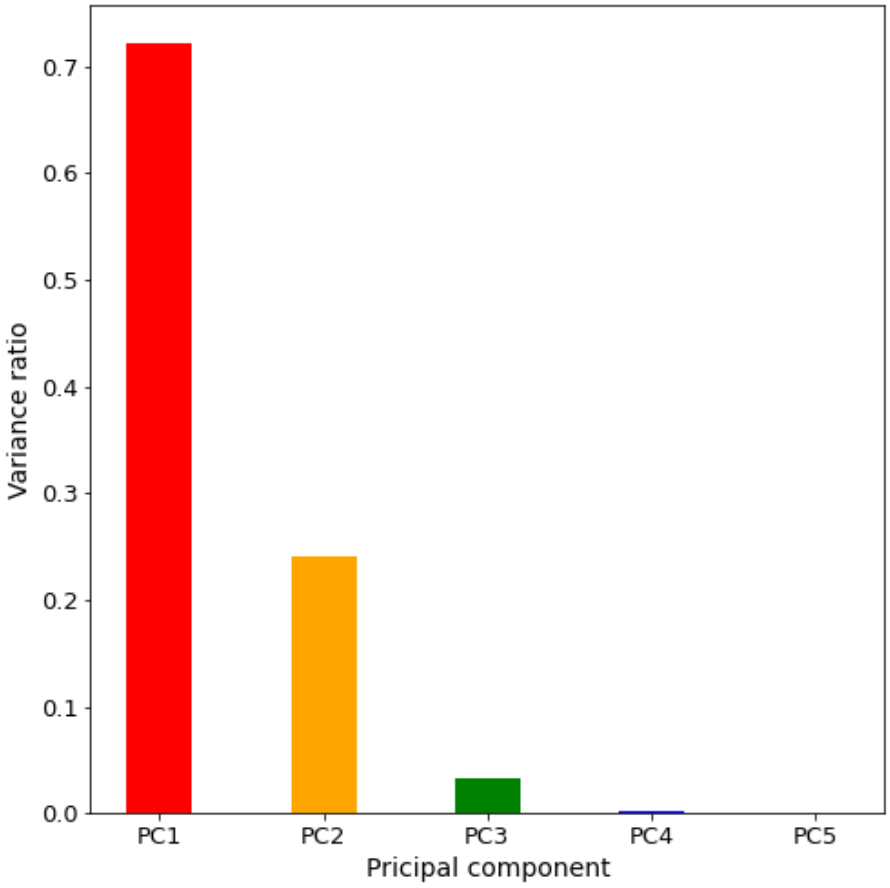
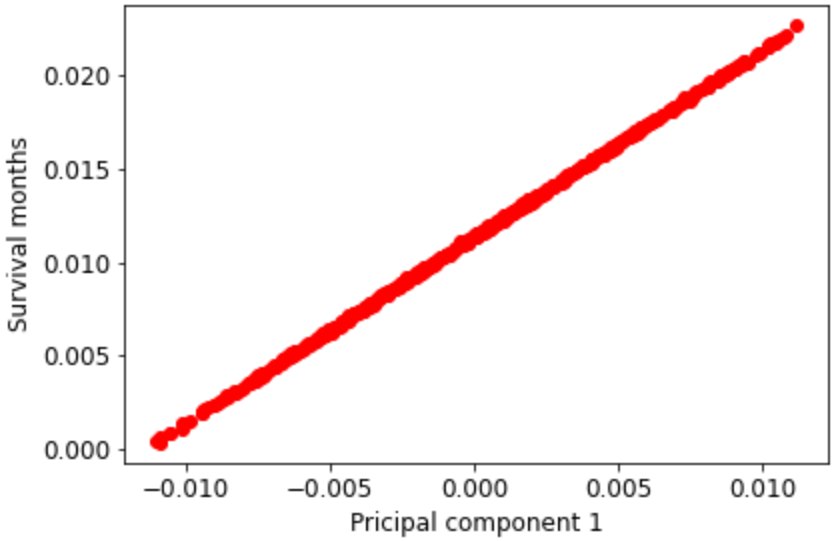
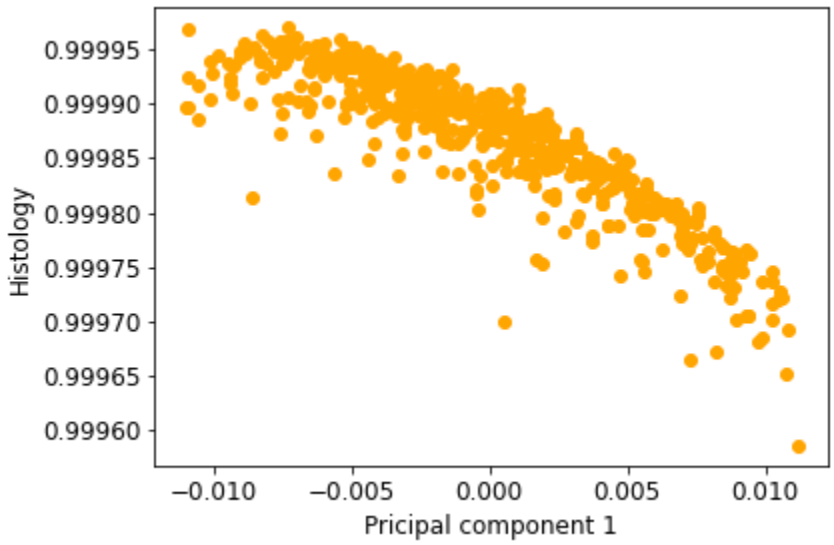


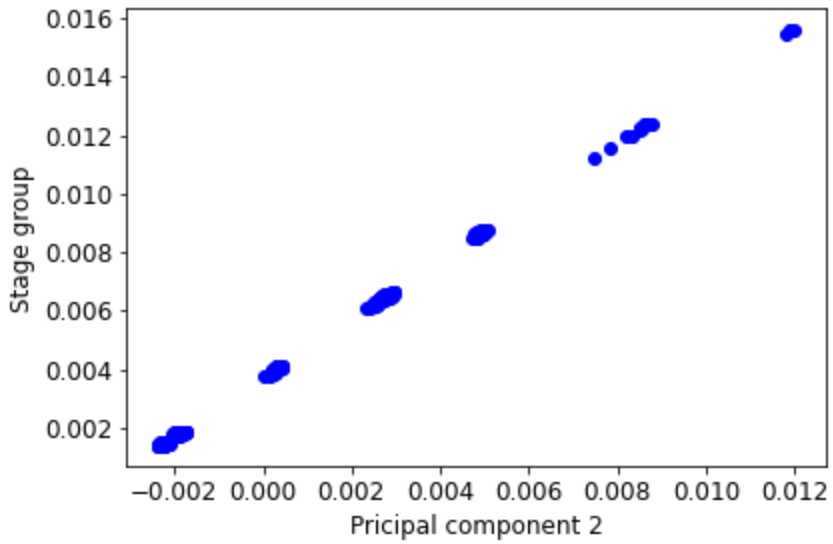
Figure 4.4. Explained variance ratio of PCA analysis for NSCLC dataset.


Survival months
0.020
0.015
0.010
0.005
0.000
−0.010
−0.005
0.000
0.005
0.010
Pricipal component 1

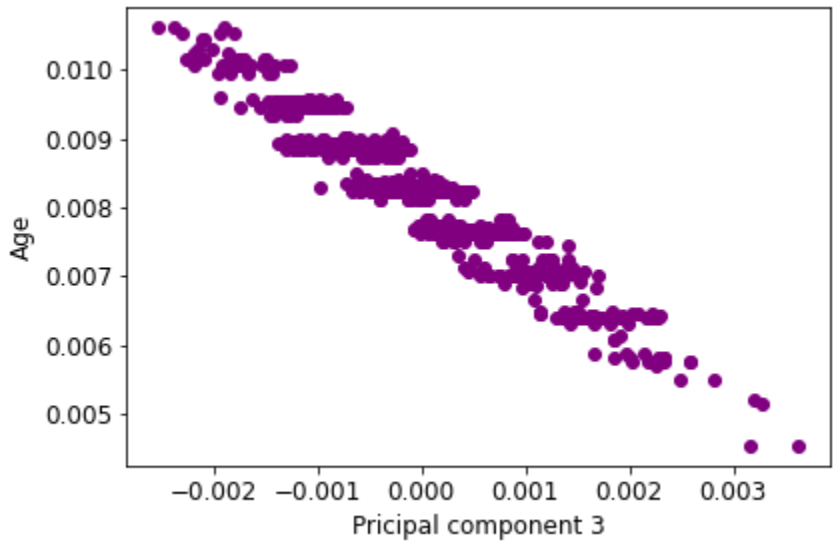
(a)


Histology
0.99995
0.99990
0.99985
0.99980
0.99975
0.99970
0.99965
0.99960
−0.010
−0.005
0.000
0.005
0.010
Pricipal component 1

(b)



(c)



(d)

Figure 4.5. Pairwise analysis between principal components and input variables: first principal component correlated with (a) survival months and (b) histology, (c) second principal component associated with stage group, and (c) third principal component attributed to age.

CHAPTER 5. TUMOR SIZE AND SURVIVAL MONTH PREDICTIONS USING T-PLSTM NETWORKS

Previous sections describe the limitations of traditional LSTM models and how I extracted longitudinal patient records to train the T-pLSTM model. This chapter shows comparisons between my advanced approach and existing algorithms so as to demonstrate the efficacy of my model in predicting patient outcomes for NSCLC. As mentioned in the section 2.4, I use RMSE, MAE, and Wilcoxon rank sum test to provide comparative analysis.

5.1. Effect of Sequence Length of Patient Records for Prediction Performance

Before investigating the efficacy of my proposed model, I compared prediction outcomes of the LSTM models with other supervised learning algorithms to verify the importance of capturing time-relevant information. I selected two additional ML algorithms, random forest (RF) and ridge regression (RR), for the experiment. Table 5.1 shows the averaged RMSE and MAE results for each algorithm optimized by random search. The results indicate that the LSTM models improved prediction performance by approximately 66% to 87% compared to the RF and RR algorithms. That is, capturing time-series information significantly outperforms the two algorithms with no consideration of time-relevant information.

Table 5.1. Average RMSE and MAE of supervised learning algorithms for tumor size prediction using fixed length patient records (3 records, from time 1 to time 3).

Methods	Hyperparameters	Training		Validation	
		RMSE	MAE	RMSE	MAE
RF	Number of trees = 100 Maximum depth = 50 Minimum number of samples to split = 8 Minimum number of samples at leaf node = 3	0.4702± 0.0184	0.1558± 0.0100	0.6534± 0.0889	0.2267± 0.0267
RR	Alpha = 0.9843	0.9279± 0.0305	0.3998± 0.0229	0.9379± 0.0903	0.4061± 0.0336
LSTM	h = 32 η = 0.0075	0.1642± 0.0132	0.0588± 0.0194	0.1544± 0.0283	0.0567± 0.0186
Bi-LSTM	h = 16 η = 0.025	0.1664± 0.0082	0.0497± 0.0089	0.1632± 0.0230	0.0488± 0.0101
T-LSTM	h = 256 η = 0.001	0.01618 ±0.0105	0.0501± 0.0069	0.1550± 0.0219	0.0482± 0.0068
BiTLSTM	h = 256 η = 0.0075	0.1662± 0.0093	0.0518± 0.0102	0.1621± 0.0205	0.0502± 0.0131

Then, I implemented and compared four LSTM methods, LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM, to investigate the effect of temporal information on their prediction performance. All models were single-layer architecture. Random search (Bergstra and Bengio, 2012) was used to optimize two hyperparameters including number of hidden neurons (h) and learning rate (η). Moreover, I have studied the effect of activation function on prediction performance. By comparing three types of popular activation functions (tanh, ReLU, Leaky ReLU), I observed that tanh has a comparable performance to Leaky ReLU. This result is in line with Baytas et al. (2017). Therefore, I used the tanh activation function to handle temporal data. I designed experiments to determine optimum number of epochs in computing better prediction

performance and the experimental results show that the best number of epochs are obtained for 750 and 1500 for LSTM/Bi-LSTM and T-LSTM/BiT-LSTM, respectively. All the LSTM methods are implemented in Python version 3.7 with Keras (Chollet et al., 2015) and TensorFlow libraries (Abadi et al., 2015). All experiments and data processing are implemented on the OU Supercomputing Center for Education & Research (OSCER) with Intel Xeon "Haswell" E5-2650v3 10-core 2.3 GHz and 32 GB of RAM.

Table 5.2 demonstrates training and validation RMSE and MAE for tumor size prediction of the considered LSTM models using fixed length patient records (3 records). I considered the patient records at time 1 and time 2 (as x_t, $t \in \{1, 2\}$), then I aimed to predict the output time 3 (as y_t, $t = 3$), shown in Figure 5.1. Figure 5.2 shows the robustness of training and validation RMSE and MAE scores for each model, which have been calculated over 50 repetitions for each LSTM model on the same data. In this process, different random states were assigned to each repetition to get consistent results from different running of the ML models. The results on the validation set show that T-LSTM outperforms BiT-LSTM in terms of MAE (3.96% vs. 4.11%) and the RMSE is similar for both methods (4.48%) although the T-LSTM method has relatively large variation (less robust) in terms of RMSE on the validation set. From predictions of all the LSTM models, I could see that although T-LSTM and BiT-LSTM have relatively larger variation of RMSE scores compared to LSTM and Bi-LSTM methods, they obtained lower RMSE and MAE and thus capturing temporal information better. The results also reveal that adding bidirectional framework improves the model's prediction performance.

Furthermore, I compared the LSTM methods using the variable sequence length patient records (i.e., our data consists of patients with 3, 4, and 5 patient records) as shown in Table 5.3. In these cases, Bi-LSTM shows better RMSE (4.27%) than other methods while the T-LSTM

model produces lower MAE (3.75%) on the validation set compared to other methods. Figure 5.3 shows that the RMSE scores of all LSTM models using different length patient records have relatively larger variation than the models built on the fixed length records (Figure 5.2).

		Features/predictors							Response variable
Patient record/time	Used for	Age	Marital Status	...	Survival month	Radiation treatment	ic tre	...	Tumor size
Patient 1, T1	Train	82	Married		93	0	1		78
Patient 1, T2	Train	82	Married		51	1	0		65
Patient 1, T3	Validation	83	Divorced		7	1	0		20
...	...								
Patient n+1, T1	Train								
Patient n+1, T3	Train								
Patient n+1, T4	Validation								
...	...								
Patient n+m, T1	Train								
Patient n+m, T4	Train								
Patient n+m, T5	Validation								

Figure 5.1. An example of a dataset for patient records.

Table 5.2. Average RMSE and MAE of four LSTM models for tumor size prediction using fixed length patient records (3 records, from time 1 to time 3). The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
LSTM	0.1615±0.0148	0.1491±0.0148	0.0487±0.0370	0.0463±0.0175
Bi-LSTM	0.1605±0.0120	0.1552±0.0066	0.0468±0.0259	0.0411±0.0068
T-LSTM	0.1584±0.0147	**0.1404±0.0045**	**0.0448±0.0351**	**0.0396±0.0097**
BiT-LSTM	**0.1565±0.0113**	0.1468±0.0031	**0.0448±0.0283**	0.0411±0.0068

Table 5.3. Average RMSE and MAE of four LSTM models for tumor size prediction using

variable length patient records (3 to 5 records). The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
LSTM	0.1538±0.0293	0.1307±0.0101	0.0443±0.0466	0.0391±0.0132
Bi-LSTM	**0.1527±0.0298**	0.1357±0.0073	**0.0427±0.0472**	0.0388±0.0094
T-LSTM	0.1540±0.0246	**0.1253±0.0054**	0.0452±0.0431	**0.0375±0.0092**
BiT-LSTM	0.1543±0.0203	0.1323±0.0058	0.0448±0.0424	0.0388±0.0094

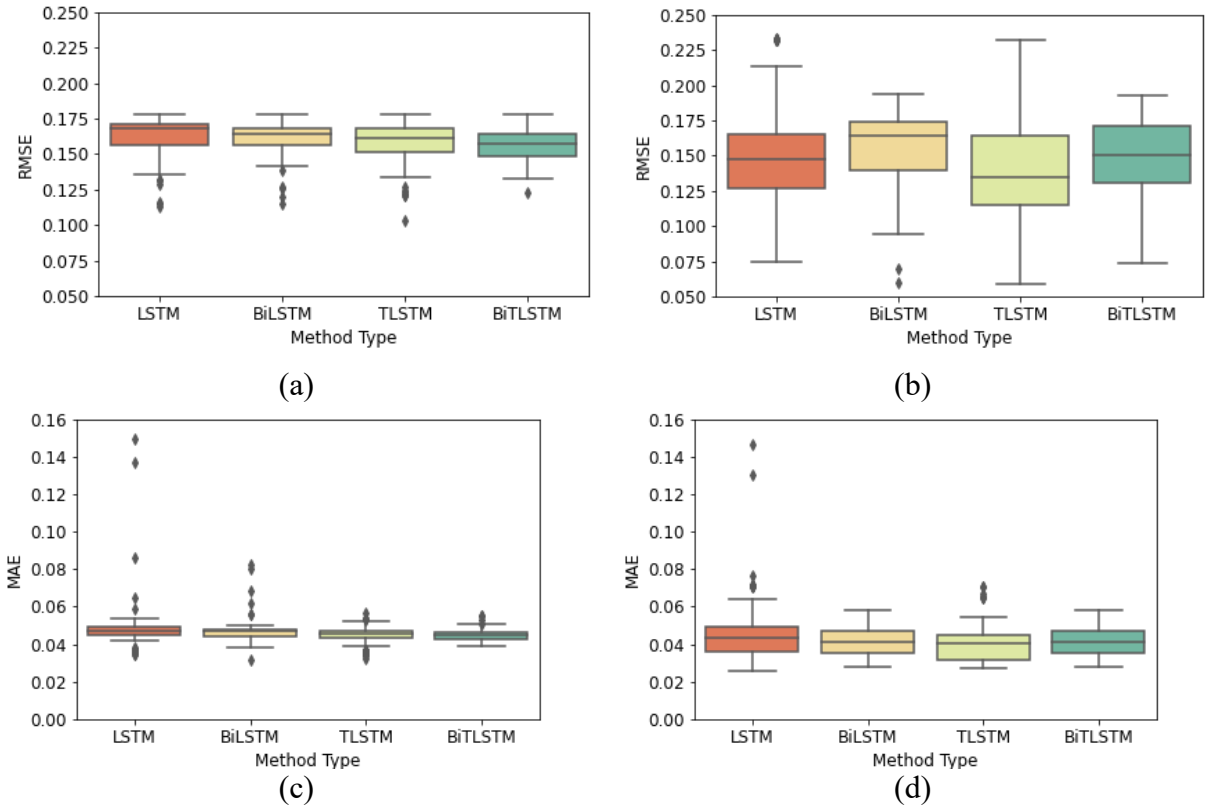


Figure 5.2. Train and validation results of LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for tumor size prediction using fixed sequence length patient records (3 records, from time 1 to time 3): (a) Training RMSE, (b) Validation RMSE, (c) Training MAE, and (d) Validation MAE.

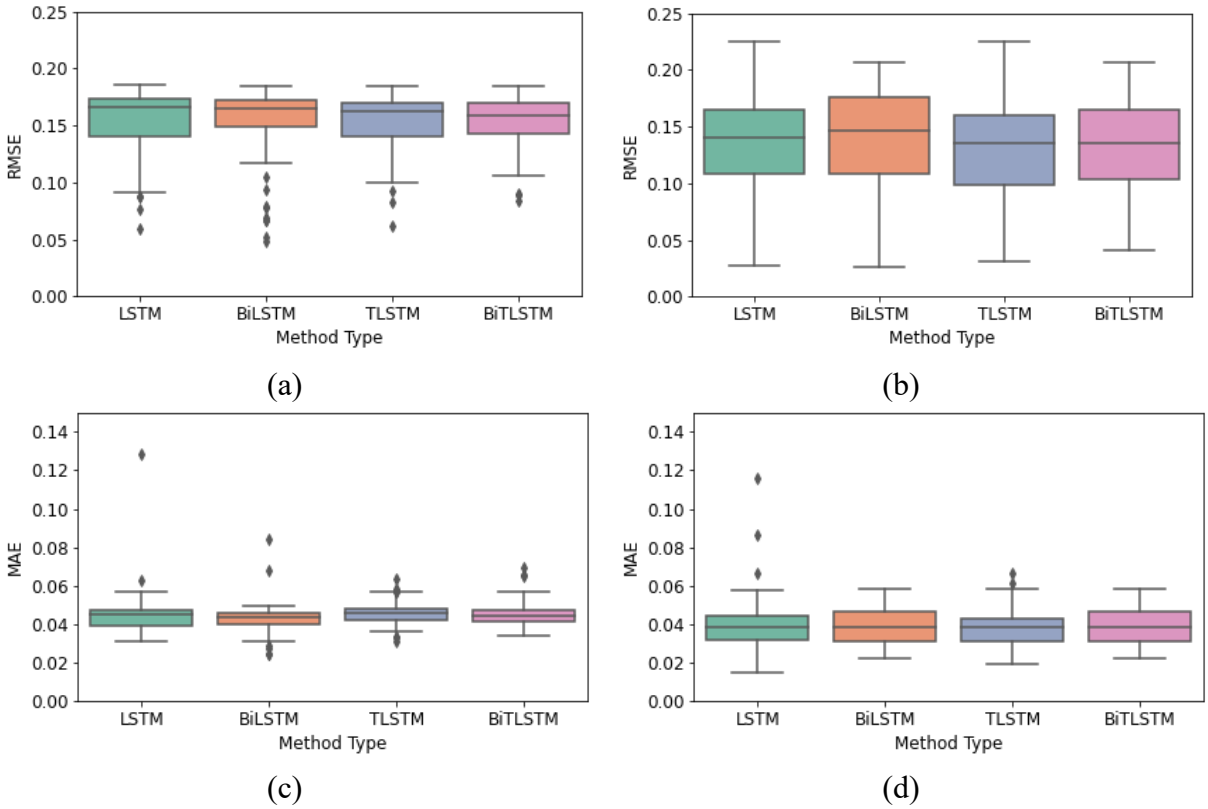


Figure 5.3. Training and validation results of LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for tumor size prediction using different sequence length patient records: (a) Training RMSE, (b) Validation RMSE, (c) Training MAE, and (d) Validation MAE.

Next, I studied the behavior of the performance metrics of the LSTM models for survival month prediction. I considered the output variable as survival month instead of tumor size. I report the average RMSE and MAE values in Table 5.4 and Figure 5.5. I observed that all LSTM models performed worse than models where the output response was tumor size. Regardless of the patient sequence records, I observed that the Bi-LSTM showed lower RMSE compared to both T-LSTM and BiT-LSTM models while LSTM showed superior performance in terms of MAE.

Huang et al. (2020) explained that the lack of clinical data might lead to degrading the effect of sequential information in clinical datasets and EHRs. As mentioned in the cohort section, I could obtain clinical records of 840 patients with complete cases, which was a small size dataset to incorporate a comprehensive set of NSCLC patients. I expect that the advanced algorithms, such as LSTM, might not capture time-relevant information from data with a small number of patients and thus it might lead to poor prediction performance. I also observed that the results of survival month prediction using different sequence length patient records were similar to the results using fixed length patient data, as shown in Table 5.4 and Figure 5.5. The comparative analysis shows that the models with patient records with different sequence lengths are less robust (larger variation of prediction results) than models using fixed length patient dataset since the patient data with more than 4 records has fewer instances than the one with 3 records. Thus, I observed that the time-aware models could successfully capture sequential information from patient data, but in the future, I need to collect more datasets from various sources in predicting outcomes to help to make clinical decisions for NSCLC patients.

Table 5.4. Average RMSE and MAE of four LSTM models for survival month prediction using fixed length patient records (3 records, from time 1 to time 3). The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
LSTM	0.1781±0.0409	0.1636±0.0421	0.1332±0.0561	**0.1219±0.0503**
Bi-LSTM	**0.1736±0.0408**	**0.1562±0.0414**	**0.1292±0.0560**	0.1362±0.0470
T-LSTM	0.1904±0.0353	0.1746±0.0370	0.1451±0.0518	0.1333±0.0471
BiT-LSTM	0.1942±0.0348	0.1788±0.0369	0.1489±0.0512	0.1362±0.0470

Table 5.5. Average RMSE and MAE of four LSTM models for survival month prediction using variable length patient records (3 to 5 records). The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
LSTM	0.1495±0.0430	0.1398±0.0459	0.0983±0.0507	**0.0926±0.0485**
Bi-LSTM	**0.1440±0.0445**	**0.1315±0.0458**	**0.0964±0.0573**	0.0998±0.0479
T-LSTM	0.1603±0.0377	0.1464±0.0462	0.1078±0.0470	0.0988±0.0488
BiT-LSTM	0.1608±0.0393	0.1454±0.0456	0.1095±0.0499	0.0998±0.0479

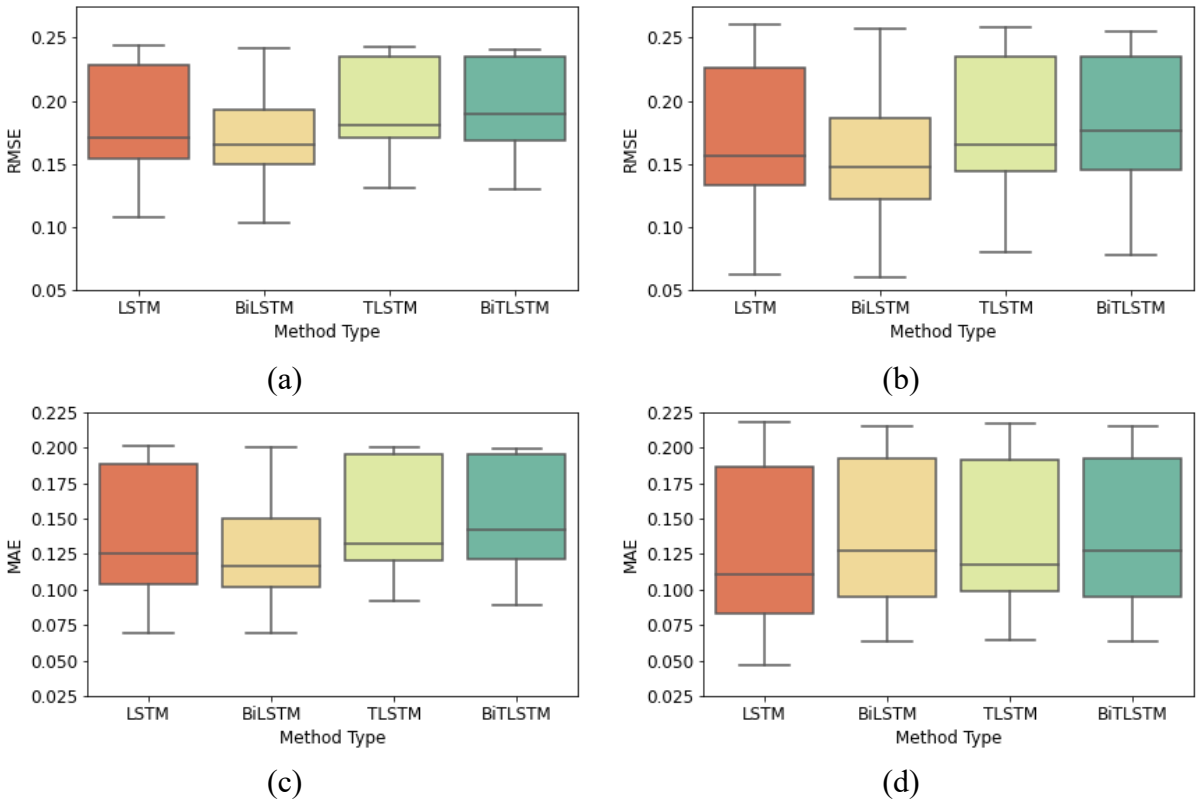


Figure 5.4. Training and validation results of LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for survival month prediction using fixed sequence length patient records: (a) Training RMSE, (b) Validation RMSE, (c) Training MAE, and (d) Validation MAE.

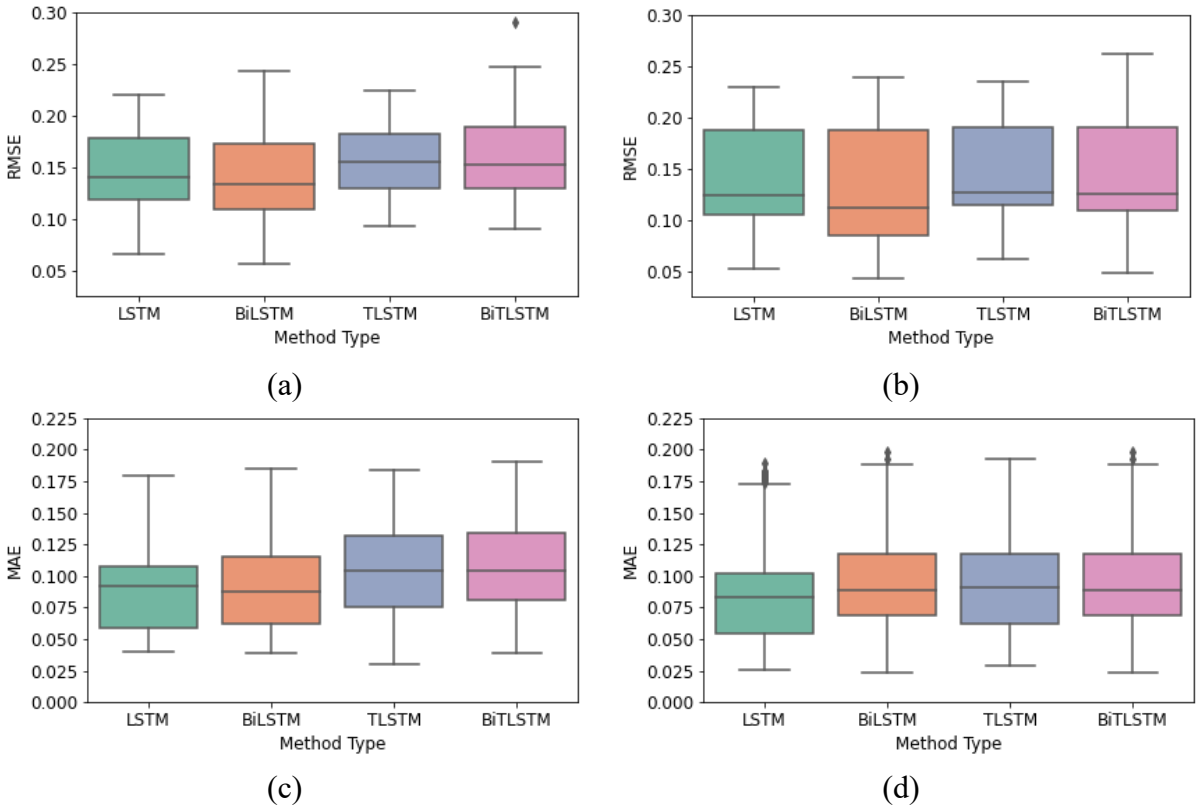


Figure 5.5. Training and validation results of LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for survival month prediction using different sequence length patient records: (a) Training RMSE, (b) Validation RMSE, (c) Training MAE, and (d) Validation MAE.

In addition, I used the pairwise Wilcoxon rank-sum test with the 5% level of significance ($\alpha = 0.05$) to compare the prediction performance between T-LSTM and BiT-LSTM versus LSTM and Bi-LSTM models. The non-parametric test is useful for comparison of prediction performance of ML models trained by datasets with non-normal distribution, but we should be careful with the power of the test method in multiple comparison problems. For instance, one comparison in my experiments has a chance with a 5% probability level to compute incorrect estimation. Assuming 100 comparisons, I use 100 confidence intervals simultaneously, each comparison with a 95%

confidence level, the expected number of false evaluations is 5 but it cannot be as my experiment was designed as a single comparison. This problem becomes worse when each comparison has statistically independent of the intervals as the probability of false estimation is much significant than the previous case (Kutner et al., 2005; Koulouris, 2020). Thus, we should be aware of that the test results cannot be as confident as the given confidence level in a statistical hypothesis. Table 5.6 and Table 5.7 show a comparison of pairwise Wilcoxon rank-sum tests between these models for tumor size prediction. The results using fixed length patient records implies that T-LSTM and BiT-LSTM outperforms traditional LSTM models while I could not see significant differences among the LSTM models using different length patient records. Survival month prediction given in Table 5.8 and Table 5.9 also indicates similarity with the results from tumor size prediction. These results may be attributed to the disadvantage of Wilcoxon rank sum test, but the RMSE and MAE results (Figure 5.4 and 5.5) show the effect of datasets is more significant.

Table 5.6. Pairwise Wilcoxon rank sum test for T-LSTM and BiT-LSTM versus LSTM and Bi-LSTM models for tumor size prediction using fixed length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-LSTM > LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
T-LSTM > Bi-LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
T-LSTM > BiT-LSTM	0.0766	Fail to reject H_0	0.1467	Fail to reject H_0
BiT-LSTM > LSTM	0.8285	Fail to reject H_0	**$\ll 0.05$**	**Reject H_0**
BiT-LSTM > Bi-LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**

Table 5.7. Pairwise Wilcoxon rank sum test for T-LSTM and BiT-LSTM versus LSTM and Bi-LSTM models for tumor size prediction using different length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-LSTM > LSTM	0.2423	Fail to reject H_0	0.1658	Fail to reject H_0
T-LSTM > Bi-LSTM	0.0524	Fail to reject H_0	0.4808	Fail to reject H_0
T-LSTM > BiT-LSTM	0.3496	Fail to reject H_0	0.3055	Fail to reject H_0
BiT-LSTM > LSTM	0.1658	Fail to reject H_0	0.8554	Fail to reject H_0
BiT-LSTM > Bi-LSTM	0.4808	Fail to reject H_0	0.6900	Fail to reject H_0

Table 5.8. Pairwise Wilcoxon rank sum test for T-LSTM and BiT-LSTM versus LSTM and Bi-LSTM models for survival month prediction using fixed length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-LSTM > LSTM	0.0967	Fail to reject H_0	0.0653	Fail to reject H_0
T-LSTM > Bi-LSTM	**≪ 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**
T-LSTM > BiT-LSTM	0.6449	Fail to reject H_0	0.6799	Fail to reject H_0
BiT-LSTM > LSTM	**≪ 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**
BiT-LSTM > Bi-LSTM	**≪ 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**

Table 5.9. Pairwise Wilcoxon rank sum test for T-LSTM and BiT-LSTM versus LSTM and Bi-LSTM models for survival month prediction using different length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-LSTM > LSTM	0.2579	Fail to reject H_0	0.2382	Fail to reject H_0
T-LSTM > Bi-LSTM	0.0600	Fail to reject H_0	0.2465	Fail to reject H_0
T-LSTM > BiT-LSTM	0.5003	Fail to reject H_0	0.5940	Fail to reject H_0
BiT-LSTM > LSTM	0.3731	Fail to reject H_0	0.1536	Fail to reject H_0
BiT-LSTM > Bi-LSTM	0.3719	Fail to reject H_0	**≪ 0.05**	**Reject H_0**

5.2. Effect of Slower Forget Gate for Prediction Performance

After reviewing that time-aware networks outperform traditional LSTM networks, I also compared them with T-pLSTM models. The experimental results to determine the number of epochs show that the best number of epochs for early stopping are obtained as 1500 for both T-pLSTM and BiT-pLSTM. Table 5.10 and Figure 5.6 show the results of T-pLSTM and BiT-pLSTM using fixed length patient records. According to this table and figure, both T-pLSTM and BiT-pLSTM models successfully handled the long-term temporal information by achieving better predictions than T-LSTM models. The best results are obtained as 4.34% RMSE of BiT-pLSTM and 4.14% MAE of T-pLSTM on the validation data. However, as shown in Table 5.11 and Figure 5.7, using different patient records' length does not indicate the effect of long-timescale information as well as the ones using fixed length patient records.

Table 5.10. Average RMSE and MAE of four time-aware LSTM networks for tumor size prediction using fixed length patient records. The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
T-LSTM	0.1561±0.0194	**0.1408±0.0076**	0.0456±0.0447	0.0430±0.0129
BiT-LSTM	0.1582±0.0166	0.1493±0.0123	0.0466±0.0369	0.0452±0.0139
T-pLSTM	0.1593±0.0187	0.1476±0.0058	0.0444±0.0335	**0.0414±0.0097**
BiT-pLSTM	**0.1546±0.0189**	0.1466±0.0045	**0.0434±0.0367**	0.0419±0.0087

Table 5.11. Average RMSE and MAE of four time-aware LSTM networks for tumor size prediction using different length patient records. The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
T-LSTM	0.1554±0.0261	0.1486±0.0037	**0.0436±0.0399**	**0.0419±0.0084**
BiT-LSTM	**0.1393±0.0356**	**0.1363±0.0135**	0.0443±0.0619	0.0458±0.0167
T-pLSTM	0.1611±0.0177	0.1549±0.0039	0.0440±0.0340	0.0427±0.0080
BiT-pLSTM	0.1472±0.0271	0.1563±0.0117	0.0454±0.0468	0.0492±0.0121

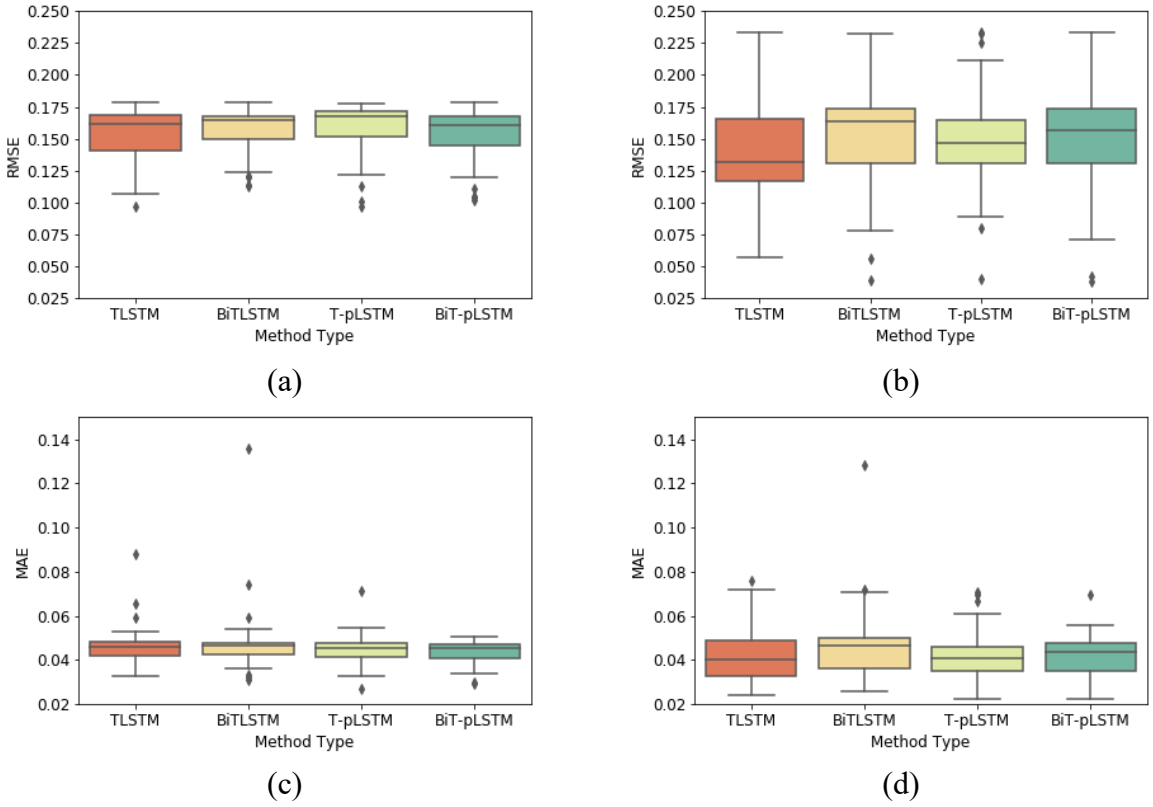


Figure 5.6. Training and test results of T-LSTM, Bi-TLSTM, T-pLSTM, and BiT-pLSTM models for tumor size prediction using fixed sequence length patient records: (a) Training RMSE, (b) Test RMSE, (c) Training MAE, and (d) Test MAE.

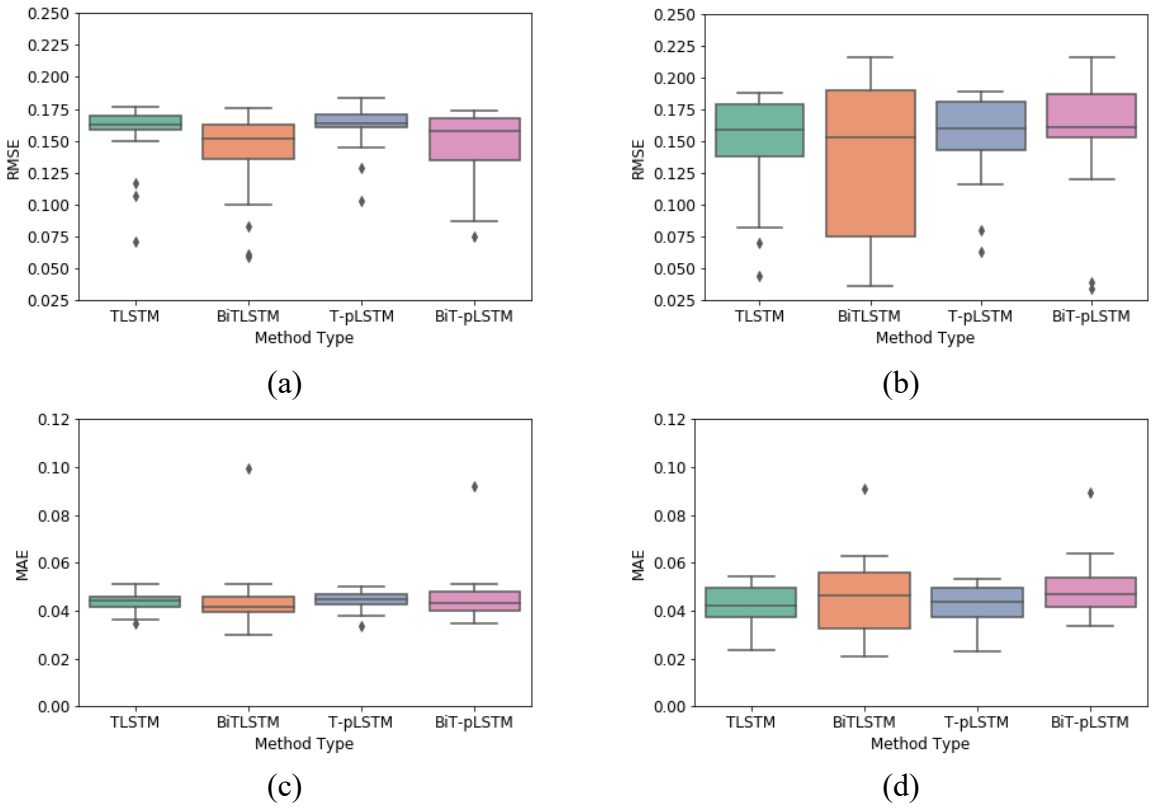


Figure 5.7. Training and test results of T-LSTM, BiT-LSTM, T-pLSTM, and BiT-pLSTM models for tumor size prediction using different sequence length patient records: (a) Training RMSE, (b) Test RMSE, (c) Training MAE, and (d) Test MAE.

Furthermore, I observed that BiT-LSTM has better prediction performance (4.40% validation RMSE) for survival month prediction using fixed length records (Table 5.12 and Figure 5.8). As I noted in the previous section, a limited number of observations might diminish the effect of a slower forgetting mechanism. As shown in Table 5.13 and Figure 5.9, the models with directionality produce higher prediction performance than other LSTM models for survival month prediction using different length records. The best models are BiT-LSTM with 9.96% RMSE and BiT-pLSTM with 9.25% MAE for test data.

Table 5.12. Average RMSE and MAE of four time-aware LSTM networks for survival month prediction using fixed length patient records. The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
T-LSTM	0.1794±0.0509	0.1679±0.0482	0.1346±0.0639	0.1256±0.0577
BiT-LSTM	**0.1722±0.0414**	**0.1600±0.0405**	**0.1255±0.0535**	**0.1155±0.0479**
T-pLSTM	0.2002±0.0429	0.1928±0.0412	0.1595±0.0549	0.1521±0.0496
BiT-pLSTM	0.1871±0.0371	0.1667±0.0364	0.1451±0.0480	0.1267±0.0452

Table 5.13. Average RMSE and MAE of four time-aware LSTM networks for survival month prediction using different length patient records. The lowest RMSE and MAE values are denoted in bold.

Methods	Training		Validation	
	RMSE	MAE	RMSE	MAE
T-LSTM	0.1355±0.0634	0.1289±0.0555	0.1055±0.0683	0.1010±0.0588
BiT-LSTM	**0.1297±0.0646**	0.1239±0.0562	**0.0996±0.0702**	0.0961±0.0611
T-pLSTM	0.1554±0.0523	0.1428±0.0471	0.1221±0.0600	0.1117±0.0520
BiT-pLSTM	0.1349±0.0493	**0.1166±0.0426**	0.1049±0.0580	**0.0925±0.0486**

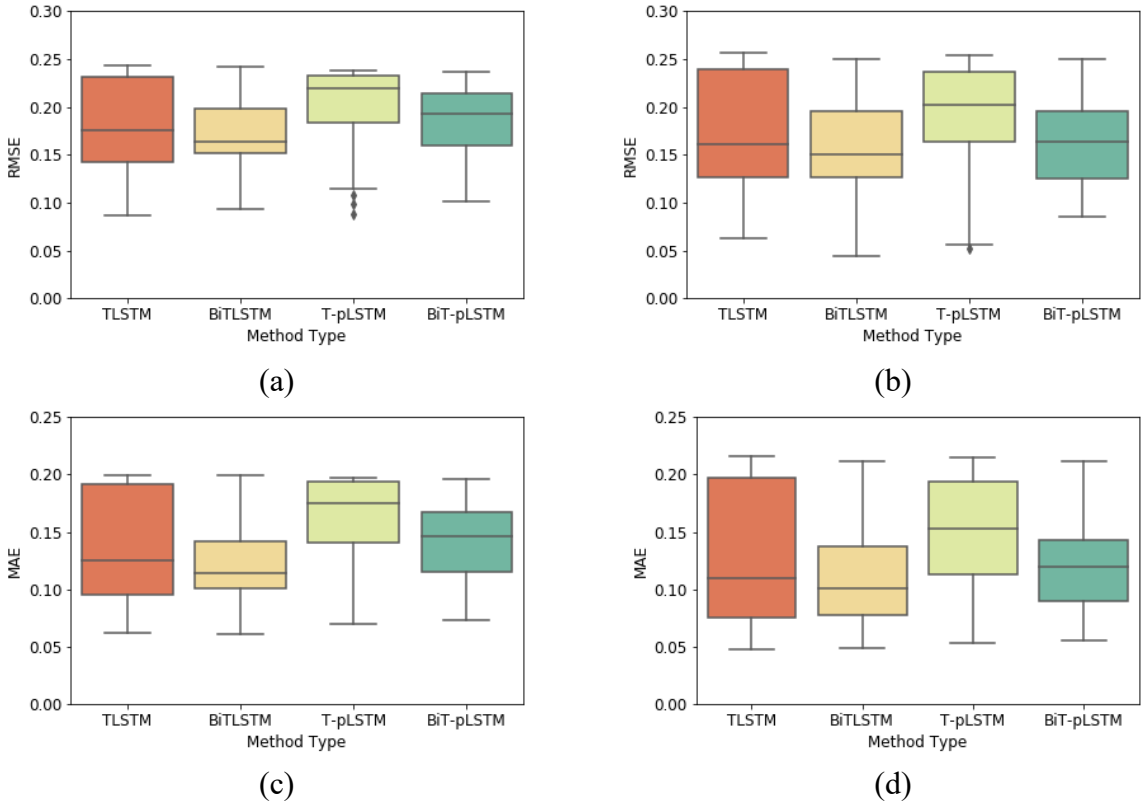


Figure 5.8. Training and test results of T-LSTM, Bi-TLSTM, T-pLSTM, and BiT-pLSTM models for survival month prediction using fixed sequence length patient records: (a) Training RMSE, (b) Test RMSE, (c) Training MAE, and (d) Test MAE.

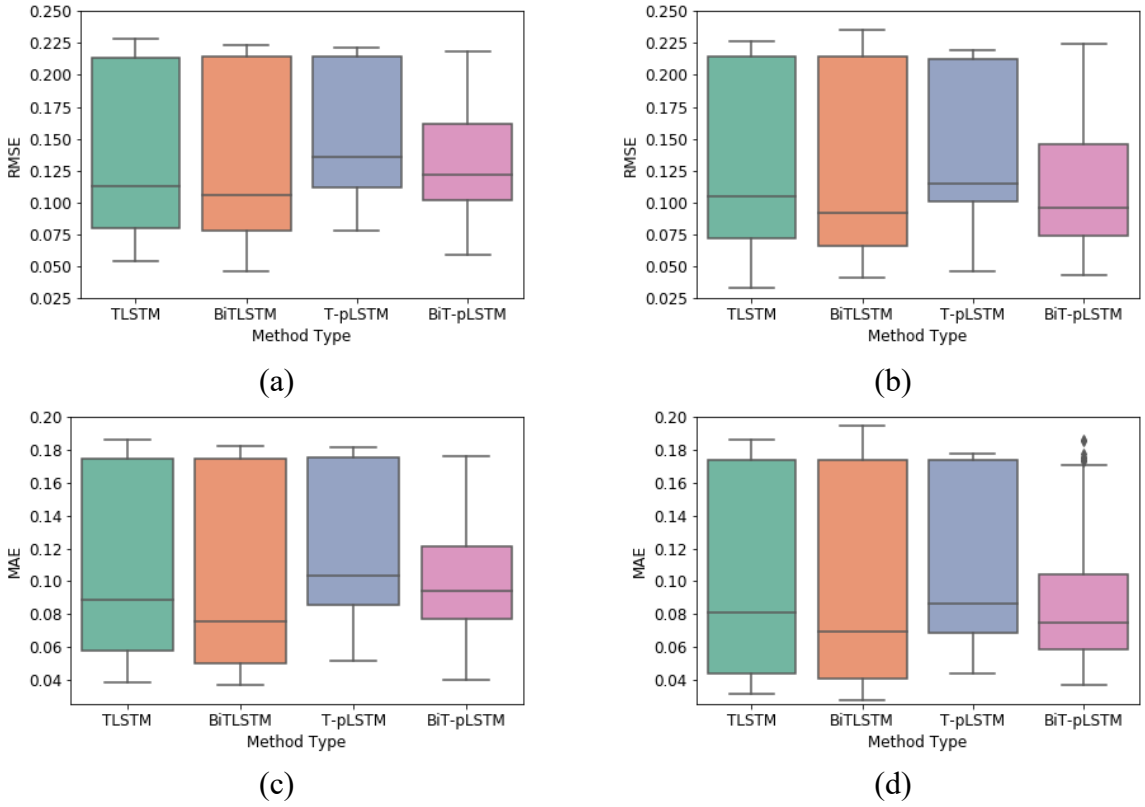


Figure 5.9. Training and test results of T-LSTM, BiT-LSTM, T-pLSTM, and BiT-pLSTM models for survival month prediction using different sequence length patient records: (a) Training RMSE, (b) Test RMSE, (c) Training MAE, and (d) Test MAE.

I adopted the pairwise Wilcoxon rank-sum test again to investigate advancement of our proposed models. For tumor size prediction, Table 5.14 and Table 5.15 present T- pLSTM and BiT-pLSTM outperform LSTM models while BiT-pLSTM only shows better performance than T-LSTM models regardless of sequence length of patient records. Even if our proposed methods show better performance for survival month prediction using fixed length patient records, I could not see improvement when using different length patient records due to the lack of patient records. Survival month prediction models given in Table 5.16 also demonstrate the superiority of T-pLSTM versus other methods using data with fixed length patient records. Also, according to Table 5.17, there is no significance between the methods except the T-pLSTM outperforms the BiT-pLSTM when the patient length records vary in the data.

Table 5.14. Pairwise Wilcoxon rank sum test for T-pLSTM and BiT-pLSTM versus LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for tumor size prediction using fixed length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-pLSTM > LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
T-pLSTM > Bi-LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
T-pLSTM > T-LSTM	0.0994	Fail to reject H_0	0.0796	Fail to reject H_0
T-pLSTM > BiT-LSTM	0.6086	Fail to reject H_0	0.0701	Fail to reject H_0
T-pLSTM > BiT-pLSTM	0.7945	Fail to reject H_0	0.6004	Fail to reject H_0
BiT-pLSTM > LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
BiT-pLSTM > Bi-LSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
BiT-pLSTM > T-LSTM	0.2855	Fail to reject H_0	0.8952	Fail to reject H_0
BiT-pLSTM > BiT-LSTM	0.4109	Fail to reject H_0	**< 0.05**	**Reject H_0**

Table 5.15. Pairwise Wilcoxon rank sum test for T-pLSTM and BiT-pLSTM versus LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for tumor size prediction using different length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-pLSTM > LSTM	0.6533	Fail to reject H_0	0.8887	Fail to reject H_0
T-pLSTM > Bi-LSTM	0.6533	Fail to reject H_0	0.8887	Fail to reject H_0
T-pLSTM > T-LSTM	0.2322	Fail to reject H_0	0.4114	Fail to reject H_0
T-pLSTM > BiT-LSTM	0.3134	Fail to reject H_0	0.6012	Fail to reject H_0
T-pLSTM > BiT-pLSTM	0.7651	Fail to reject H_0	0.1258	Fail to reject H_0
BiT-pLSTM > LSTM	0.8309	Fail to reject H_0	0.8597	Fail to reject H_0
BiT-pLSTM > Bi-LSTM	0.8309	Fail to reject H_0	0.8597	Fail to reject H_0
BiT-pLSTM > T-LSTM	0.2789	Fail to reject H_0	**< 0.05**	**Reject H_0**
BiT-pLSTM > BiT-LSTM	0.3506	Fail to reject H_0	0.1560	Fail to reject H_0

Table 5.16. Pairwise Wilcoxon rank sum test for T-pLSTM and BiT-pLSTM versus LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for survival month prediction using fixed length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-pLSTM > LSTM	**< 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**
T-pLSTM > Bi-LSTM	**< 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**
T-pLSTM > T-LSTM	0.0596	Fail to reject H_0	**< 0.05**	**Reject H_0**
T-pLSTM > BiT-LSTM	**< 0.05**	**Reject H_0**	**≪ 0.05**	**Reject H_0**
T-pLSTM > BiT-pLSTM	**< 0.05**	**Reject H_0**	**< 0.05**	**Reject H_0**
BiT-pLSTM > LSTM	0.8441	Fail to reject H_0	0.6231	Fail to reject H_0
BiT-pLSTM > Bi-LSTM	0.8441	Fail to reject H_0	0.6231	Fail to reject H_0
BiT-pLSTM > T-LSTM	0.9477	Fail to reject H_0	0.9217	Fail to reject H_0
BiT-pLSTM > BiT-LSTM	0.6702	Fail to reject H_0	0.2870	Fail to reject H_0

Table 5.17. Pairwise Wilcoxon rank sum test for T-pLSTM and BiT-pLSTM versus LSTM, Bi-LSTM, T-LSTM, and BiT-LSTM models for survival month prediction using different length patient records (at a specific significance rate $\alpha = 0.05$).

Comparison	Validation RMSE	Hypothesis	Validation MAE	Hypothesis
T-pLSTM > LSTM	0.5188	Fail to reject H_0	0.5363	Fail to reject H_0
T-pLSTM > Bi-LSTM	0.6671	Fail to reject H_0	0.5275	Fail to reject H_0
T-pLSTM > T-LSTM	0.2822	Fail to reject H_0	0.2822	Fail to reject H_0
T-pLSTM > BiT-LSTM	0.1745	Fail to reject H_0	0.2064	Fail to reject H_0
T-pLSTM > BiT-pLSTM	**$\ll 0.05$**	**Reject H_0**	**$\ll 0.05$**	**Reject H_0**
BiT-pLSTM > LSTM	0.1662	Fail to reject H_0	0.2422	Fail to reject H_0
BiT-pLSTM > Bi-LSTM	0.0929	Fail to reject H_0	0.2477	Fail to reject H_0
BiT-pLSTM > T-LSTM	0.5723	Fail to reject H_0	0.7368	Fail to reject H_0
BiT-pLSTM > BiT-LSTM	0.7572	Fail to reject H_0	0.7880	Fail to reject H_0

CHAPTER 6. LESSONS FROM APPLICATION OF MACHINE LEARNING IN NSCLC

6.1. Conclusions

I considered longitudinal patient records with heterogeneous time steps and long-term dependencies for prognostic study of NSCLC. Multivariate sequential information involves interrelationships among consecutive clinical events which highly impact prognosis predictions. In other words, capturing temporal information from clinical data is crucial to accurately predict patient outcomes, and thus provide a proper treatment plan. In addition, the effect of previous clinical events persists to future treatment reassignment, but previous studies overlooked retaining clinical information over longer timescales. In this study, I presented an advanced time-aware model to handle long-timescale patient records with irregular time intervals in predicting patient outcomes. LSTM networks use forget gates following an exponential function with a fast decay rate. Our proposed model includes a power law forget gate with a trainable recurrent coefficient that represents slower information decay. This approach effectively captures time-relevant features which are needed for better clinical decision-making. I evaluated the performance of our models for both tumor size and survival month predictions. My analysis shows that the developed models yield better performance than the standard LSTM and Bi-LSTM models through incorporating temporal knowledge in learning prediction models. I expect that my predictive models using EHRs, or patient data will be helpful for clinicians to guide treatment assignment decisions for NSCLC patients during the long-term treatment plans.

6.2. Limitations and Future Works

In this study, I developed a new LSTM model, called T-pLSTM, for sequential patients' data using the SEER Research Plus database, which is a large database of cancer patient records. I have extracted longitudinal patient records from the database for NSCLC. Despite the large database, the extracted dataset was a small subset. The limited number of patient records diminished the significance of multivariate sequential information for our prediction task. Based on our comparative analysis, accessing more data from various sources in the future might improve the prediction performance of our model. Another limitation regarding datasets is that SEER Research Plus data lacks some important predictors such as smoking status and drug treatment. Incorporating additional predictors into the dataset can result in further improvement of prediction performance. Although I provide these suggestions for further study, it is really challenging to collect appropriate datasets with no missing values and well-distributed variables in this domain. An alternative approach can be using machine learning to generate synthetic data and enlarge the dataset for advanced machine learning algorithms. Furthermore, I only use single layer LSTM architecture to predict patient outcomes for NSCLC in this work. Thus, one research direction would be to extend our approach to multi-layer LSTM models in the future. The effectiveness of the model will be estimated by cross validation to mitigate overfitting problem so as to generate a model with stability. With respect to output response, I will add vital status as additional output in predicting survival month for more practical use of the algorithm. The T-pLSTM model can also be applied to generate robust and efficient patient representations for other prognostic studies. This study only considers LSTM methods and thus in the future, other methods such as GRU with decay function and transformer-based methods can be used to deal with irregular time series data in the healthcare domain. In spite of the limitations, my work is the first to consider time irregularity and

long-timescale information simultaneously in developing advanced machine learning models for clinical research as described in Chapter 1. Chapter 3 explains how we can improve prediction performance of the time-aware LSTM networks and feasibility of the model was demonstrated in Chapter 4. Based on the results, I guarantee that the proposed model will be a necessary first step for additional research to capture valuable time-related information from patient records in making clinical decisions from machine learning models using longitudinal patient records.

REFERENCES

Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viėgas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. Lung cancer survival prediction using ensemble data mining on SEER data. *Scientific Programming*, 20: 29–42, 2012.

American Lung Association, 2022. https://www.lung.org/lung-health diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet. Last accessed 31st March 2022.

Saqib E Awan, Mohammed Bennamoun, Ferdous Sohel, Frank M Sanfilippo, Benjamin J Chow, Girish Dwivedi. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLoS One* 14 (6): 1–13, 2019.

Patrick Baeuerle and Olivier Gires. Epcam (CD326) finding its role in cancer. *British Journalof Cancer*, 96 (3): 417–423, 2007.

Tian Bai, Brian L. Egleston, Shanshan Zhang, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD proceedings. International Conference on Knowledge Discovery and Data Mining*, pages 43–51, 2018.

Seo Jin Bang, Yuchuan Wang, and Yang Yang. Phased-LSTM based predictive model for longitudinal EHR data with missing values. In *KDD proceedings. International Conference on Knowledge Discovery and Data Mining*, 2017.

Carly C. Barron, Philip J. Bilan, Theodoros Tsakiridis, and Evangelia Tsiani. Facilitative glucose transporters: Implications for cancer detection, prognosis and treatment. *Metabolism*, 65 (2): 124–139, 2016.

James A. Bartholomai and Hermann B. Frieboes. Lung cancer survival prediction via machine learning regression, classification, and statistical techniques. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 632–637, 2018.

Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, 2017.

Rachel Bennett. Designing reliable machine learning algorithms for early prediction of preeclampsia. Master's Thesis. University of Oklahoma, 2021.

James Bergstra, Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13 (10): 281–305, 2012.

Silvia Cateni, Marco Vannucci, Marco Vannocci, Valentina Colla. Variable selection and feature extraction through artificial intelligence techniques. *Multivariate Analysis in Management, Engineering and the Sciences: IntechOpen*, pages 103–18, 2013.

Tingting Chai, Roland R Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7 (3): 1247–1250, 2014.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* 8: 1–12, 2018.

Dechang Chen, Kai Xing, Donald Henson, Li Sheng, Arnold M Schwartz, Xiuzhen Cheng. Developing prognostic systems of cancer patients by ensemble clustering. *BioMed Research International* 2009: 1–8, 2009.

Ron Chen, Purvesh Khatri, Pawel K Mazur, Melanie Polin, Yanyan Zheng, Dedeepya Vaka, Chuong D Hoang, Joseph Shrager, Yue Xu, Silvestre Vicent, Atul J Butte, E Alejandro Sweet-Cordero. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer Research*, 74 (10): 2892–2902, 2014a.

Yen-Chen Chen, Wan-Chi Ke, Hung-Wen Chiu. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine*, 48: 1–7, 2014b.

Po Chun Chen, Ta Chung Chi, Shang Yu Su, Yun Nung Chen. Dynamic time-aware attention to speaker roles and contexts for spoken language understanding. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 554–560, 2017.

Hsiang-Yun Sherry Chien, Javier S. Turek, Nicole Beckage, Vy A. Vo, Christopher J. Honey, and Ted L. Willke. Slower is better: Revisiting the forgetting mechanism in LSTM for slower information decay, *arXiv*, 2021.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems,* pages 3512–3520, 2016.

Fran¸cois Chollet et al. Keras. https://keras.io, 2015.

Christine M. Cutillo, Karlie R. Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D. Mandl Tyler Beck, Elaine Collier, Christine Colvis, Kenneth Gersing, Valery Gordon, Roxanne Jensen, Behrouz Shabestari, and Noel Southall. Machine intelligence in healthcare – perspectives on trustworthiness, explainability, usability, and transparency.*npj Digital Medicine*, 3 (47): 1–5, 2020.

George Dimitoglou, James A. Adams, Carol M. Jim. Comparison of the C4.5 and a naive Bayes classifier for the prediction of lung cancer survivability. *Journal of Computing* 4 (8): 1–9, 2012.

Clément-Duchêne C, Carnin C, Guillemin F, Martinet Y. How accurate are physicians in the prediction of patient survival in advanced lung cancer? *Oncologist* 15 (7): 782–789, 2010.

Dmitriy Fradkin, Machine learning methods in the analysis of lung cancer survival data. *DIMACS Technical Report, pages* 2005–35, 2006.

Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Mich`el Schummer, and

David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16 (10): 906–914, 2000.

Sumeet Hindocha, Thomas G. Charlton, Kristofer Linton-Reid, Benjamin Hunter, Charleen Chan, Merina Ahmed, Emily J. Robinson, Matthew Orton, Shahreen Ahmad, Fiona McDonald, Imogen Locke, Danielle Power, Matthew Blackledge, Richard W. Lee, Eric O. Aboagye. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: development and validation of multivariable clinical prediction models. *eBioMedicine* 77: 1–13, 2022.

Sepp Hochreiter and Ju¨rgen Schmidhuber. Long short-term memory. *Neural Computation* 9 (8): 1735–1780, 1997.

Philip C. Hoffman, Ann M. Mauer, and Everett E. Vokes. Lung cancer. *The Lancet* 355 (9202): 479–485, 2000.

Ching Chieh Huang, Soa Yu Chan, Wen Chung Lee, Chun Ju Chiang, Tzu Pin Lu, and Skye Hung Chun Cheng. Development of a prediction model for breast cancer based onthe national cancer registry in Taiwan. *Breast Cancer Research* 21 (92): 1–9, 2019.

Zhangheng Huang, Chuan Hu, Changxing Chi, Zhe Jiang, Yuexin Tong, and Chengliang Zhao. An artificial intelligence model for predicting 1-year survival of bone metastases in non small-cell lung cancer patients based on XGBoost algorithm. *BioMed Research International*, 2020: 1–13, 2020.

V. Jayarama Krishnaiah, Gugulothu Narsimha, N. Subhash Chandra, Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies* 4 (1): 39–45, 2013.

Nikolaos Koulouris. Preventing multiple comparisons problems in data exploration and machine

learning. PhD Dissertation. University of California San Diego, 2020.

Michael H. Kutner, Chris Nachtsheim, John Neter, William Li. *Applied linear statistical models*. Ch. 17, p.744–745, McGraw-Hill Irwin, 2005.

National Cancer Institute. Surveillance, epidemiology and end results (SEER) program (www.seer.cancer.gov) limited-use data (1973–2006), 2008.

National Cancer Institute. Surveillance, epidemiology, and end results (SEER) program (www.seer.cancer.gov) seer*stat database: Incidence - seer research plus data, 18 registries, nov 2020 sub (2000-2018) - linked to county attributes - time dependent (1990–2018) income/rurality, 1969-2019 counties, 2021.

Yu Heng Lai, Wei Ning Chen, Te Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports*, 10: 1–11, 2020.

Joon Lee, David M. Maslove, and Joel A. Dubin. Personalized mortality prediction drivenby electronic medical data and a patient similarity metric. *PLoS One*, 10 (5): 1–13, 2015.

Zachary C Lipton, David Kale, Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. *Proceedings of the 1st Machine Learning for Healthcare Conference*, PMLR 56: 253–270, 2016.

Cheng Lu, Kaustav Bera, Xiangxue Wang, Prateek Prasanna, Jun Xu, Andrew Janowczyk,Niha Beig, Michael Yang, Pingfu Fu, James Lewis, Humberto Choi, Ralph A Schmid, Sabina Berezowska, Kurt Schalper, David Rimm, Vamsidhar Velcheti, and Anant Mad-abhushi. A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. *Lancet Digit Health*, 2 (11): 594–606,2020.

Chip M Lynch, Behnaz Abdollahi, Joshua D Fuqua, Alexandra R de Carlo, James A Bartholomai, Rayeanne N Balgemann, Victor H van Berkel, Hermann B Frieboes. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics* 108: 1–8, 2017.

Warren S. McCulloch, Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* 52 (1-2): 99–115, 1990.

Will McGinnis. Basen encoding and grid search in categorical variables, 2016. URL https://www.pybloggers.com/2016/12/basen-encoding-and-grid-search-in-category_ encoders/. Last accessed 31st March 2022.

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar. *Foundations of machine learning*. Ch. 1, p.1–3, MIT press, 2012.

Maryam Panahiazar, Vahid Taslimitehrani, Naveen Pereira, and Jyotishman Pathaka. Using EHRs and machine learning for heart failure survival analysis. *Stud Health Technol Inform*, 216: 40–44, 2015.

Lynn A. Gloeckler Ries and Milton P. Eisner. *Cancer of the Lung, Chapter 9, SEER Program*. National Cancer Institute, 2007.

Tong Ruan, Liqi Lei, Yangming Zhou, Jie Zhai, Le Zhang, Ping He, and Ju Gao. Representation learning for clinical time series prediction tasks in electronic health records. *BMC Medical Informatics and Decision Making* 19: 1–14, 2019.

David M. Reif, Alison A. Motsinger, Brett A. McKinney, James E. Crowe, Jason H. Moore. Feature selection using a random forests classifier for the integrated analysis of multiple data types. *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pages 1–8, 2006.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

Nishant Sahni, Gyorgy Simon, Rashi Arora. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *Journal of General Internal Medicine* 33 (6): 921–928, 2018.

Mike SchusterKuldip and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45 (11): 2673–2681, 1997.

Kien Wei Siah, Sean Khozin, Chi Heem Wong, Andrew W Lo. Machine learning and stochastic tumor growth models for predicting outcomes in patients with advanced non- small-cell lung cancer. *JCO Clinical Cancer Informatics* 3: 1–11, 2019.

Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 63 (1): 11–30, 2013.

Wenzheng Sun, Mingyan Jiang, Jun Dang, Panchun Chang, Fang-Fang Yin. Effect of machine learning methods on predicting NSCLC overall survival time based on radiomics analysis. *Radiation Oncology* 13: 1–8, 2018.

Leonid Surguchev, Lun Li. IOR evaluation and applicability screening using artificial neural networks. *SPE/DOE Improved Oil Recovery Symposium*, 2000.

Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, NIPS, pages 3104–3112, 2014.

Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi. Can machine learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 12 (4): 1–14, 2017.

Frank Wilcoxon, Individual comparisons by ranking methods. *Biometrics* 1 (6): 80–83, 1945.

Svante Wold, Kim Esbensen, Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2 (1–3): 37–52, 1987.

Hao Xiong, Changjae Kim, and Jing Fu. A data-driven approach to forecasting production with applications to multiple shale plays. *SPE Improved Oil Recovery Conference*, 2020.

Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H. Mak, Hugo J.W.L. Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research* 25 (11): 3266–3275, 2019.

Shudong Yang, Xueying Yu, Ying Zhou. LSTM and GRU neural network performance comparison study: taking yelp review dataset as an example. *2020 International Workshop on Electronic Communication and Artificial Intelligence,* 2020.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126of *Proceedings of Machine Learning Research*, pages 566–588. PMLR, 07–08 Aug 2020.