Lecture Notes in Operations Research

Robin Qiu · Kelly Lyons · Weiwei Chen *Editors*

AI and Analytics for Smart Cities and Service Systems

Proceedings of the 2021 INFORMS International Conference on Service Science

MOREMEDIA

Springer

Lecture Notes in Operations Research

Lecture Notes in Operations Research is an interdisciplinary book series which provides a platform for the cutting-edge research and developments in both operations research and operations management field. The purview of this series is global, encompassing all nations and areas of the world.

It comprises for instance, mathematical optimization, mathematical modeling, statistical analysis, queueing theory and other stochastic-process models, Markov decision processes, econometric methods, data envelopment analysis, decision analysis, supply chain management, transportation logistics, process design, operations strategy, facilities planning, production planning and inventory control.

LNOR publishes edited conference proceedings, contributed volumes that present firsthand information on the latest research results and pioneering innovations as well as new perspectives on classical fields. The target audience of LNOR consists of students, researchers as well as industry professionals.

More information about this series at http://www.springer.com/series/16741

Robin Qiu · Kelly Lyons ·
Weiwei Chen
Editors

AI and Analytics for Smart Cities and Service Systems

Proceedings of the 2021 INFORMS International Conference on Service Science

Springer

Editors
Robin Qiu
Division of Engineering
and Information Science
Pennsylvania State University
Malvern, PA, USA

Weiwei Chen
Department of Supply Chain Management
Rutgers, The State University of New Jersey
Piscataway, NJ, USA

Kelly Lyons
Faculty of Information (Cross-appointed to
Computer Science), Faculty Affiliate,
Schwartz Reisman Institute
University of Toronto
Toronto, ON, Canada

ISSN 2731-040X ISSN 2731-0418 (electronic)
Lecture Notes in Operations Research
ISBN 978-3-030-90274-2 ISBN 978-3-030-90275-9 (eBook)
https://doi.org/10.1007/978-3-030-90275-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Deep Learning and Prediction of Survival Period for Breast Cancer Patients

Shreyesh Doppalapudi[1], Hui Yang[2], Jerome Jourquin[3], and Robin G. Qiu[1(✉)]

[1] Big Data Lab, Division of Engineering and Information Science, The Pennsylvania State University, Malvern, PA 19355, USA
{dshreyesh,robinqiu}@psu.edu

[2] Dept of Industrial Engineering, The Pennsylvania State University, University Park, PA 16802, USA
huy25@psu.edu

[3] Susan G. Komen, PO Box 801889, Dallas, TX 75380, USA
jjourquin@komen.org

Abstract. With the rise of deep learning, cancer-specific survival prediction is a research topic of high interest. There are many benefits to both patients and caregivers if a patient's survival period and key factors to their survival can be acquired early in their cancer journey. In this study, we develop survival period prediction models and conduct factor analysis on data from breast cancer patients (Surveillance, Epidemiology, and End Results (SEER)). Three deep learning architectures - Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) are selected for modeling and their performances are compared. Across both the classification and regression approaches, deep learning models significantly outperformed traditional machine learning models. For the classification approach, we obtained an 87.5% accuracy and for the regression approach, Root Mean Squared Error of 13.62% and R^2 value of 0.76. Furthermore, we provide an interpretation of our deep learning models by investigating feature importance and identifying features with high importance. This approach is promising and can be used to build a baseline model utilizing early diagnosis information. Over time, the predictions can be continuously enhanced through inclusion of temporal data throughout the patient's treatment and care.

Keywords: Deep learning · Breast cancer · Survival period prediction · SEER cancer registry · Factor analysis · Feature importance

1 Introduction

Cancer is the second leading cause of death in the United States. According to the latest statistics from the U.S. National Cancer Institute, about 1.8 million new cancer cases are expected in 2020, with about 600,000 expected new deaths [1]. Breast Cancer is the most common cancer in the United States, with 276,480 estimated new cases and 40,170 estimated deaths in 2020 [2]. Though cancer is prevalent throughout the United States, the relative five-year survival rate for breast cancer patients varies by breast

R. Qiu et al. (Eds.): ICSS 2021, LNOR, pp. 1–22, 2021.
https://doi.org/10.1007/978-3-030-90275-9_1

cancer stage. Note that the average risk of a woman in the United States developing breast cancer sometime in her life is about 13% [3].

Estimated national expenditures for cancer care in the United States in 2017 were $147.3 billion [4]. The costs are expected to increase in the coming years with an aging population and newer, more advanced treatment options. An increase in the number of patients requiring care also drives up health care costs for treating the disease. Research on disease prognosis from early information and clinical data is still needed to better manage resources and time, for both the patient and the healthcare professionals. For patients, better prognosis early can help plan time spent in care and costs, while for healthcare professionals it can help plan medical resources, treatment and care strategy for each patient. These reasons make cancer survivability a topic of great importance to both researchers from the medical domain, computer intelligence, and medical professionals.

This study is aimed at predicting the survival period of a breast cancer patient, using the data obtained from Survey Epidemiology, and End Results (SEER) cancer registry [20]. The research problem is formulated through both classification and regression techniques. We investigate three different deep learning models and their performance at providing high dimensionality and highly non-linear decision boundaries and identifying features that drive survival prediction. This study contributes to identifying an appropriate approach to estimating and understanding survivability that could be widely and practically appropriate for medical use.

2 Related Works

Chronic diseases that may threaten the lives of patients and require continuous care and surveillance are the focus of a majority of previous research. These studies aim at projecting a timeline of survival to help both patients and healthcare professionals plan care, time, and resources.

Some researchers have focused on using the results from various laboratory tests and medical images such as Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI) scans to predict disease survival for patients. Macyszyn *et al.* [5] have used a combination of Linear SVM for predictions and Kaplan-Meyer survival curves to plot the survival period and map the prognosis from the MRI images of patients with glioblastoma. Qian *et al.* [6] proposed Temporal Reflected Logistic Regression (TRLR), a model to predict survival probabilistic score on a set of patients diagnosed with heart failure at Mayo Clinic, Rochester. They showed that the new model outperformed the existing Seattle Heart Failure Model (SFHM). Marshall *et al.* [7] proposed a custom Bayesian model, Continuous Dynamic Bayesian Model, that differs in data availability at different points during the treatment of the patient, making effective use of the temporal data sequence. Zhang *et al.* [8] used LASSO as the regularization and feature selection method in a multilayer perceptron model to predict the 5-year survival of patients suffering from chronic kidney disease. Bellot *et al.* [9] developed Hierarchical Bayesian models for cardiovascular disease, which utilized a mixture of localized and individual patient data to build average survival paths for patients and specific survival paths for individual patients. They developed a personalized interpreter for the model to highlight the factors affecting the survival prediction for each patient.

Cancer has been the focus of a large body of research targeting multiple aspects: diagnosis, prognosis, best treatment estimation and survival period estimation. Gray *et al.* [10] tested the performance of the Predict algorithm (version 2), built for 5-year and 10-year prognostic indicators for breast cancer, using the Scottish Cancer Registry (SCR). They found that the algorithm overestimates some 5-year prognostic cases while the overestimation is lower for 10-year prognosis. Song *et al.* [11] built nomographs to predict overall and cancer-specific survival of patients with chondrosarcoma and identified the prognostic factors for 3-year and 5-year survival. Lynch *et al.* [12] compared multiple supervised machine learning methods to predict survival time of lung cancer patients. Tested algorithms include Linear Regression, Gradient Boosting Machine, Random Forest, Support Vector Machine and a custom stacking ensemble combining all the other machine learning algorithm models. The custom ensemble performed the best with an RMSE (Root Mean Squared Error) value of 15.30%. Sadi *et al.* [13] proposed the use of clustering cancer registry data before feature selection for classification model on survival period prediction. They used the Egyptian National Cancer Institute (NCI) cancer registry data to perform clustering and then build classification models inside each cluster to get a better sense of hyperparameter tuning and achieve better performance.

The SEER dataset provides comprehensive information on the early diagnosis of patients for multiple types of cancer along with the survival time for patients. It has been a part of a multitude of research around prediction of survival period. Bartholomai and Frieboes [14] compared multiple traditional machine learning models in an attempt to predict the exact survival period for a lung cancer patient by modeling the problem using both classification and regression techniques. They used Principal Component Analysis (PCA) to identify the features that have the biggest impact on the survival of lung cancer patients. Hegselmann *et al.* [15] developed models with a focus on producing reproducible models for further research use, to predict 1-year and 5-year survival for patients with lung cancer using logistic regression and multilayer perceptron models. Naghizadeh and Habibi [16] compared various ensemble machine learning models to predict survival in cancer comorbidity cases for various cancers. Dai *et al.* [17] built 1-year and 3-year nomographs to predict survival for triple negative breast cancer patients with a histology of infiltrating duct carcinoma. Imani *et al.* [18] used random forests as a modeling approach to predict survival in patients with breast cancer recurrences and identified the important variables in defining survivability for those patients. Kleinlein and Riano [19] built joint and stage-specific machine learning models to predict 5-year survival for breast cancer patients. They also tested the performance of these models over time from training to demonstrate the feature importance in selection for models. Shukla *et al.* [20] performed clustering using Self-Organizing Maps (SOM) and DBSCAN to cluster the patients into related cohorts and build multilayer perceptron to predict 3-year, 5-year and 7-year survival of breast cancer patients.

Most of previous research into survival prediction in breast cancer has focused on the binary classification model of whether a patient survives a preset period of time or not. To the best of our knowledge, little research has focused on the predicting the exact survial periods for breast cancer patients. This study aims to bridge this gap by utilizing deep learning models to predict survival periods in breast cancer cases, and further

provide interpretability to these models through feature importance analysis, revealing the features with the largest impact on the survival predictions.

3 Dataset

In this research study, the data was collected from Surveillance, Epidemiology, and End Results (SEER) program cancer registry. The SEER program is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS) and was established to provide consistent statistics around cancer incidence in the United States. SEER collects its information from several population-based registries spread geographically across the United States and estimated to cover about 34.6% of the United States population. The SEER dataset is one of the biggest and most comprehensive databases on the early diagnosis information for cancer patients in the United States [21]. The SEER cancer registry includes information on primary tumor, tumor histology, morphology, the first course of treatment and follow-up with patients to record the vital status. All the records are de-identified, with very little demographic information and all the entries are encrypted to provide another layer of security for the patients' identity. It includes statistics on incidence of multiple types of cancer such as breast cancer, lung cancer, intestinal cancer, leukemia, lymphoma, prostate cancer and others.

3.1 Data Collection and Cleaning

The incidence files in the SEER cancer dataset are available as a set of ASCII text files with each row in a text representing a record in the registry and the information encrypted in codes. The whole database was segmented by sets of registries and the cancer type. We selected the November 2018 submission (which was made publicly available in April 2019) for our research.

This submission contained information for the diagnosis years of 1975–2016. For this research, we only selected the files for breast cancer and built Python codes to read ASCII text files and return comma separated values (.csv) files, which would be easier to ingest for data engineering purposes. The Python codes ignored all records with diagnosis year before 2004 to minimize the impact of now-defunct data fields on our modeling process. (The registry went through a major data overhaul in 2004 and new fields were added to replace the older fields.)

Further, the codes representing missing values in the registry were replaced by null values to standardize the data loss at both individual registry level and the national registry level. Also, the records of patients with cancer comorbidities were eliminated. Only patients with breast cancer and no comorbidities were considered to reduce the effect of other cancers on the survival analysis. Records with missing diagnosis year and survival time information were eliminated as these values could not be imputed, because any imputation for these fields affects the survival analysis directly. This cleaning resulted in a dataset with 197,038 records for analysis and modeling.

3.2 Data Preprocessing

The features were divided into categorical and quantitative variables. Also, features that are built on the target variable, such as variables defining cause of death classifications, were identified through domain knowledge gained by a better understanding of the dataset. These variables were not included in the modeling analysis to avoid model bias.

Different pre-processing techniques were used for categorical and quantitative variables. Variables with more than 20% missing values were eliminated from modelling, because they were either data collected during different time periods (before 2004 diagnosis year) or were not relevant measurements for breast cancer. For the remaining variables, the median and mode were used to impute missing values for quantitative and categorical variables, respectively. After the imputation, categorical variables were binarized using one-hot encoding.

Figure 1 shows the distribution of patients across different survival periods. For the classification approach of predicting survival period, the survival time in months was segmented into three bins '< =5 years', '5–10 years', and '>10 years' used as the target class for multiclass classification modeling. These classes were decided through knowledge gained from previous research approaches to this problem, where 5-year and 10-year survival models are the most common. Yet, imbalance exists in the data coming from '>10 years' class as shown in Fig. 2. To adjust for this imbalance, we upsampled other classes to create a balanced dataset for model tuning and evaluation. For the regression approach, we selected the survival time in months as the target variable.

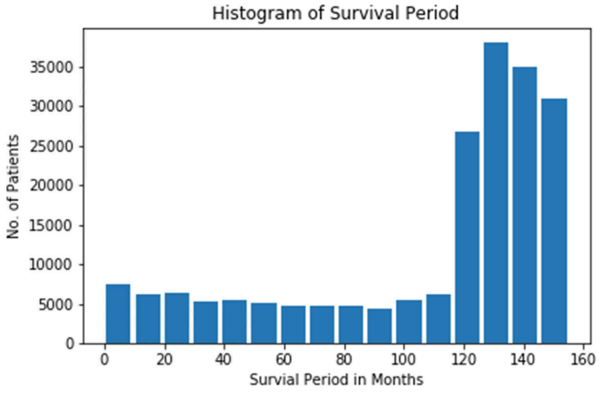


Fig. 1. A histogram to understand the distribution of breast cancer patients across various survival period.

The features have different scales, which could lead to a longer time for the model to converge and reach the global minima while also increasing the probability for the model to get stuck in local minima. Hence, the predictors were all normalized for the classification approach, while for the regression approach, the target variable was also scaled for easier comparison between errors across model iterations and between different models. The entire dataset was partitioned in the ratio of 80:20 into training and test

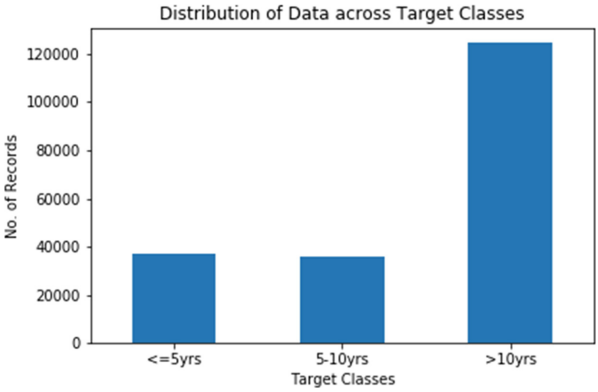


Fig. 2. Distribution of Records across the selected three target classes.

sets. The models were trained on the training set and the performance was observed on the test set.

4 Research Methodology

The data preprocessing methods remain similar for the two different approaches to the problem (i.e., classification and regression), except for the target variable in the dataset preparation. However, the model building methodologies differ for both approaches with the different tuning approaches and evaluation metrics for both approaches. Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) are the three selected deep learning architectures to model the problem.

4.1 Deep Learning Architectures

This research study focuses on the performance of deep learning models in the space of cancer survival period prediction. For this purpose, we have selected three of the most popular deep learning architectures in practice today- ANN, RNN and CNN.

Neural networks are sub-field of the rapidly growing deep learning space, which is in itself a sub-field of machine learning. The rapidly growing popularity of deep learning can be attributed to its performance to model unstructured data such as text and images and to infer meaning at par with or better than human accuracy.

A basic building block of a neural network is a perceptron. A perceptron provides each input with a weight and creates a function to represent the input space. This representation is then passed through an activation function to produce a probability of the input sequence representing the output or not. The simplest perceptron model can mimic a logistic regression model through the use of just one perceptron. A neural network is formed by arranging groups of perceptrons in multiple layers to represent the output with activation functions defined for each layer. A perceptron is known as a neuron inside a neural network, with the name representing the tendency of neural network to mimic the performance of neurons inside the human brain. Neural networks have three major

units - input layer, hidden layers and an output layer. The size of the input and output layers are always defined by the problem, and the architecture of the hidden layers is always defined by a machine learning practitioner which helps to model the problem more efficiently. As the number of hidden layers increases, the defined representation becomes more complex and highly non-linear.

The earliest defined neural network is ANN [22], which consists of neurons arranged in multiple layers, with each neuron in one layer fully connected to each neuron in the next layer. The simplest ANN architecture consists of an input layer, a hidden layer, and an output layer. As the number of hidden layers increase, the number of parameters to train in the network increases and the representation becomes more complex.

A neural network where the output of a neuron is fed back as input after certain time units is known as RNN [23]. This ability makes it easy for them to model temporal sequences of data such as text and time-series data. But, with the output of neurons being fed back to the input neuron creates a gradient vanishing problem when the network is very large and the input sequence is too long. To overcome this problem, Long Short Term Memory Networks (LSTM) [24] were proposed as a variant of the simple RNN architecture. LSTM introduced the concept of a forget gate that limits the sequence size that is under consideration at any particular time, thereby overcoming the vanishing gradient problem. We use LSTM networks for our research and hereby in this paper, RNN refers to LSTM architecture.

A neural network model where the weights are shared across a few neurons in a layer is known as CNN [25]. The shared weights concept makes CNN both space and shape invariant and makes them the ideal network to model data from images and videos. This helps in bringing down the number of training parameters that would be required if the fully connected layers were used to model images. In this study, we are working with 1-dimensional data and hence, use the 1-D convnet architecture. In this paper, CNN refers to the 1-D convnet architecture.

4.2 Model Architecture and Parameters

4.2.1 Classification Model

The architecture and parameters of the classification model under study are given in Table 1. With '< =5 years', '5 – 10 years', and '>10 years' as the target classes, this model attempts to address a multiclass classification problem. Each model has a unique model architecture with a varying distribution of neurons in each layer, number of layers, and number of iterations (epochs) over which the model is trained. In addition, to prevent overfitting of the data, regularization was performed through the addition of Dropout layers between the hidden layers. A few hyper-parameters remain similar across all models, such as ReLU (Rectified Linear Unit) as the activation function in the hidden layers, Softmax as the activation function in the output layer, Adam (Adaptive Moments Estimation) as the optimizer with a 0.001 learning rate and Categorical Cross Entropy as the loss function.

ANN Model: The ANN model has 5 hidden layers with 600, 300, 100, 50 and 20 neurons in each layer, respectively. The model was run over 100 epochs with 10% Dropout layer after the first two hidden layers.

Table 1. Parameters for classification models

Parameters	ANN	RNN	CNN
Target classes	[<=5 Years, 5 – 10 Years, >10 Years]		
No. of hidden layers	5	4	5
No. of neuron in each hidden layer	[600, 300, 100, 50, 20]	[256, 128, 64, 32]	[512, 256, 128, 64, 32]
Activation functions in each layer	ReLU in hidden layers and Softmax on output layer		
Loss function	Categorical cross entropy		
Optimizer	Adam (Adaptive Moments Estimation) with 0.001 learning rate		
No. of epochs	150	100	100
Dropout layers for regularization	10% dropout layer after first two hidden layers	10% dropout layer after first two hidden layers	10% dropout layer after first and third hidden layers

RNN Model: The RNN model has 4 hidden layers with 256, 128, 64 and 32 neurons in each layer, respectively. The model was run over 100 epochs with 10% Dropout layer after the first two hidden layers.

CNN Model: The CNN model has 5 hidden layers with 512, 256, 128, 64 and 32 neurons in each layer, respectively. The model was run over 100 epochs with 10% Dropout layer after the first and third hidden layers.

4.2.2 Regression Model

The architecture and the parameters of the regression model under study are given in Table 2. With the normalized survival period as the target, this model attempts to address a regression problem. Mean squared error was selected as the loss function since it is a regression problem with each model having a varying distribution of neurons in each layer, number of layers, and number of iterations (epochs) over which the model is trained. Again, to prevent overfitting, Dropout layers were added to the models between the hidden layers to perform regularization of the network. A few hyper-parameters still remain similar across models such as ReLU (Rectified Linear Unit) as the activation function in the hidden layers, no activation function in the output layer, and Adam (Adaptive Moments Estimation) as the optimizer with a 0.001 learning rate.

ANN Model: The ANN model has 3 hidden layers with 100, 60, 20 neurons in each layer, respectively. The model was run over 20 epochs with no Dropout layer.

RNN Model: The RNN model has 2 hidden layers with 50, 20 neurons in each layer, respectively. The model was run over 25 epochs with 20% Dropout layer after the first and last hidden layers.

CNN Model: The CNN model has 2 hidden layers with 40, 20 neurons in each layer, respectively. The model was run over 50 epochs with no Dropout layer.

Table 2. Parameters for Regression models

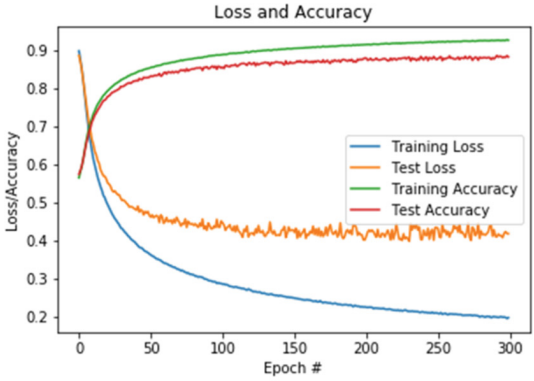
Parameters	ANN	RNN	CNN
No. of hidden layers	3	2	2
No. of neuron in each hidden layer	[100, 60, 20]	[50, 20]	[20, 40]
Activation functions in each layer	ReLU in hidden layers		
Loss function	Mean squared error		
Optimizer	Adam (Adaptive Moments Estimation) with 0.001 learning rate		
No. of epochs	20	25	50
Dropout layers for regularization	None	20% dropout layer after every hidden layer	None

4.3 Model Tuning

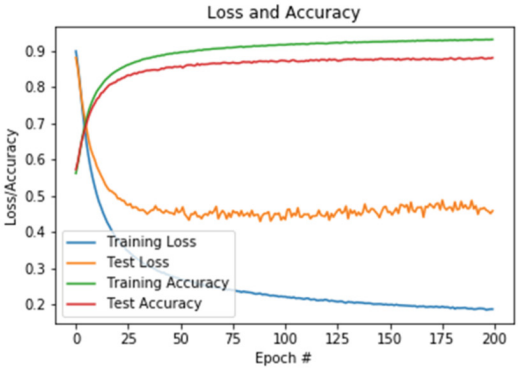
The models were initially tuned over a large number of epochs to observe the performance of the models on both the training and test sets.

4.3.1 Classification Model

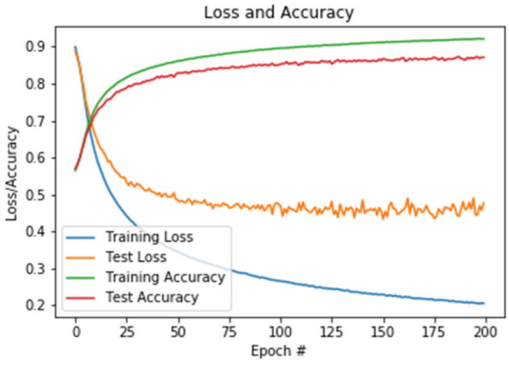
The plots of the training and test loss and accuracy vs. the number of epochs are shown in Fig. 3. The optimal numbers of epochs for the ANN, RNN, and CNN models are determined to be approximately 150, 100, and 100, respectively. The training loss and test loss can be seen to diverge with the test loss stabilizing due to regularization and the training loss decreasing at a rapid pace for all the models near the optimal number of epochs.



(a) ANN Model

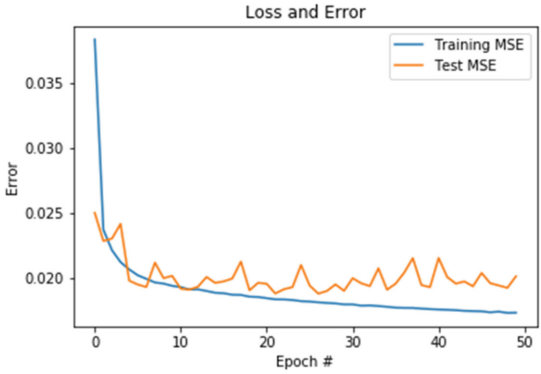


(b) RNN Model

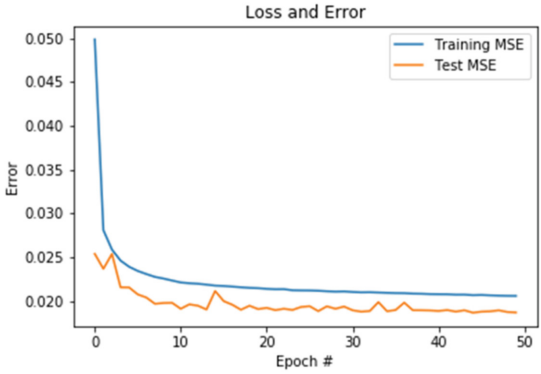


(c) CNN Model

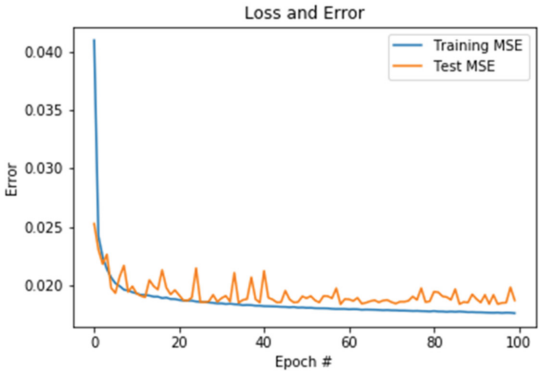
Fig. 3. Epoch Plots for Classification models. Epoch plots are required to find the optimal number of Epochs to train the models by observing training loss and test loss and finding the point where over-fitting starts.



(a) ANN Model



(b) RNN Model



(c) CNN Model

Fig. 4. Epoch Plots for Regression models. Epoch plots are required to find the optimal number of Epochs to train the models by observing training loss and test loss and finding the point where over-fitting starts.

4.3.2 Regression Model

The plots of the training and test loss and accuracy vs. the number of epochs are shown in Fig. 4. The optimal numbers of epochs for the ANN, RNN, and CNN models are determined to be approximately 20, 25, and 50, respectively. The training loss and test loss can be seen to diverge near the optimal number of epochs with the test loss increasing and the training loss decreasing at a rapid pace for ANN model, while test loss stabilized and training loss decreased at a rapid pace for both the RNN and CNN model.

4.4 Models for Comparison with Previous Research

To put our results in perspective, we compared our results with the results from traditional machine learning models approaches. Feature selection for classification models was done using one-way ANOVA tests for continuous variables and χ^2 test for categorical variables. Similarly, for regression models, one-way ANOVA tests were used for categorical variables and univariate linear regression for continuous variables. These tests give variables that are most relevant to the breast cancer classification and regression models.

4.4.1 Classification Model

For the classification model, we have implemented the following models:

- *Random Forest Classifier* with 500 tree count, depth of 3 variables and minimum node size of 50.
- *Support Vector Machines* with a linear kernel, primal optimization and a regularization parameter of 1.7e07.
- *Naïve Bayes Classifier* with a Complement Multinomial distribution assumed for the features and Laplace smoothing parameter of 2.37.

The hyper-parameters for these models were selected through a random search for the optimal hyper-parameters.

4.4.2 Regression Model

For the regression model, three models were implemented:

- *Random Forest Regressor* with 500 tree count, depth of 3 variables and minimum node size of 50.
- *Gradient Boosting Machine* with 1000 tree count, interaction depth of 1, 0.21 learning rate and minimum node size of 100.
- *Line Regression Model* with default parameters.

The hyper-parameters for these models were selected through a random search for the optimal hyper-parameters.

4.5 Feature Importance

There has been a growing demand to increase the interpretability of machine learning models. This is especially true for the models that operate in heavily regulated industries such as healthcare, mortgages, government and military. Understanding feature importance in models is a step towards being able to interpret the learning patterns that the machine learning model interpreted from the data and would help in making the models more interpretable, bias-free and easier to audit. These reasons contribute to understanding feature importance being an important step in building machine learning models in the healthcare sector.

To understand important features towards predicting survival period, we employed two different types of model-agnostic machine learning model interpretation techniques - one to explain the predictions locally at each prediction level and another to explain predictions on a global level. At a local level, we aim to understand for each prediction which feature plays an important role in outputting that particular prediction. At a global level, we try to understand the impact each feature has on the final prediction. In other words, we look at the final output as a function of each predictor individually.

For local level interpretability, we used Shapley Addictive exPlanations (SHAP) [26] values to understand the features driving a set of sample predictions. SHAP is a framework for machine learning interpretation that combines many algorithms such as LIME [27], DeepLIFT [28], Shapley Sampling Values [29], Shapley Regression Values [30], Quantitative Input Influence [31] and layer-wise relevance propagation [32]. SHAP works as a model agnostic interpretation and has no performance issues even on blackbox models such as deep neural networks, random forests or gradient boosting machines. SHAP provides a feature ranking for each prediction based on the amount of influence each predictor had on the final prediction. SHAP value for a particular predictor is provided as a percentage importance considering the most important predictor to have a base value of 100%.

For global level interpretability, we have built Partial Dependence Plots (PDP) [33] to understand the effect that each feature in the dataset has on the final prediction across the complete training set. A partial dependence plot helps to visualize how the value of final prediction changes with the changes in the feature value. Also, the distribution of each feature was plotted to understand if the effect of each prediction is significant or not. Partial dependence plots work on the assumption of binary classification. Hence, to adapt it to the multiclass classification problem, we have employed One-vs-All approach to describe the feature importance for each class.

For this study we investigated the interpretability of the results from our deep learning models with the feature importance, which helps identify the most important features at the local level and global level of interpretability.

4.6 Experimental Setting

The processes of data extraction, cleaning, visualization, transformation and model building were done in Python. For those purposes, various open-source technological stacks such as Numpy [34], pandas [35], Scikit-learn [36], TensorFlow [37], Keras [38] and Matplotlib [39] were used. For providing the model interpretability and feature importance using SHAP and Partial Dependence Plots, InterpretML [40] was used.

5 Results and Discussion

We have used different evaluation metrics to evaluate the performance of the models for the classification and regression approaches due to the inherent differences in these approaches.

5.1 Evaluation Metrics

Precision, Recall, F-1 Score, Accuracy, and Cohen's κ were selected as the evaluation metrics for the classification models built in this study. These evaluation metrics are defined by the following equations [1–5]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\kappa = \frac{P_O - P_e}{1 - P_e} \tag{5}$$

where TP is True Positives, FP is False Positives, TN is True Negatives, FN is False Negatives, P_O is the probability of observed agreement, and P_e is the probability of random agreement between raters.

For the regression modeling, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) value were used as the evaluation metrics for the regression model. These evaluation metrics are defined by the following equations [6–8]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{8}$$

where, y_i is the i^{th} observed value, $\hat{y}_i$ is the i^{th} predicted value, $\bar{y}$ is the mean of observed values, and n is the total number of values in the dataset.

5.2 Classification Model Results

Deep learning models significantly outperform other baseline models of Random Forest Classifier, SVM classifier and the Naïve Bayes Classifier (Table 3). ANN turns out to be the best performing model with achieving 87.50% accuracy, slightly higher than RNN and CNN.

As shown in Table 3, the performance of deep learning models decreases as the survival period of the class increases. ANN, RNN and CNN models have much lower performance on the '>10 years' class. In contrast, the traditional machine learning models have a much lower performance on the bounded class '5–10 years' in comparison with the unbounded '<=5 years' and '>10 years'. Note the exceptionally low performance of the random forest classifier on the middle class. Among the traditional machine learning models, the ensemble classifier Random Forest Classifier performs much better that SVM and Naïve Bayes Classifier. Naïve Bayes Classifier has the worst performance among all the classifiers at 48.82% accuracy.

The low recall for the '>10 years' and the low precision for the '5–10 years' for the deep learning models suggest that the models find it difficult to differentiate between medium length survival period prediction and long survival period prediction. This brings down the overall performance of the models over '5–10 years' and '>10 years' classes in comparison with the '<=5 years' class.

Table 3. Performance comparison of classification models

Model	Target class	Precision	Recall	F-1 score
ANN model	<=5 years	90.94%	94.29%	92.59%
	5–10 years	84.06%	91.67%	87.70%
	>10 years	87.73%	76.54%	81.75%
	Accuracy	87.50%		
	Cohen's κ	81.24%		
RNN model	<=5 years	89.05%	95.11%	91.98%
	5–10 years	85.48%	91.31%	88.30%
	>10 years	87.95%	75.97%	81.52%
	Accuracy	87.46%		
	Cohen's κ	81.18%		
CNN model	<=5 years	90.17%	93.85%	91.97%
	5–10 years	83.85%	89.60%	86.63%
	> 10 years	85.70%	76.32%	80.74%
	Accuracy	86.58%		
	Cohen's κ	79.87%		
Random forest classifier	<=5 years	63.84%	65.62%	64.72%

(*continued*)

Table 3. (*continued*)

Model	Target class	Precision	Recall	F-1 score
	5–10 years	52.16%	39.07%	44.68%
	>10 years	56.82%	69.56%	62.55%
	Accuracy	58.04%		
	Cohen's κ	37.07%		
SVM classifier	<=5 years	58.32%	62.98%	60.56%
	5–10 years	47.09%	29.89%	36.57%
	>10 years	53.65%	69.07%	60.39%
	Accuracy	53.93%		
	Cohen's κ	30.91%		
Naïve Bayes classifier	<=5 years	54.08%	56.69%	55.35%
	5–10 years	41.59%	30.35%	35.09%
	>10 years	48.67%	59.55%	53.56%
	Accuracy	48.82%		
	Cohen's κ	23.24%		

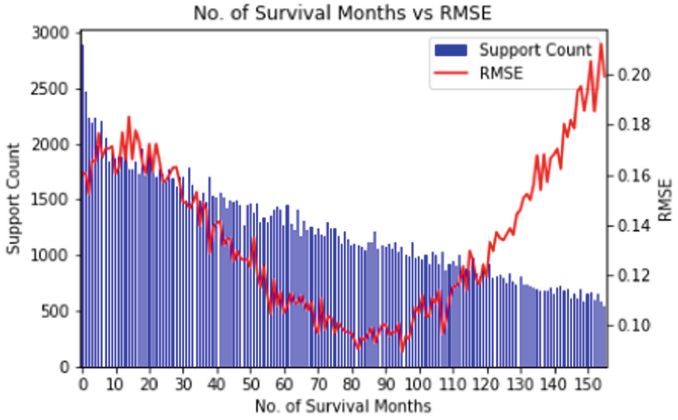
5.3 Regression Model Results

As shown in Table 4, deep learning models significantly outperform other baseline methods with a lower RMSE and higher R^2 value. CNN turns out to be the best performing model achieving 13.62% RMSE and 76.37% R^2 value. CNN's performance is just marginally better than that of ANN and RNN models.

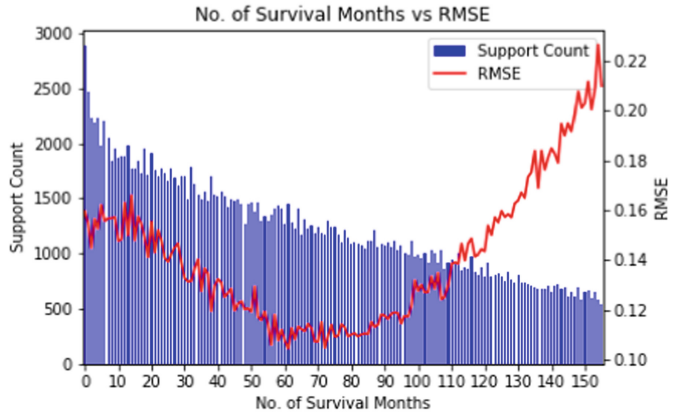
We have plotted the RMSE vs. Survival Period in months along with support count in Fig. 5 for all the models to better understand the error progression over survival period. Figure 5 shows that the error in prediction for survival period is high for both breast cancer patients with long survival periods and for those with very short survival period. The error is lowest for the mid-level survival patients. This indicates that the models

Table 4. Performance comparison of regression models

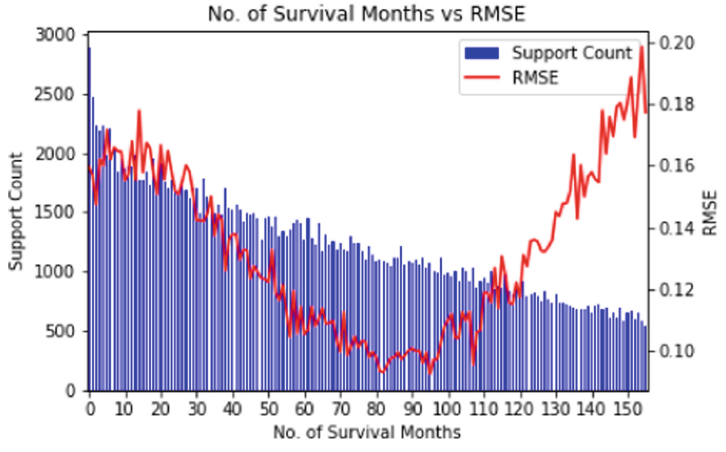
Model	MSE	RMSE	R^2 value
ANN	0.020	13.98%	75.09%
RNN	0.020	13.99%	75.04%
CNN	0.019	13.62%	76.37%
Linear regression	0.022	15.01%	71.27%
Gradient boosting Machine	0.021	14.58%	72.68%
Random forest regression	0.021	14.16%	74.01%



(a) ANN Model



(b) RNN Model



(c) CNN Model

Fig. 5. RMSE vs. No. of Survival Months plot for the deep learning models. The error in prediction is the lowest for mid-level survival period patients.

find it difficult to accurately estimate the survival period for low survival period patients since breast cancer patients have a high mean survival period. The confidence in the prediction is the highest for mid-level survival and again decreases for long survival period patients as the number of cases decreases for high number of survival months. This indicates that the regression approach cannot be useful in an unbounded approach to modeling the patient survival period. The regression approach works best when the survival period is bounded on the lower and higher side, away from which the error increases to an unacceptable level while the support for the survival period goes below an acceptable threshold value. For example, in the case of breast cancer, the survival period could be bounded between 60 and 120 months (5 and 10 years, respectively), as the RMSE is above 12% (which is assumed to be an acceptable threshold of error) for survival period less than 60 months or greater than 120 months. The other survival period needs more accurate substitution to help in predicting the exact survival period.

Also, notice the almost identical error progression occurs for all the deep learning models in the figure. This leads us to believe that various deep learning models can model a similar non-linear function from the predictors provided with the data to represent the survival period.

5.4 Discussion

As shown above, deep learning methods significantly outperform the baseline models which include traditional machine learning methods such as linear regression, random forests, gradient boosting machine, support vector machine and naïve Bayes across both classification and regression approaches. This again represents the ability to learn highly non-linear decision boundaries from high dimensional data, an aspect where neural networks tend to do better than other traditional machine learning approaches.

The performance of both RNN and CNN is not better than the performance of ANN across both classification and regression approaches. The absence of temporal or sequential data in the dataset could be responsible for this differential. RNN and CNN can shine with multi-dimensional data such as text or images where they optimize much faster. Their performance is comparable to ANN without temporal data and represent a loss of time in the project due to the longer training time that they take on a similar network size as an ANN. This reinforces that selection of model architecture depends more closely on the data than any other external factors.

This also opens another avenue to improve the performance of survival models - through the use of temporal data. The built models will be enhanced through the availability of temporal data, which can help in providing more accurate predictions. The build models can be further enhanced through the patient treatment period, resulting in the increase of prediction accuracy for cases with long survival period.

Prediction of survival period should still be bounded within an acceptable threshold for the regression approach as many other factors currently not represented in the data could drive the survival period for longer survival cases. The acceptable bounds for prediction should be defined based on a specific cancer type as the behavior and survival period distribution differs for each type of cancer. As a result, cancer-specific survival models help model the intricacies of each cancer with respect to the survival period.

Also, specifically for breast cancer, the error in regression for low survival period is high, which could be augmented by combining with a classification approach. The classification approach provided high performance on the low survival period class while the regression approach lagged in performance on low survival periods. A combination of both approaches could help provide a sufficient threshold for regression prediction which performs best for mid-range survival period. This hybrid approach helps in defining a tier-structure for predictions that can prove to be a more refined approach for healthcare professionals and patients to better plan medical resources, treatment and care strategy at an individual level. Also, the combined approach helps to define a threshold for high survival periods, where the performance of both models is lower. This threshold should define the maximum period for which the predictions can be made accurately either through a classification or regression approach.

5.5 Feature Importance

Feature importance was estimated at both the local and global prediction levels. The feature importance varied across multiple classes. Age at the time of diagnosis (*AGE_DX*) was an important driving factor in prediction, with higher value resulting in a low survival period prediction and lower age driving higher survival period. Also, the number of malignant comorbidity tumors (*MALIGCOUNT*) was another driving factor, especially for low survival period predictions with higher count leading to more confidence in the prediction of low survival period. Similarly, the number of benign tumors (*BENBORDCOUNT*) drives low survival period prediction, with a lower number of benign tumors leading to low survival period prediction. The tumor size (*CSTUMSIZ*) and extension (*CSEXTEN*), were important features for prediction of higher survival period, with lower tumor and lesser extension leading to higher survival period prediction. The number of radiation rounds (*RADIATNR*) represented another feature contributing highly to high survival period prediction, with an increasing number of rounds of radiations showing higher prediction contribution to higher survival period prediction.

Another important factor contributing to low survival period prediction is the grade of the cancer (*GRADE*), with a higher grade driving lower survival period prediction. Laterality of tumor (*LATERAL*) is an important factor for high survival period prediction, with higher laterality pushing for a higher survival period prediction. Derived AJCC T value (*DAJCCT*) and N value (*DJACCN*) provide another facet of higher survival period prediction with higher T & N values contributing to higher survival period prediction. A lower number of lymph nodes involved ((*CSLYMPHN*)) leads to higher survival period prediction.

Also, some features such as progesterone status (*PRSTATUS*), primary site for surgery (*SURGSITF*), first primary cancer (*FIRSTPRIM*), marital status (*MAR_STAT*) and race (*RACEIV*) are driving predictions at the local level, but the dependence is not seen at the global level. This indicates that these features are important to the model, but they drive the prediction through interaction with other variables, which cannot be observed through Partial Dependence Plots. Hence, these features need more scrutiny to understand their interactions with other variables and the effect of these interactions on the final prediction.

Finally, there are a few variables that are considered important by the models, but they represent information that is already convened through other variables. Hence, even if they are considered important to the predictions, they still represent information already conveyed through other variables. These variables include year of birth (*YR_BRTH*), age recode (*Age_1_REC*), adjusted AJCC stage variable (*ADJAJCCSTG*), administration of chemotherapy (*CHEMO_RX_REC*) and primary site (*PRIMSITE*).

6 Conclusions

This paper aims at providing new ways to predict breast cancer survival period by applying deep learning methods. We were able to model the survival period for breast cancer patients using ANN, RNN, and CNN and compare those models with baseline models consisting of linear regression, random forests, gradient boosting machine, support vector machine, and naïve Bayes model. We achieved 87.50% accuracy for the classification approach with the best-performing ANN model and 13.62% Root Mean Squared Error (RMSE) and 76.37% R^2 value for the regression approach with the best-performing CNN model. The deep learning models significantly outperformed the traditional machine learning baseline models for both the classification and regression approaches.

The performance of classification model was shown to be decreasing as the survival period in the class increases. This leads to the best performance on '<=5 years' class and lowest performance on '>10 years' class. In the regression approach, it was shown that the prediction error was the lowest for mid-range survival period and high for both low survival period and high survival period cases. Hence, a threshold of acceptable error specific to a cancer needs to be defined to estimate the maximum threshold of prediction to receive acceptable values and combine predictions with the classification models to improve the usability of prediction on low survival period cases.

Note that the performance of deep learning models does not differ much between models, reinforcing the selection of model architecture based on intrinsic characteristics of data. Adoption of hybrid architecture with different architectures in different layers could improve the models' performance, leading to an increased ability to model more complex decision boundaries, but with decreasing interpretability.

Finally, we analyzed feature importance at both the individual prediction level and the model as a whole. Important features were identified that drive both the low survival period prediction and high survival period prediction, along with the features that represent information through interactions with other features.

This study constitutes a proof-of-concept that our methodology can serve as a basis for developing models that will support healthcare professionals and patients plan medical resources, treatment and care strategy with just information collected early in the patient's cancer journey.

Acknowledgement. The authors of this work would like to acknowledge the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC award #1624727 and in part by Susan G. Komen Foundation for funding this research. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

Data Availability. The SEER cancer registry is made available through the NCI and the process to access the data along with the documentation is provided at https://seer.cancer.gov/data/.

References

1. National Cancer Institute (NCI): 'Cancer Stat Facts: Cancer of Any Site' (2019). https://seer.cancer.gov/statfacts/html/all.html
2. National Cancer Institute (NCI): 'SEER Cancer Stat Facts: Breast Cancer' (2019). https://seer.cancer.gov/statfacts/html/breast.html
3. Susan G. Komen: 'Breast Cancer Statistics' (2020). https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/
4. National Cancer Institute (NCI), National Institutes of Health (NIH): 'Cancer Statistics' (2019), Available: https://www.cancer.gov/about-cancer/understanding/statistics
5. Luke, M., et al.: Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro-Oncol. 18(3), 417–425 (2015)
6. Mingjie, Q., Pathak, J., Pereira, N.L., Zhai, C.: Temporal reflected logistic regression for probabilistic heart failure survival score prediction. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 410–416. IEEE (2017)
7. Marshall, A.H., Hill, L.A., Kee, F.: Continuous dynamic bayesian networks for predicting survival of ischaemic heart disease patients. In: 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 178–183. IEEE (2010)
8. Zhang, H., Hung, C.-L., Chu, W.C.-C., Chiu, P.-F., Tang, C.Y.; Chronic kidney disease survival prediction with artificial neural networks. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1351–1356. IEEE (2018)
9. Bellot, A., Van der Schaar, M.: A hierarchical bayesian model for personalized survival predictions. IEEE J. Biomed. Health Inform. 23(1), 72–80 (2018)
10. Gray, E., Marti, J., Brewster, D.H., Wyatt, J.C., Hall, P.S.: Independent validation of the PREDICT breast cancer prognosis prediction tool in 45,789 patients using Scottish cancer registry data. Br. J. Cancer **119**(7), 808–814 (2018)
11. Song, K., et al.: Can a nomogram help to predict the overall and cancer-specific survival of patients with chondrosarcoma? Clin. Orthop. Relat. Res. **476**(5), 987 (2018)
12. Lynch Chip, M., et al.: Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int. J. Med. Inform. 108, 1–8 (2017)
13. Said, A.A., Abd-Elmegid, L.A., Kholeif, S., Abdelsamie Gaber, A.: Classification based on clustering model for predicting main outcomes of breast cancer using hyper-parameters optimization. Int. J. Adv. Comput. Sci. Appl. 9(12), 268–273 (2018)
14. Bartholomai, J.A., Frieboes, H.B.; Lung cancer survival prediction via machine learning regression, classification, and statistical techniques. In: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 632–637. IEEE (2018)
15. Hegselmann, S., Gruelich, L., Varghese, J., Dugas, M.: Reproducible survival prediction with SEER cancer data. In: Machine Learning for Healthcare Conference, pp. 49–66 (2018)
16. Naghizadeh, M., Habibi, N.: A model to predict the survivability of cancer comorbidity through ensemble learning approach. Expert Syst. **36**(3), e12392 (2019). Agrawal, Ankit, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. "A lung cancer outcome calculator using ensemble data mining on SEER data." In Proceedings of the tenth international workshop on data mining in bioinformatics, pp. 1–9. 2011
17. Dai, D., Jin, H., Wang, X.: Nomogram for predicting survival in triple-negative breast cancer patients with histology of infiltrating duct carcinoma: a population-based study. Am. J. Cancer Res. **8**(8), 1576 (2018)

18. Imani, F., Chen, R., Tucker, C., Yang, H.: Random forest modeling for survival analysis of cancer recurrences. In: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), pp. 399–404. IEEE (2019)
19. Kleinlein, R., Riano, D.: Persistence of data-driven knowledge to predict breast cancer survival. Int. J. Med. Inform. **129**, 303–311 (2019)
20. Shukla, N., Hagenbuchner, M., Win, K.T., Yang, J.: Breast cancer data analysis for survivability studies and prediction. Comput. Methods Program. Biomed. 155, 199–208 (2018)
21. SEER Program, National Cancer Institute (NCI): 'SEER Incidence Data, 1975–2017' (2019), Available: https://seer.cancer.gov/data/
22. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**(4), 115–133 (1943)
23. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science (1985)
24. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
25. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (1980)
26. Lundberg, S.M., Lee, S.-L.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
28. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3145–3153 (2017)
29. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2013)
30. Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Appl. Stoch. Model. Bus. Ind. **17**(4), 319–330 (2001)
31. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617. IEEE (2016)
32. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS ONE **10**(7), e0130140 (2015)
33. Friedman, J.H.:Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
34. Travis, E.O.: A guide to NumPy. Trelgol Publ (2006)
35. McKinney, W.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56 (2010)
36. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
37. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283 (2016)
38. Chollet, F., et al.: Keras (2015). https://github.com/fchollet/keras
39. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007)
40. Harsha, N., Jenkins, S., Koch, P., Caruana, R.: Interpretml: a unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)