Lifespan Prediction for Lung and Bronchus Cancer Patients via Machine Learning Techniques

Rouzbeh Talebi Zarinkamar and Rene V. Mayorga

***Abstract*—Patients' accurate survival predictions can influence treatment planning and costs, particularly lung cancer, which is one of the leading causes of cancer-related death. Machine Learning (ML) techniques are powerful in increasing the accuracy of such predictions. However, only a few studies have used an ML approach for actual lifespan prediction for cancer patients using the Surveillance, Epidemiology, and End Results (SEER) program database. This study intends to apply several well-known ML models, namely, a developed Deep Neural Networks (DNN), Linear Regression, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forest (RF), and Adaboost, to predict the actual survival time on a monthly basis for lung cancer patients. The results indicate that the models give better performance for low to average survival times (0 to 25 months) that make up the majority of the data. The best model was the developed DNN with a Root Mean Square Error (RMSE) value of 12.672. In contrast, the Adaboost model was the worst-performing technique since it had weak discrete power for the data.**

***Index Terms*—Machine learning, lung and bronchus cancer, survival prediction, performance evaluation.**

I. Introduction

Lung cancer is in the top three common cancer types with a high mortality rate, causing about one-fifth of malignancies in men and one-ninth in women [1]. Moreover, based on the National Cancer Institute (NCI) in the USA, almost 23% of all cancer deaths in 2020 belongs to lung and bronchus cancer.

Survival time estimation is crucial in assessing patients' prognoses. However, it is a severe challenge for clinicians. Clement-Duchene *et al.* [2] claimed that clinicians predicted patients' survival time approximately 21.5 months, but in reality, they survived almost 11.7 months on average. Muers *et al.* [3] discovered that only 10% of physicians' prediction is accurate when they predict survival time less than a month, and 59% and 71% of estimations are correct for the survival time prediction within three months and four months, respectively.

The accuracy of survival time estimation is crucial and could affect the treatment costs, treatment decision making, treatment planning, and therapeutic results [4], [5]. The application of Machine Learning (ML) techniques is helpful to increase the accuracy of such predictions [6]. ML can use historical medical data and enable the computers to discover the outcomes by exploring the relationships and patterns among variables [7]. Therefore, having enough data is crucial for ML models to learn and perform precisely. However, having access to patients' historical data is challenging. One exception is the largest publicly available database called the Surveillance, Epidemiology, and End Results (SEER) program from the National Cancer Institute (NCI) at the National Institutes of Health (NIH) in the USA.

Most of the existing studies have used ML classification techniques on SEER datasets to predict whether a patient would survive after a specific number of years from diagnosis [8]–[10]. In this case, the binary of survival and death prediction is not precise enough to aid treatment planning [11]. On the other hand, predicting actual survival time can support more appropriate decision making for both clinicians and involved families [12]. However, this area is less examined to date.

This study tends to answer the question of whether actual survival prediction by ML techniques can have better precision. Lung and bronchus cancer patients diagnosed during the years 2004–2011 were selected from the SEER cancer database.

Related Work

In the 1950s, the TNM staging system (stage T, N, and M) was introduced to predict cancer patients' survivability [13]. Then, in 1994, Burke [14] used statistical techniques and proved that the performance of those techniques gives better results than the TNM staging system.

ML techniques were getting more popular, and the subsequent studies applied those techniques for the comparison of ML against statistical models. Ali *et al.* [15] and Delen [16] proved that ML techniques are more accurate than statistical models.

In recent years, ML techniques have been widely used for cancer prognosis using SEER datasets. Lynch *et al.* [7] evaluated several ML techniques to predict the actual survival time of lung cancer patients, including linear regression, Decision Trees (Dt), Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and a custom ensemble. Wen *et al.* [17] developed a Neural Network (NN) model along with several ML techniques such as k Nearest Neighbors (KNN), DT, Na ïve Bayes (NB), and SVM to predict the survivability of the patients who have prostate cancer.

Because each ML technique has different performances in the same case, the ensemble methods offer a better result by

Manuscript received October 12, 2021; revised April 26, 2022. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was also supported by a grant from the Saskatchewan Centre for Patient-Oriented Research (SCPOR), Saskatchewan, Canada.

The authors are with the Faculty of Engineering and Applied Science, Department of Industrial Systems Engineering, University of Regina, Canada (e-mail: Rouzbeh.Talebi@gmail.com, Rene.Mayorga@uregina.ca).

doi: 10.18178/ijmlc.2022.12.6.1107

combining multiple ML models. Edeki *et al.* [18] proposed Random Forest, a Bagging ensemble learner method, with several classification models for breast cancer prediction and declared that Random Forest achieved the best performance. Wang *et al.* [11] used tree ensemble-based models for colorectal cancer. They used a two-stage model, whereas, in the first stage, the models predict whether the patients survive after five years from diagnosis or not. In the next step, the ML regression models predict the actual survival time of the patients who were predicted to die. Recently, Deep Artificial Neural Network has been proved as a robust ML technique in addition to the previously mentioned ML models. Song *et al.* [19] used Deep Learning, then compared it to the Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF) to predict the survival of pancreatic neuroendocrine tumors (PNETs). Han *et al*. [20] developed a novel Deep-Learning-based prediction for synovial sarcoma cancer. Wen *et al.* [17] applied a Deep Artificial Neural Network and several ML models to classify survivability and mortality for the SEER prostate cancer database. The experimental results showed that Deep Neural Networks had the best performance among the algorithms. Al-Bahrani *et al.* [21] developed a deep neural network prediction model for SEER colon cancer patients. Due to its ability to learn non-linear structures, Deep Learning has become a powerful ML method with many advantages.

Some studies applied various ML models on lung cancer diagnosis, recognition and prediction using different datasets such as X-ray images [6], [22], [23]. For example, Bharati *et al.* [22] used a hybrid deep learning model using X-ray images to detect lung diseases. Bharati *et al.* [6] developed an ML approach using different ML classification models to predict lung cancer disease. On the other hand, only a few studies have used ML techniques to predict lung cancer patients' survival prediction using the SEER database.

II. Materials and Methods

For this study, we chose Deep Neural Networks (DNN), Support Vector Machine (SVM), Linear Regression, Random Forest, Gradient Boosting Machine (GBM), and Adaboost model. There are many ML techniques that exist, but we chose the most common methods used for regression. Furthermore, these models use algorithms that can achieve higher performance than many other algorithms because of their ability to learn non-linear and complex relationships.

For models' evaluation and comparison, Root Mean Square Error (RMSE) is selected as an indicator. This metric is so famous for regression model evaluation, and it provides an excellent estimation of the model accuracy [7]. The Root Mean Square Error is the sample standard deviation of the differences between observed and foretold values. Equation 1 shows how the RMSE is calculated.

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

Symbols *n*, *y*, ˆ*y* denote the number of predicted samples, actual value, and predicted value, respectively.

Random sampling for the training and testing datasets may cause the ML models to generalize with bias. The K-fold cross-validation is an efficient method to avoid such an over-fitting problem. In our study, we utilize 10-fold cross-validation for all ML models.

Fig. 1 shows the whole process from data preparation to the comparison of all models.

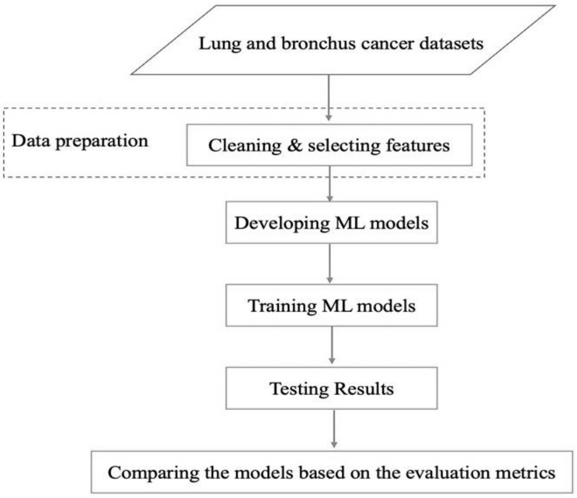
Lung and bronchus cancer datasets
Data preparation
Cleaning & selecting features
Developing ML models
Training ML models
Testing Results
Comparing the models based on the evaluation metrics

Fig. 1. The flowchart of whole process.

We implement the models in Python version 3.7, an open-source programming language. The implementation is carried out on a laptop with a 2.6 GHz 6-core 9th-generation Intel Core i7 processor and 16 GB 2400 MHz DDR4 memory.

A. Deep Neural Networks

The human brain, because of billions of nerve cells, can learn complicated tasks much better than computers. The Idea of Artificial Neural Networks (ANN) using a mathematical model came from those nerves to perform almost the same as the human brain [24]. An ANN model is constructed by three base layers, including the input layer, hidden layer, and output layer. The hidden layer plays a significant role in discovering the relationship between the input layer and the output layer. The complex form of ANN is called Deep Neural Network (DNN) that has two or more hidden layers [25]. It is capable of performing tasks with high complexity. In this study, we build a DNN model in Keras Sequential Model API in Keras Python library. The proposed DNN for classification tasks has an input layer with 100 neurons and linear activation function, three hidden layers (100, 200, and 100 neurons for each hidden layer, respectively) with linear activation function and an output layer with a sigmoid activation function. The specific hyperparameters (architecture) of the DNN algorithm, i.e., the length of neurons, the number of layers, and the activation function, are achieved manually through many experiments from the given dataset. Before training, the model needs to configure the learning process with a compiled method. To compile it, an optimizer and loss function is 'Root Mean Square Propagation' and 'Mean Squared Error', respectively. After creating the regression model, we fit the model with given datasets to predict the lifespans.

B. Support Vector Machine

Support Vector Machine (SVM) is a commonly used supervised ML technique for both classification and regression tasks [26]. This technique aims to build a hyperplane with maximum margin in multi-dimensional space to classify the targets precisely [27]. The hyperplane could be a point, a line, and space in one, two, and three-dimensional spaces. When the algorithm is not able to draw a line for separation, a kernel method is applied to separate non-linearly. When the data is noisy, the SVM model could suffer from over-fitting. We use a grid search function with 10-fold cross-validation in Python to find the best hyper-parameters with different values of C (0.1, 1, 100) and kernels (RBF and linear).

C. Linear Regression

One of the simplest and widely used correlational techniques is Linear Regression. The Linear Regression algorithm's goal is to find the relationship between the dependent and independent continuous variables. The core idea of the model is to fit the best line to minimize the prediction errors that come from the difference between the predicted points and the line [28]. A drawback of Linear Regression is that the linear line does not work for real data [7]. We use the default Linear Regression parameters in Python.

D. Random Forest

Random forest (RF) is a supervised ML technique using ensemble methods. It makes a forest by decision Tree (DT) models as a base learner and combines and develops them to achieve high-performance results [29]. Each DT model is feed by a bootstrapped sample. Then, some DT models are selected randomly to calculate the mean of the results. This process continues until the best results are obtained [30]. In this study, a grid search with 10-fold cross-validation was used to find the best parameters. Maximum depth, the minimum number of samples split of trees, and the number of trees in the forest were tested with the values of {110, 120, 130, 140, 150}, {3, 4, 5} and {50, 100}, respectively.

E. Adaboost

The AdaBoost (Adaptive Boosting) ensemble is a sequential process model that brings multiple weak learners together (usually decision trees) to build a robust predictor [31]. The models are built sequentially using the training set. Each model's responsibility is to reduce the previous model's error, and this process continues until the minimum error is obtained. In the first step, all the models are allocated with the same weights. Then, in each iteration, the algorithm increases and decreases the weight of the models that predicted inaccurately and accurately, respectively [32], [33]. Consequently, the final ensemble model uses the summation of weights obtained by previous models to predict.

F. Gradient Boosting Machine

Gradient Boosting Machine (GBM) is another form of ensemble method using multiple small weak learners and brings them together into a precise result. It is a widely used technique due to its effectiveness in calculating complex samples. This ability came from how the GBM algorithms identify the deficiency of weak learners by using gradients in the loss function [7]. The prediction is concluded by the summation of all weak learners' predictions. To tune the hyper-parameters, a grid search along with 10-fold cross-validation was used. The number of trees was kept at 100, and the minimum observation was 50. The depth was tested at {1,2,3,5, and 10}, and shrinkage between {0.01, 0.1, 0.2, and 0.5}.

TABLE I. SELECTED FEATURES AFTER DATA PROCESSING FOR LUNG AND BRONCHUS CANCER

No.	Variable	Description
1	Tumor_Size	Measurement of tumor size.
2	Diag_Confirmation	The method used to confirm the presence of the cancer being reported.
3	Reg_nodes_exam	Records the total number of regional lymph nodes that were removed and examined by the pathologist.
4	Surgery_Prime_Site	Lung surgery to the prime site
5	Grade	The appearance of cancer cells and how fast they may grow.
6	CS_REG_NODE	This item reflects the validity of the classification of the item CS Lymph Nodes only according to diagnostic methods employed.
7	Sex	Gender of the patients
8	Primary_Site	Location of the tumor within the lungs.
9	Laterality	Laterality describes the side of a paired organ or side of the body on which the reportable tumor originated.
10	Histologic_Type	Describes the microscopic composition of cells and/or tissue for a specific primary.
11	CS_LYMPH_NODE	This item records the highest specific lymph node chain that is involved by the tumor.
12	Race	Race recode is based on the race variables and the American Indian/Native American IHS link variable.
13	Surg_Reg_Dis	The surgical procedure of Other Site describes the surgical removal of distant lymph node(s) or other tissue(s) or organ(s) beyond the primary site.
14	Scope_Reg_LN_Sur	(Scope of Regional Lymph Node Surgery) The procedure of removal, biopsy, or aspiration of regional lymph nodes.
15	Reason_NO_Surg	The reason that surgery was not performed
16	Radiation	Indication of whether the patient has received radiation.
17	Radiation_Sequence	Order of surgery and radiation therapy administered for patients who received both.
18	Sequence	Order of lung cancer occurrence with respect to other cancers for this patient.
19	Historic_Stage	It is a simplified version of the stage: in situ, localized, regional, distant, & unknown.
20	T	AJCC component describing tumor size.
21	N	AJCC component describing lymph node involvement.
22	NUM_INSITU	Count of a patient's total reported in situ/malignant cancers
23	NUM_BENIGN	Count of a patient's total reported benign/borderline cancers.
24	Age	Age at the time of diagnosis.
25	Year	Year of diagnosis
26	CS_METS	Information on distant metastasis.

III. Results

A. Selection of Patient Attributes

This study acquired lung cancer patients' data diagnosed from 2004-2011 in the SEER database. This study received the lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI), which is the largest publicly available cancer database (http://www.seer.cancer.gov). The data preprocessing is crucial and should be carried out for data mining algorithms. Several attributes were excluded due to value repetition and null values. Thus, based on the literature review and experts, 26 attributes (shown in Table I).

The description presented in Table I directly came from" SEER Research Data Record Description" [34]. Then, a One-Hot Encoding method is applied for nominal variables. Based on Lynch *et al*., (2017) [7] work, the survival range is considered from 0 to 72.

Fig. 2 indicates the distribution of survival time. After data preprocessing, 18013 samples are retained from the SEER database.

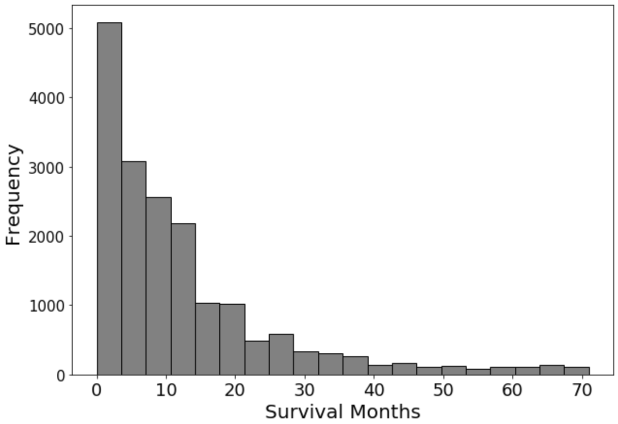

Fig. 2. The range of survival time for patients (2004-2011).

B. Prediction of Survival Times

We applied the regression-based ML techniques to the given data, and their respective performance is measured. Table II shows the metrics, including RMSE, the mean and standard deviation of predicted values, and standard deviation of residual between actual and predicted values and coefficient of determination. It is essential to mention that the mean and standard deviation of the dataset are 12.49 and 13.82, respectively.

TABLE II. Comparison of Results via Different Metrics

Models	RMSE	Standard Deviation	Standard Deviation of Residuals	Mean
DNN	12.672	3.382	12.664	13.837
SVM	13.606	2.16	13.608	8.740
Linear Regression	12.744	5.311	5.311	5.311
Random Forest	13.816	7.587	13.818	13.837
Adaboost	13.160	4.702	13.163	12.7
GBM	12.704	5.069	12.706	12.54

Table II presents that the developed DNN gives the best results among all ML models with the RMSE value of 12.672. The standard deviation of predicted values gives further insight besides the RMSE value. The SVM model has the lowest standard deviation of predicted values with a score of 2.16. In contrast, Random Forest gives the highest deviation with the value of 7.587. Furthermore, the standard deviation of the residuals shows that how accurate the models perform for all the samples. Surprisingly, simple Linear Regression outperforms with the values of 5.31.

A scatter plot can give further insight into the standard deviation for the residuals. One application of the scatter plot shows the correlation between predicted and actual values and how the predicted values fit the actual values. Fig. 3 indicates the scatter plots for all six different regression-based ML models. It shows predicted values fit well with the actual values for low (almost 0 months) to moderate values (almost 25 months). However, the predicted values for the values past the 25 months do not fit well with the actual values.

IV. Discussion

Six ML models were applied for lung and bronchus cancer prediction, namely, Deep Neural Network (DNN), Support Vector Machine (SVM), Linear Regression, Random Forest, Adaboost, and Gradient Boosting Machine (GBM). As presented in Table II, the DNN model was the most robust model with an RMSE value of 12.672. GBM and Linear Regression followed this with the values of 12.704 and 12.744, respectively.

Trailing behind the other models were Random Forest, SVM, and Adaboost with RMSE values of 13.816, 13.606, and 13.160, respectively. Surprisingly, the simple Linear Regression model outperformed some advanced techniques, including Random Forest, SVM, and Adaboost. Adaboost could not find sufficient discrete branching or splitting points due to the size of the data and its vague nature.

The reason for the high standard deviation of residuals with a value of almost 13 is 60% of testing patients' data survive less than 13 months. This value is higher than the lifespan of 60% of the population. This significant deviation could happen from the longer survival months, which makes a prediction so tricky. In contrast, for the patients with a survival time of fewer than 30 months in the dataset set, the RMSE value for the DNN model is 6.5 months.

The performances of the models in this study are not directly comparable to most previous works since the cancer types and the dataset are different. Most of them built classification models to predict categorical survival times rather than applying the regression models to analyze the continuous spectrum of survival time. Therefore, their metric indicators, such as accuracy, sensitivity, and specificity, cannot be compared by the regression results presented in this study. Lynch *et al.* [7] used several ML models for actual survival time prediction for lung cancer patients using different datasets. Our results show that the RMSE with the value of 12.672 is better than the value of 15.30 given in [6].

Hopefully, the Deep Neural Network (DNN) can improve by developing the power of computation. However, the performance of the techniques applied here has not outperformed dramatically. It could have different reasons, such as data limitations. The survival prediction needs to have some crucial lung cancer features, including ALK, EGFR, PD-L1, and ROS1 [35], [36], which were not provided by the SEER database. Also, the patient's lifestyle (smoking or not), economic situation, and patient's mental conditions (optimistic or pessimistic) could be highly effective in the patient's lifespan. Besides, our lung and bronchus cancer datasets suffer from a considerable amount of missing data.

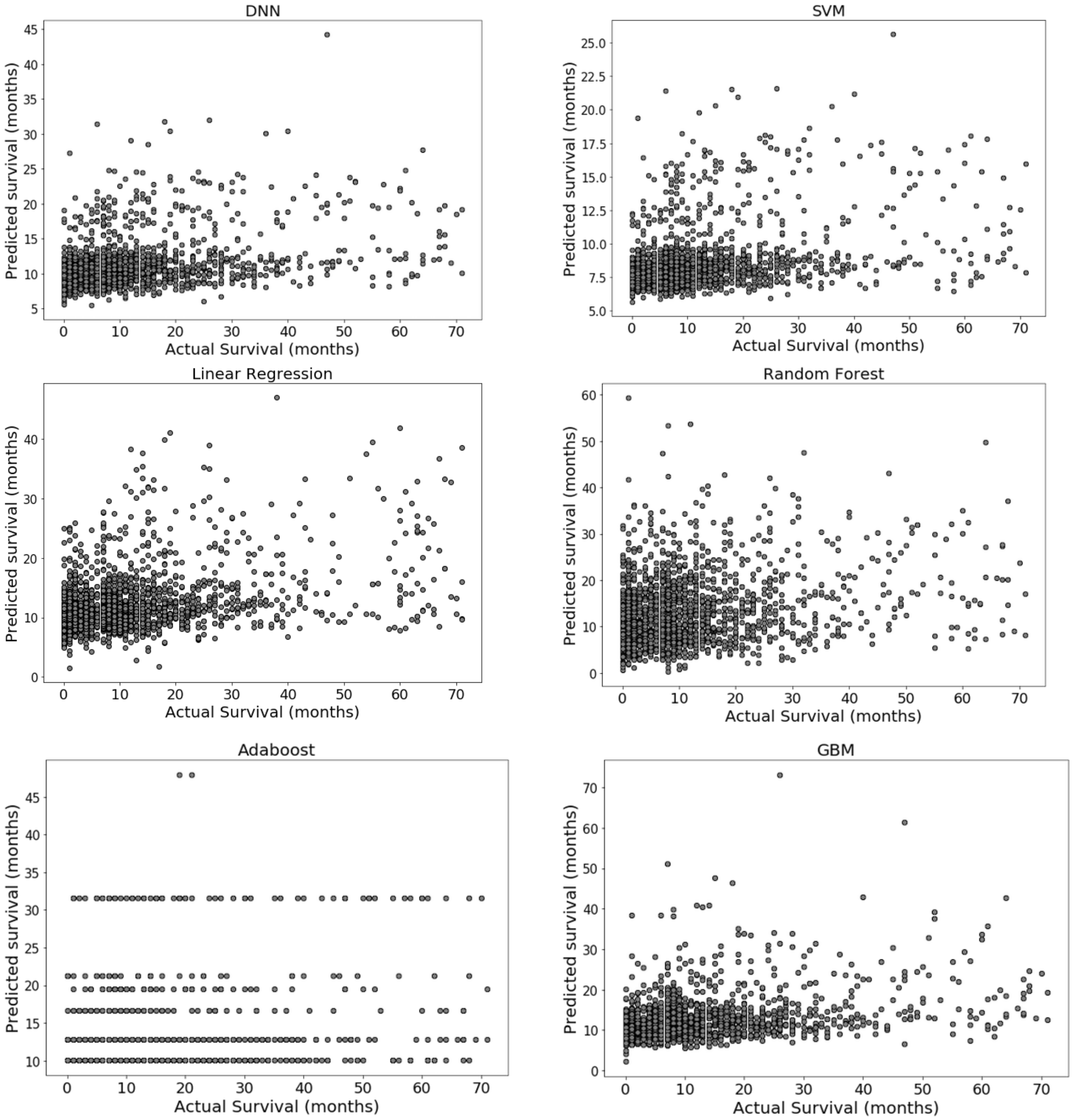

Fig. 3. Scatter plots between predicted values and actual values for all models.

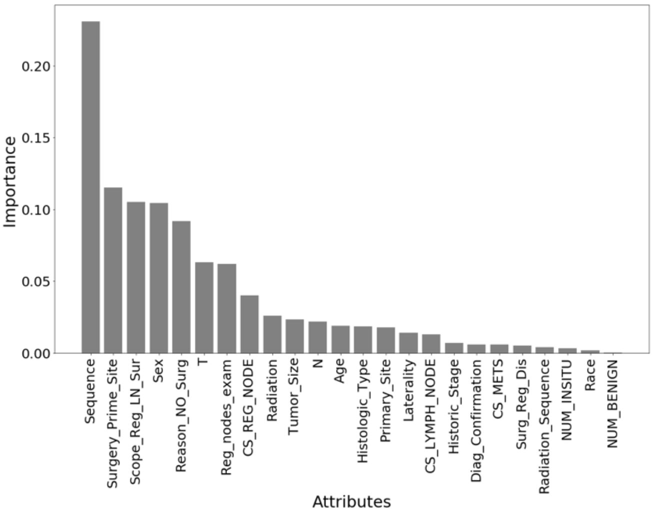

Fig. 4. Sorting the importance of the attributes in survivability prediction.

In addition to analyzing the ML models' performance evaluation, it is valuable to discover the essential factors for lung and bronchus cancer survival prediction. Different features can have a different degree of importance based on their prediction strength. This study used the Random Forest model to rank the futures by their predicted power values. Fig. 4 illustrates the sorting of 26 features based on their prediction power for survivability prediction. Among the features, 'Sequence', 'Surgery Primary Site', 'Scope Reg LN Sur', 'Sex' and 'Reasopn NO Surg' are the most significant features. On the other hand, 'NUM BENIGN', 'Race', 'NUM INSITU', 'Radiation Sequence', 'Surg Reg Dis' and 'CS METS' have less importance among all variables.

V. Conclusions and Future Work

Precise survival prediction in cancer prognosis is critical but essential, as it affects treatment decision making and planning. This study developed several popular ML models, namely, Deep Neural Networks (DNN), Linear Regression, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forest (RF), and Adaboost, to predict the actual survival time on monthly basis for lung cancer patients. The results indicate that the models give better performance for the patients whose survival time is between 0 and 25 months, making up the majority of the data. In addition, the DNN model outperformed other ML models with a Root Mean Square Error (RMSE) value of 12.672. In contrast, the Adaboost model gives the lowest performance since it had weak discrete power for the data.

The future studies would focus on the following problems. They could contain complete cancer information that is not available in the latest version of the SEER dataset. Furthermore, future studies could apply to other high mortality cancers such as prostate and breast cancer. Finally, it would be interesting to use images related to cancer therapy records for survival prediction.

Acknowledgments

The authors would like to express their gratitude to Dr. Osama Souied from the Allan Blair Cancer Centre, at the University of Saskatchewan, for helpful advice on features selection for cancer survival. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was also supported by a grant from the Saskatchewan Centre for Patient-Oriented Research (SCPOR), Saskatchewan, Canada.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contribution

Rouzbeh Talebizarinkamar does the works on conceptualization, formal analysis, methodology, investigation, writing-original draft. Dr. Rene V. Mayorga does the works on Formal analysis, methodology, supervision, paper review & editing, funding acquisition.

References

[1] National Cancer Institute. (2019). *SEER Training Modules, Introduction to Lung Cancer.* [Online]. Available: https://training.seer.cancer.gov/lung/intro/

[2] C. Cl ément-Duchêne, C. Carnin, F. Guillemin, and Y. Martinet, "How accurate are physicians in the prediction of patient survival in advanced lung cancer?" *Oncologist*, vol. 15, no. 7, pp. 782–789, 2010.

[3] M. F. Muers, P. Shevlin, and J. Brown, "Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model," *Thorax*, vol. 51, no. 9, pp. 894–902, 1996.

[4] J. Jiang *et al.*, "Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm," *Scientific Reports*, 2019.

[5] P. Glare, C. Sinclair, M. Downing, P. Stone, M. Maltoni, and A. Vigano, "Predicting survival in patients with advanced disease," *Eur. J. Cancer*, vol. 44, no. 8, pp. 1146–1156, May 2008.

[6] S. Bharati, P. Podder, R. Mondal, A. Mahmood, and M. Raihan-Al-Masud, "Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer," *Adv. Intell. Syst. Comput.*, vol. 941, pp. 447–457, Dec. 2018.

[7] C. M. Lynch *et al.*, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *Int. J. Med. Inform.*, vol. 108, pp. 1–8, 2017.

[8] A. V. Karhade *et al.*, "Development of machine learning algorithms for prediction of 5-Year Spinal Chordoma Survival," *World Neurosurg.*, vol. 119, pp. e842–e847, 2018.

[9] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005.

[10] Q. C. B. S. Thio *et al.*, "Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma?" *Clin. Orthop. Relat. Res.*, vol. 476, no. 10, pp. 2040–2048, 2018.

[11] Y. Wang, D. Wang, X. Ye, Y. Yuyan, Y. Yin, and Y. Jin, "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction," *Inf. Sci. (Ny).*, vol. 474, pp. 106–124, 2019.

[12] R. Talebizarinkamar, "A machine learning approach for lung and bronchus cancer survival prediction," Master Thesis, Department of Industrial Systems Engineering, University of Regina, SK, Canada, 2020.

[13] P. Gao *et al.*, "Is the prediction of prognosis not improved by the seventh edition of the TNM classification for colorectal cancer? Analysis of the surveilla006Ece, epidemiology, and end results (SEER) database," *BMC Cancer*, vol. 13, pp. 1–6, 2013.

[14] H. B. Burke, "Artificial neural networks for cancer research: outcome prediction," in *Seminars in Surgical Oncology*, 1994, vol. 10, no. 1, pp. 73–79.

[15] A. Ali, A. Tufail, U. Khan, and M. Kim, "A survey of prediction models for breast cancer survivability," *ACM Int. Conf. Proceeding Ser.*, vol. 403, pp. 1259–1262, 2009.

[16] D. Delen, "Analysis of cancer data: A data mining approach," *Expert Syst.*, vol. 26, no. 1, pp. 100–112, 2009.

[17] H. Wen, S. Li, W. Li, J. Li, and C. Yin, "Comparision of four machine learning techniques for the prediction of prostate cancer survivability," in *Proc. 2018 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2018*, 2019, pp. 112–116.

[18] C. Edeki and S. Pandya, "Comparison of data mining techniques used to predict cancer survivability," *IJCSIS 2012*, vol. 10, no. 6, 2012.

[19] Y. Song, S. Gao, W. Tan, Z. Qiu, H. Zhou, and Y. Zhao, "Multiple machine learnings revealed similar predictive accuracy for prognosis of PNETs from the surveillance, epidemiology, and end result database," *J. Cancer*, vol. 9, no. 21, pp. 3971–3978, 2018.

[20] I. Han, J. H. Kim, H. Park, H. S. Kim, and S. W. Seo, "Deep learning approach for survival prediction for patients with synovial sarcoma," *Tumor Biol.*, vol. 40, no. 9, pp. 1–9, 2018.

[21] R. Al-Bahrani, A. Agrawal, and A. Choudhary, "Survivability prediction of colon cancer patients using neural networks," *Health Informatics J.*, vol. 25, no. 3, pp. 878–891, 2019.

[22] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Informatics Med. Unlocked*, vol. 20, p. 100391, Jan. 2020.

[23] S. Bharati, P. Podder, and P. K. Paul, "Lung cancer recognition and prediction according to random forest ensemble and RUSBoost algorithm using LIDC data," *Int. J. Hybrid Intell. Syst.*, vol. 15, no. 2, pp. 91–100, Jan. 2019.

[24] D. Anderson and G. McNeill, "Artificial neural networks technology," *Kaman Sci. Corp.*, vol. 258, no. 6, pp. 1–83, 1992.

[25] J. Prakash and P. K. Kankar, "Health prediction of hydraulic cooling circuit using deep neural network with ensemble feature ranking technique," *Measurement*, p. 107225, Nov. 2019.

[26] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, no. April, pp. 101–111, 2019.

[27] W. S. Noble, "What is a support vector machine?" *Nat. Biotechnol.*, vol. 24, no. 12, p. 1565, 2006.

[28] D. L. Naik and R. Kiran, "Naïve Bayes classifier, multivariate linear regression and experimental testing for classification and characterization of wheat straw based on mechanical properties," *Ind. Crops Prod.*, vol. 112, pp. 434–448, Feb. 2018.

[29] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *J. Clean. Prod.*, vol. 203, pp. 810–821, 2018.

[30] Y. Zhou and G. Qiu, "Random forest for label ranking," *Expert Syst. Appl.*, vol. 112, pp. 99–109, 2018.

[31] Y. Freund and R. E. Schapire, "Schapire R: Experiments with a new boosting algorithm," in *Proc. Thirteenth International Conference on ML*, 1996.

[32] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, Feb. 2010.

[33] N. Zheng and J. Xue, *Statistical Learning and Pattern Analysis for Image and Video Processing*, Springer Science & Business Media, 2009.

[34] National Cancer Institute. Seer research data record description. [Online]. Available: https://seer.cancer.gov/data-software/documentation/seerstat/nov2016/TextData.FileDescription.pdf

[35] D. Rangachari *et al.*, "Correlation between classic driver oncogene mutations in EGFR, ALK, or ROS1 and 22C3–PD-L1 ≥50% Expression in Lung Adenocarcinoma," *J. Thorac. Oncol.*, vol. 12, no. 5, pp. 878–883, May 2017.

[36] M. Mino-Kenudson, "Immunohistochemistry for predictive biomarkers in non-small cell lung cancer," *Translational Lung Cancer Research*, vol. 6, no. 5. AME Publishing Company, pp. 570–587, 01-Oct-2017.

Rouzbeh Talebi Zarinkamar received his MASc degree in industrial systems engineering from the University of Regina, Canada. He has contributed to the data science and machine learning field in recent years by realizing several research studies. During his master's program at the University of Regina, he has gained two years of experience in collecting, preprocessing and analyzing data. The Master's program also provided him with an opportunity to obtain extensive knowledge in applied mathematics and statistics as a backbone of his research. Throughout his research, he attained valuable experience in developing machine learning models, data extraction, analysis, and manipulation.

Rene V Mayorga is a professor in the Department of Industrial Systems Engineering, at University of Regina, Canada. His research activities are dedicated to the development of artificial/computational sapience (Wisdom) as new disciplines/fields, and to Intelligent / Sapient (Wise) Systems applied on diverse areas. Over the years he has been in the editorial board of several international journals. He was the Editor in Chief for *Applied Bionics and Biomechanics* from 2003 to 2016. He is the co-editor of *"Toward Artificial Sapience: Principles and Methods for Wise Systems"*, Springer 2008. He has published papers widely in scientific journals, international conferences proceedings, books, and monographs. Also, he has edited several international conference proceedings. Over the years he has also served in several occasions as general chair and program chair of several international conferences.