TWO STAGED PREDICTION OF GASTRIC CANCER PATIENT'S SURVIVAL VIA MACHINE LEARNING TECHNIQUES

Peng Liu[1], Liuwen Li[1], Chen Yu[2], and Shumin Fei[†1]

[1]Key Laboratory of Measurement and Control of CSE (Ministry of Education), School of Automation, Southeast University, Nanjing, Jiangsu 210000 China
[2]Department of Integrated TCM &Western Medicine Jiangsu Cancer Hospital, Nanjing, Jiangsu 210000 China

ABSTRACT

Cancer is one of the most common causes of death in the world, while gastric cancer has the highest incidence in Asia. Predicting gastric cancer patients' survivability can inform patients care decisions and help doctors prescribe personalized medicine. Classification techniques have been widely used to predict survivability of cancer patients. However, very few attention has been paid to patients who cannot survive. In this research, we consider survival prediction to be a two-staged problem. The first is to predict the patients' five-year survivability. If the patient's predicted outcome is death, the second stage predicts the remaining lifespan of the patient. Our research proposes a custom ensemble method which integrated multiple machine learning algorithms. It exhibits a significant predictive improvement in both stages of prediction, compared with the state-of-the-art machine learning techniques. The base machine learning techniques include Decision Trees, Random Forest, Adaboost, Gradient Boost Machine (GBM), Artificial Neural Network (ANN), and the most popular GBM framework--LightGBM. The model is comprehensively evaluated on open source cancer data provided by the Surveillance, Epidemiology, and End Results Program (SEER) in terms of accuracy, area under the curve, F-score, precision, recall rate, training and predicting time in the classification stage, and root mean squared error, mean absolute error, coefficient of determination (R^2) in the regression stage.

KEYWORDS

Gastric Cancer, Cancer Survival Prediction, Machine Learning, Ensemble Learning, SEER

1. INTRODUCTION

"Gastric Cancer" refers to a malignant tumor that originates from gastric mucosal epithelium. It is one of the most common cancers worldwide [1]. It used to be the most malignant type of cancer until 1980s, when Lung cancer became the most deadly neoplasm [2]. Gastric cancer is more likely to happen to people over 50 years old, and the ratio of men to women with gastric cancer is 2:1. Although the incidence has decreased, the number of new cases is increasing each year due to global population aging. Moreover, due to the changes in dietary structure and pressure from work, and for reasons that remain unclear, gastric cancer incidence among young people have grown [3]. In the foreseeable future, gastric cancer will remain one of the major causes of cancer related fatalities. With a high incidence and mortality rates, gastric cancer causes a large amount of medical expenditures and has been a heavy burden on patients' families. Prediction of gastric

Dhinaharan Nagamalai et al. (Eds): AIAP, ICCIoT, CNSA, SIGML, IT, ICBB, DMDB - 2020
pp. 105-116, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100308

cancer survival rate and remaining lifespan has become essential in cancer studies. It may provide advices for better clinical decisions treatment [4].

Survivability often refers to the possibility of a patient being alive after five years since the time of cancer diagnosis. It is an indicator in medical science commonly used for evaluation of treating effects. Most cancer survivability studies use five-year survivability as their predicting target. However this type of prediction may not provide enough information for the doctors to make better medical decisions. If a patient's survival prediction is negative, the actual survival time of the patient remains unclear. For high mortality cancers including gastric cancer, most patients would not survive after five years. Thus survival time prediction should be studied to provide more precise information for medical decision making. [5]

Cancer survivability prediction used to be challenging due to the lack of publically available large scale medical data. The Surveillance, Epidemiology, and End Results Program (SEER), is an open source database is an open-source database which provides de-identified, coded, and annotated information on cancer statistics of the United States [6, 7]. The scale of data is large enough to be analyzed. Machine Learning represents a group of methods, including decision trees [8], artificial neural networks (ANN)[9], random forests[10], Adaboost[11], and gradient boost machine[12]. These methods have been widely used to discover a function to represent the relationship between a group variables and an outcome, and have been widely applied to predict cancer patients' outcome. LightGBM is a novel gradient boost decision tree (GBDT) algorithm which outperforms current GBDT methods in computation speed and memory consumption without compensating its prediction performance [13].

The main contributions of this work are:

1) We consider the survavibility prediction problem to be two-staged: the first is to predict patient's five-year survivability, the second is to predict the remaining life span of patients whose first-stage outcome is 'death';
2) We propose a custom ensemble method which integrated six machine learning algorithms. It exhibits a significant predictive improvement in both stages of prediction;
3) We provide the output scatterplot to compare the outcomes and the correlation scatterplot to analyze relationships between base learners.

The proposed framework is the custom ensemble of six base learners. The weights of each base learner is determined using a loop. In the first stage, the ensemble's prediction accuracy is 85% comparing to the maximum base learner's accuracy 84% and the average base learners' accuracy 82.9%. In the regression stage, the ensemble's root mean squared error (RMSE) is 11.6 compared to the best base learner's RMSE 11.67 and the average RMSE 12.05.

The remainder of this article is organized as follows. Section 2 introduces the related work on the application of machine learning algorithms to cancer survivability prediction. Section 3 introduces the base machine learning methods we used and compared with. Section 4 provides the detailed two-stage model we used, and the details of the experimental procedures. Section 5 presents the experimental results while section 6 presents the discussion of the results. Section 7 concludes the paper and presents possible future researches.

2. RELATED WORKS

Machine learning techniques has been widely applied to predict outcomes for medical purposes[8-12]. Ensemble learning methods that train a number of weak base learners and then combine their outputs are popular in medical prediction researches [22]. Many researchers conducted their researches on cases collected from SEER database [5-6, 14-22].

In [14], researchers compared four machine learning techniques' performance on survivability prediction of prostate cancer. Neural network out performed decision tree, Naïve Bayes, and support vector machine learning. Some scholars carried out survival month's prediction using various machine learning techniques including linear regression, decision trees, gradient boosting machines, support vector machines, and a custom ensemble approach [6]. The best performing method was custom ensemble and the most influential method was gradient boosting machine. In [15] a Gaussian k-base NB Classifier system was proposed to enhance classification accuracy, comparing to Naïve Bayes classifier and linear regression algorithm. Researchers obtained their dataset from UCI repository and SEER database. They proposed an online gradient boosting learning with adaptive linear regressor and compared its performance with state of the art machine learning algorithms[15]. Scholars conducted a research on cancer comorbidity survival prediction including breast and genital in women, and prostate cancer comorbidity in men. Gradient boosting, random forest, artificial neural networks and decision tree learning algorithms were used. The article focused on data pre-processing including searching and labeling cancer comorbidity cases [14].

Some Research used machine learning techniques and statistical methods to perform the survival analysis for patients with spinal ependymoma. Their research was also based on SEER data. They discovered that lower grade histology and higher extend of surgical resection were the key prognostic factors. They compared statistical method with machine learning techniques. They concluded that therapeutic factors are associated with improved overall survival. Machine learning methods performed better in classification tasks, however the dataset were heterogeneous and complex with numerous missing values [18]. Stage-specific survival prediction has become a research interest. Fifteen recently published breast cancer survival prediction papers were analyzed together. Stage-specific prediction models and joint models were created and evaluated. They concluded that data-driven knowledge obtained with machine learning methods must be subject to over time validation before it could be clinically and professionally applied [19]. A statistical multivariate Cox proportional hazards model was developed to train and predict cases from Taiwan Cancer Registry (TCR) data. The model was applied to SEER database as well for validation purposes [20].

Some researchers considered SEER colorectal patients' survival prediction to be two-staged: the first stage was to predict survival, the second was to predict remaining life span of patients whose predicted outcome is death. The first stage adopted a tree ensemble classification method that took into account the imbalanced data. The regression stage used a tree-based selective ensemble regression method called SRRT-SEM [5]. Machine learning algorithms also had potential to improve lung cancer stage classification but might be prone to overfitting. Use of ensembles, cross-validation, and external validation could aid generalizability [21].

3. MACHINE LEARNING METHODS

This paper compares different machine learning algorithms' performances with our proposed custom ensemble learning method. The comparing machine learning methods are decision trees, random forests, adaboosts, artificial neural networks, and two gradient boosting methods. Decision Tree Learning is one of the most classic supervised models. It can be visualized as a graph. The structure of a decision tree is similar to an actual tree: the internal nodes test certain characteristics of the data; the branches represent outcomes of the test; each leaf node is a classification result. The taller and wider the tree grows, the better it can fit the training data. However, it could be overfitting the training set causing bad prediction to the test set. On the other hand, minor changes in the training data can cause large variations in the structure. However, the stability of the method alone is questionable. [8] Multilayer Perceptron (MLP) is a feed forward artificial neural network which has been very popular in pattern recognition areas.

The mathematical model is composed of the input layer, the hidden layers, and the output layer of artificial neurons. The structure of MLP is the simulation and abstraction of human brain's reaction system [9]. Random Forests are very similar to Bagging with one additional step of clustering the training sets before sampling from the training set. Assume one of the characteristics has strong impact on the tree structure, using all dimensions of the input would produce strongly biased base learners. To resolve this issue, we only use a part of the features of each sample to train the decision trees [10]. The boosting method used in this research is Adaboost. Boosting is also an ensemble learning. The difference between bagging and boosting is that instead of randomly taking samples from the dataset and voting for the final label, boosting assigns an initial weight for each group of sample. The weight for wrong predicting samples will increase to re-train the classifier in the next iteration. Bagging has a parallel connecting structure while boosting connects the classifiers in series [11]. GBM is the abbreviation for gradient boosting machine. It is also an ensemble learning method. Unlike adaboosts, which adjust the weights of the samples after each iteration, GBM uses the difference between the last iteration's output and the target value to be the new target for the next iteration [12]. LightGBM is a novel GBM method that implements Gradient-based One-side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques. It can speed up the training process of conventional GBDT by 20 times while achieving almost the same accuracy [13].

4. METHODOLOGY AND EXPERIMENTS

4.1. Two-staged Custom Ensemble Model

Most cancer prognosis researches are limited to predicting whether a patient can live for a specific amount of time. The patient is then classified as 'survival' or 'death'. Since gastric cancer has high mortality incidence, most cases would be classified as 'death'. The remaining survival time for these patients remains unknown. Therefore, we propose a two-staged classification model consisting of a classification model that predicts the patient's survivability, and a regression model that predicts the remaining lifespan of the patients whose predicted outcome is 'death'.[5]

Both stages have similar procedures except for the base machine learning types. Classifiers are adopted in the classification stage to predict survival condition, and regressors are used in the regression stage to predict survival months.

4.2. Data Acquisition

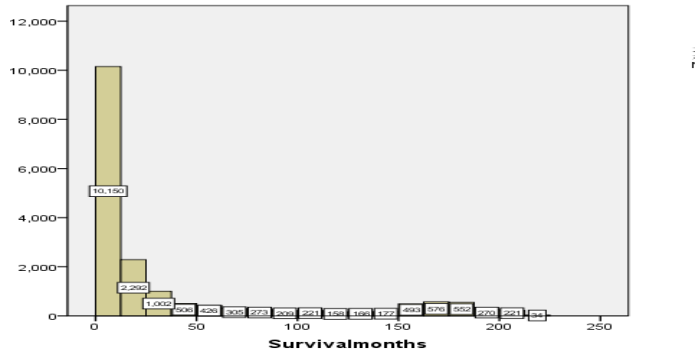


Figure 1. Histogram of the Dataset

The clinical data is collected from SEER database ranging from 1998 to 2002. SEER program collects cancer statistical data throughout the United States. The ultimate goal of SEER program is to reduce cancer burden in the United States. SEER develops a software, which provides easy

access for us to analyze SEER data. 17 features were selected in our research. Some are discrete and some are continuous. Table 1 include the chosen features and a brief description for each feature. Most of the features are discrete. These features are processed with One-hot encoding. The numeric features remains unchanged. The survival month's tab is transferred to the target for our prediction. Patients who lived over 60 months are labeled 1, others are labeled 0. Dropping patient ID and survival months, the rest of the chosen features are used to train and test the classifiers. Only stage II to stage IV patients are selected, because stage I patients can be cured at much greater chance thus should be treated differently. After removing registries with missing value, there are 18032 cases left. Figure 1 shows the distribution of the patients' survival time. Most of the patients did not live up to 5 years.

SEER database provides lots of attributes. Some of the attributes are similar to each other, while some have limited connection to our prediction. Table 1 lists the key attributes we select.

Table 1. Selected SEER attributes and their descriptors.

Feature Name	Description	Type
Patient ID number	Patient ID, unique for each patient	discrete
State-County	Origin of the Patient	discrete
Age	Age of Patient	numeric
Race	Race of Patient	discrete
Sex	Gender of Patient	discrete
Grade	Grading and differentiation codes	discrete
Radiation Sequence with Surgery	Whether and when the patient received radiation	discrete
Radiation Recode	Type of radiation	discrete
Primary Site_Labeled	Site of cancer	discrete
HISTOLOGY RECODE	Based on Histologic Type ICD-O-3.	discrete
RX Summ--Surg Prim Site (1998+)	Description of the surgery performed	discrete
Reginal nodes positive(1988+)	Number of regional lymph nodes examined to contain metastasis	numeric
Chemotherapy	Whether the patient has received chemotherapy	discrete
Survival Months	Time between death and diagnosis	numeric
Year of diagnosis	The year when patient was diagnosed	discrete

4.3. Scikit-Learn

Scikit-Learn is an open source machine learning package built in python environment. It covers almost all the major machine learning methods. It can be easily used and the parameters can be tuned.

4.4. Experimental Procedures

The training validating and testing data are then separated at the ratio of 3:1:1. The datasets in stage 1 include 10818 training data, 3606 validating, and 3607 testing. The datasets in stage 2 include 8592 training data, 2685 validating, and 2685 testing. For the tree-based ensemble learning algorithms, the number of estimators are all set to be 500 to be fair. Other parameters are fine tuned through GridSearching.

4.5. Custom Ensemble Procedures

After data preprocessing, the datasets are separated into training, validating, and testing sets. In the first stage, all gastric cancer cases in the training set are used to train the base machine learning classifiers introduced in section 3. The validation sets are fed into the classifiers to obtain predicted outcome and performance metrics. The custom ensemble approach sums of all predicted probabilities at specific weights. The probabilities are then transferred to binary predictions. This is a soft voting approach. The predicted outcome for each method are stored, and a loop is used to determine the weights for each method to reach the best performances in the validation sets by comparing accuracies of different combinations. The performance metrics for custom ensemble method and the base machine learning methods are then calculated among the testing sets to compare with each other.

In the regressing stage, cases whose survival time are greater than 60 months are abandoned from the datasets. Only the patients whose labels are 'death' are analyzed. The procedures are similar to the first stage. After going through training validation and testing sets, performance metrics and output values are compared and analyzed.

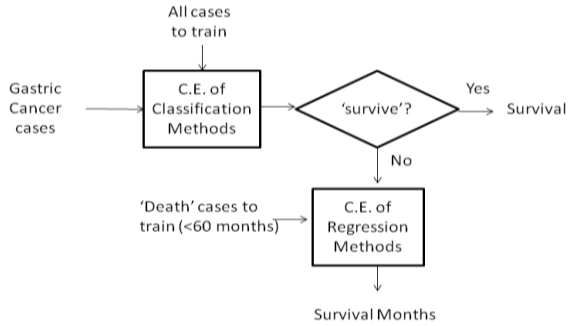


Figure 2. Flow chart of our two-stage prediction

4.6. Performance Metrics

The classification accuracy is quantified as recognition accuracy, precision, recall, and F-score. The calculation formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F} - \text{score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \quad (4)$$

True Positive (TP) is the number of correctly identified patients who lived longer than 60 months, True Negative (TN) is the correctly identified patients who did not survive up to 60 months. Where FP (False Positive) is the amount of patients incorrectly to be predicted to survive, and FN(False Negative) are the Patients whose label are 1 but are predicted to be 0. Accuracy refers to the ratio of correct predictions to the total sample. Precision considers the positive samples only, which is the ratio of correctly predicted positive sample to all the samples predicted to be in group 1. The recall rate refers to the ratio of accurately predicted positive samples to the actual total actual positive samples. Time to train and time to predict is the amount of time each method takes to train the model and to predict the results. AUC is the area under the ROC curve. F-score

is the harmonic mean of the precision and recall. Reducing either of them would cause a smaller f-score. [7]

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2} \quad (5)$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_{test}^{(i)} - \hat{y}_{test}^{(i)}\right| \quad (6)$$

$$R^2 = 1 - \frac{\sum_i(\hat{y}_{test}^{(i)} - y_{test}^{(i)})^2}{\sum_i(\bar{y}_{test}^{(i)} - y_{test}^{(i)})^2} \quad (7)$$

The regression performance is evaluated through Root Mean Squared Error, Mean Absolute Error, and coefficient of determination (R^2). The RMSE of a model is the average distance between the model's prediction and the actual outcome. The MAE measures the absolute average of the difference between prediction and the actual label. The R^2 value maps the accuracy to a value between 0 and 1, so that it can be used to compare predicting performances on different datasets [5][6].

5. RESULTS

5.1. Stage for classification

Table 2 shows confusion matrix for the testing set in classifying stage. There are 3607 testing cases in total. From the confusion matrix we can see that the traditional gradient boosting machine has outstanding performance predicting 'survive' cases, while the artificial neural networks are good at predicting cases whose actual outcome is 'death'.

Table 2. Classification Outcome vs. Actual Outcome

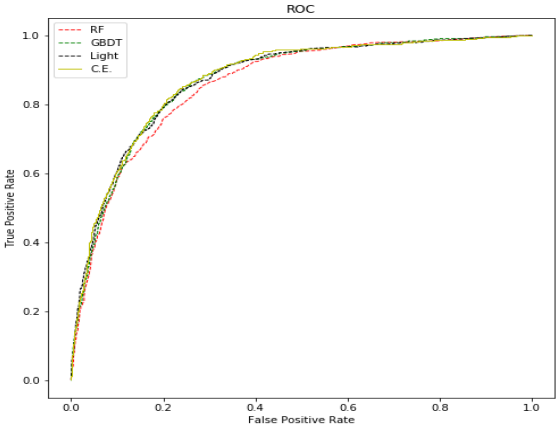
	Actual 0	**Actual 1**	**Predicted**
Decision Trees	2708	465	0
	161	273	1
Random Forests	2691	424	0
	178	314	1
Adaboost	2710	424	0
	159	314	1
GBM	2665	**379**	0
	204	**359**	1
ANN	**2806**	657	0
	63	81	1
LightGBM	2734	442	0
	135	296	1
Custom Ensemble	2730	403	0
	135	296	1

Table 3 contains the performance metrics calculated for each method. Among the base learners, Light GBM has the best accuracy score and the highest precision rate. Its computational speed is also fast ranking the second among all methods. Traditional GBM has very close accuracy score to lightGBM. The F-score and recall rate are even better that the F-score is very close to our custom ensemble approaches, and the recall rate is even better than the ensemble method's. The computation time is very close to the lightGBM approaches. Adaboosts have almost the same accuracy to GBM method. The precision rate ranks second to lightGBM and the area under curve is even higher than that of lightGBM. Random Forest Classifier has mediocre performance among all methods. Artificial neural network and decision tree learning has weaker performances but provide variations in the base learners.

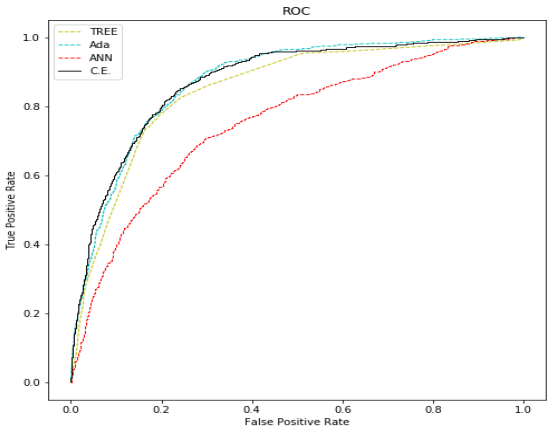
Table 3. Performance Metrics of the Classification Stage

Model	Tree	RF	ADA	GBM	ANN	Light	C.E.
Accuracy	0.826	0.833	0.838	0.838	0.800	**0.840**	**0.850**
F-Score	0.466	0.511	0.519	**0.552**	0.187	0.50	**0.553**
Precision	0.629	0.638	0.664	0.638	0.563	**0.686**	**0.707**
Recall	0.370	0.425	0.425	**0.486**	0.110	0.401	0.453
AUC	0.848	0.857	**0.873**	0.868	0.754	0.870	**0.874**
Train time	0.038	3.31	2.36	0.34	0.750	0.345	N.A.
Pred time	0.0004	0.269	0.19	0.003	0.003	0.003	N.A.
Weight	0.105	0.158	0.19	0.474	0.105	0.158	N.A.

We can see that the custom ensemble outcome has better performance in most performance metrics. Outstanding base learners have heavier weighting factors in the ensemble that good performing base learners are important to a good ensemble. Their diversity are also important that weaker base learners also participate in the voting.



(a)



(b)

Figure 3. ROC curves for all methods

Figure 3 contains the ROC curves for all methods. Adaboosts, Gradient Boost Machine, Light GBM and our custom ensemble approach has close ROC curves. To avoid overlapping, the curves are plotted in 2 subplots. Except ANN, all other machine learning techniques are very close on the graph. The ensemble approaches in subplot (a) has very close performances while

our custom ensemble approach is slightly better and random forests approach is slightly worse. In subplot (b), decision trees method is slightly below Adaboosts while ANN has the worst ROC curve among all methods.

5.2. Stage for Regression

Table 4. Performance Metrics of the Regression Stage

Method	Tree	RF	Ada	GBM	ANN	Light	C.E.
RMSE	12.1	12.24	12.34	11.67	12.30	11.67	11.6
MAE	8.56	8.86	9.73	8.33	8.55	8.29	8.39
R^2	0.126	0.105	0.090	0.187	0.097	0.186	0.19
Weight	0.071	0.071	0.071	0.286	0.071	0.429	N.A.

Light GBM has the smallest root mean squared error and mean absolute error among all base learners, but the traditional GBM method has better R2 score. Our ensemble approach can improve the prediction performance that the RMSE and R2 are both better. From Figure 4, we can see that decision trees and adaboosts output discrete values. The predicted survival time lines up with the actual values for medium low to high values (~10 to ~45 months) for random forests, artificial neural networks, traditional GBM, and Light GBM methods. The prediction for lower range (0 to 10 months) however do not well line up with actual values. The custom ensemble approach has fewer outliers comparing to base machine learning algorithms.

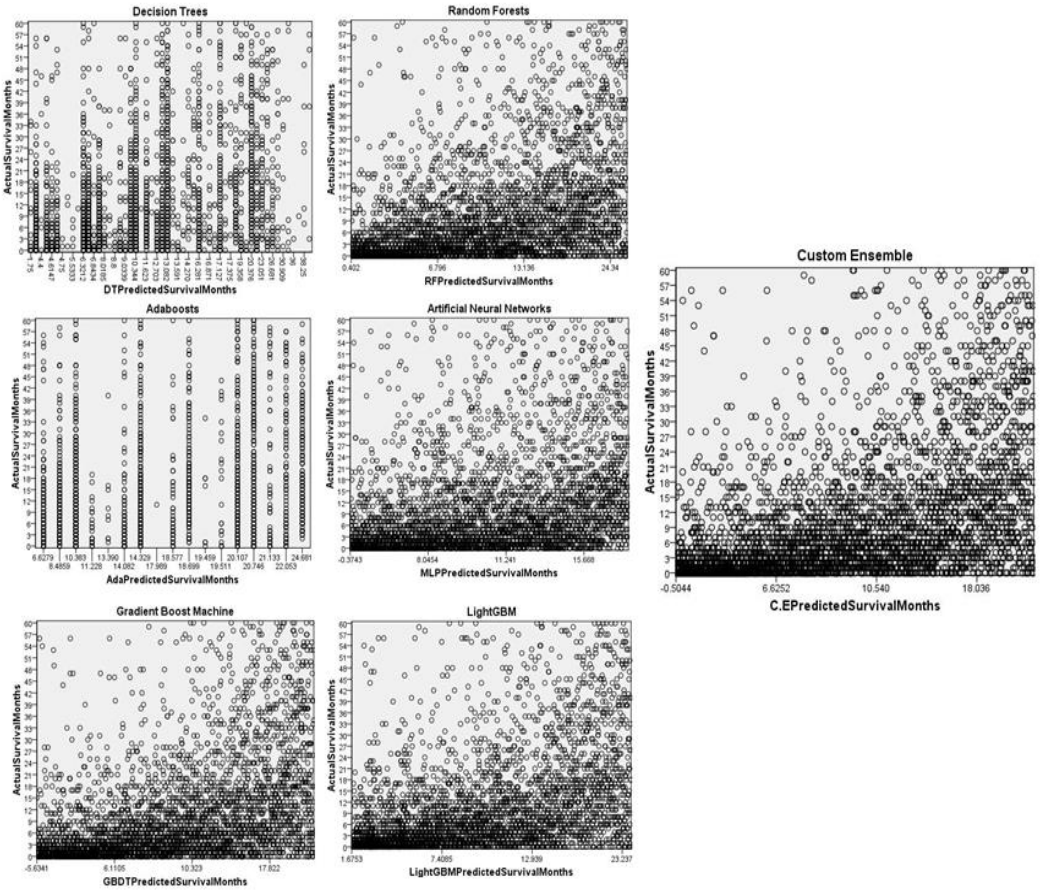


Figure 4. Predicted survival time to actual for each method

To compare machine learning methods, we draw the correlation scatterplots in Fig. 5. We can see that random forests and GBM has the moderate correlation. Decision trees related correlation plots also shows some linearity. Overall, the plots suggest limited interaction effects. The variation between methods provides diversity in the ensemble.

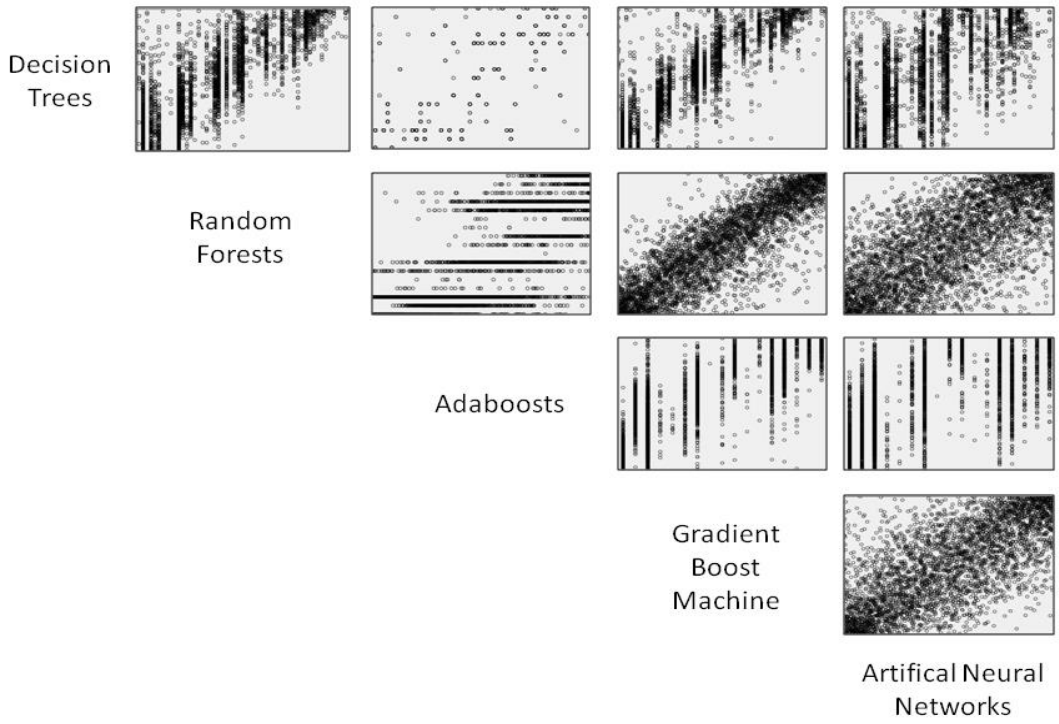


Figure 5. Correlation scatterplot comparing base machine learning methods

6. DISCUSSION

In both stages, our custom ensemble approach has the best performance that the RMSE and R2 are both better. LightGBM is the most able base model with an classification accuracy of 84% and the RMSE value 11.67, as shown by the results in Table 3 and Table 4. Other ensemble learning methods outperforms simple methods in the first stage regarding the accuracy score. In the regression stage, decision trees have better RMSE values than adaboosts and random forests. The results imply that one method may be strong at classification tasks but weak at regression tasks. The ROC curves in Figure 3 are very close to each other except for ANN. The ANN method was kept in the base learners for its diversity.

The weighting factor is determined using a loop, and outstanding base learners have heavier weighting factors in the ensemble since good performing base learners are important to a good ensemble. Diversity is important that weaker base learners also participate in the voting. The correlation scatterplot Figure. 5 in the second stage shows limited interaction between each other. This provides diversity in our ensemble.

7. CONCLUSION

Predicting Survivability of Gastric Cancer patients can help the doctors and the patients' family. For high incidence cancers, this may not provide sufficient amount of information to support better medical decisions. Our two-stage survivability prediction model is proposed to deal with

this problem. The first stage predicts patients' five year survivability. If the prediction is 'death', which is common in gastric cancer database, the second stage predicts the remaining lifespan of the patient.

Random Forests, adaboosts, and gradient boost machine are the four ensemble learning methods we usually use in survivability analyses. Decision Trees and Artificial Neural Networks are the basic machine learning methods. LightGBM is also utilized in our research providing good classification and regression outcome at high computational speed. These methods provide good accuracies and diversity. The results in our research indicate that LightGBM is the best performing base learner in both stages. Other ensemble tree algorithms, including Random Forests, Adaboosts, and GBM also have competitive prediction performance. Decision trees and neural networks provides variations in the prediction. Researches conducted by other scholars usually focus on fine-tuning one method to obtain better classification accuracy. By voting multiple machine learning methods' outcome at specific weighting factors, we obtain a more adept approach.

In the future, we will try to find a method to obtain fixed weighting factors. Right now, the weighting factors are determined in each iteration. This could cause problems if we want to apply our research to real life. We will also extend our research to other cancers including those with lower mortality rate and multiple primary cancers. We will keep searching for more adept machine learning methods.

REFERENCES

[1] Jemal, A & Bray, F. & Center, M. & Ferlay, J. & Ward, E. & Forman, D., (2011) "Global Cancer Statistics," CA Cancer J Clin. 61(2):69.

[2] Parkin, DM. (1999) "Epidemiology of cancer: global patterns and trends," Toxicol Lett. 102-103:227.

[3] Correa, P. (2011) "Gastric cancer: two epidemics?" Dig Dis Sci.;56(5):1585.

[4] Glare, Paul.& Sinclair, Christian.& Downing, Michael. et al. (2008) "Predicting survival in patients with advanced disease". Eur. J. Cancer 44(8) 1146-1156.

[5] Wang, Yuyan & Wang, Dujuan & Ye, Xin et al, (2019) "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction". Information Sciences, 474:106-124.

[6] Lynch, Chip & Abdollaha, Behnaz & Fuqua, Joshua, et al. (2017) "Prediction of lung cancer patient survival via supervised machine learning techniques".Int J of Med Informatics. 108.

[7] NCI_SEER_Overview, Overview of the SEER Program. Surveillance Epidemiology and End Results, (2015) (Available from: http://seer.cancer.gov/about/).

[8] Liu, Ya-Qin & Wang, Cheng & Zhang, Lu, (2009) "Decision tree based predictive models for breast cancer survivability on imbalanced data", in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedical Engineering, pp. 1–4.

[9] Ng, E.Y.K.& Acharya, U.R. & Keith, L.G., et al., (2007) "Detection and differentiation of breast cancer using neural classifiers with first warning thermal sensors", Inf. Sci. 177 (20) 4526–4538.

[10] Nguyen, Cong & Wang, Yong & Nguyen, H.N., (2013) "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic", J. Biomed. Sci.Eng. 6 (5) 551–560.

[11] Thongkam, Jaree & Xu, Guandong & Zhang, Yanchun, et al. (2008) “Breast Cancer Survivability via AdaBoost Algorithms[C]” Australasian Workshop on Health Data & Knowledge Management. Australian Computer Society, Inc.

[12] Park, Kanghee, et al. (2013) "Robust predictive model for evaluating breast cancer survivability." Engineering Applications of Artificial Intelligence 26.9:2194-2205.

[13] G. Ke, et al. (2017) “A highly efficient gradient boosting decision tree. ”In Advances in Neural Information Processing Systems, 3149–3157.

[14] Naghizadeh, Majid & Habibi, Narges, (2019) “A model to predict the survivability of cancer comorbidity through ensemble learning approach.” Expert Systems, 36:e12392.

[15] Kaviarasi, R & Gandhi, R, (2019) “Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System.” Journal of Medical Systems, 43(7):201.

[16] Lu, Hongya & Wang, Haifeng & Won, Yoon Sang, (2019) “A Dynamic Gradient Boosting Machine Using Genetic Optimizer for Practical Breast Cancer Prognosis.” Expert Systems with Applications.

[17] Wen, H. & Li, S. & Li, W. & Li, J. & Yin, C., (2018) “Comparison of Four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability. ”112-116. 10.1109/ICCWAMTIP.8632577.

[18] Ryu, Sung Mo & Lee, Sun-Ho & Kim, Eun-Sung, Whan, Eoh, (2019) “Predicting Survival of Patients with Spinal Ependymoma Using Machine Learning Algorithms with the SEER Database”, World Neurosurgery, 124:331-339.

[19] Kleinleina, R. & Riaño, D. (2019) “Persistence of data-driven knowledge to predict breast cancer survival.” International Journal of Medical Informatics 129 303–311.

[20] Huang, Ching-Chieh & Chan, Soa-Yu & Chiang, Chun-Ju & Lee, Wen-Chung et al., (2019) “Development of a prediction model for breast cancer based on the national cancer registry in Taiwan,” Breast Cancer Research 21:92.

[21] Brooks, Gabriel & Bergquist, Savannah & Landrum, Mary &Rose, Sherri, (2019) “Classifying Stage IV Lung Cancer From Health Care Claims: A Comparison of Multiple Analytic Approaches.” JCO CLINICAL CANCER INFORMATICS May.

[22] Li, Leijun. & Hu, Qinghua & Wu, Xiangqian et al., (2014) “Exploration of classification confidence in ensemble learning”, Pattern Recognit. 47 (9) 3120–3131 .