공학박사 학위논문

인공지능 기반 종양학 임상 의사결정 지원 시스템: 다기관 확장을 위한 기술적 문제 및 과제

Artificial intelligence (AI)-based clinical decision support system in oncology: Technical issues and challenges for multi-center extension

울산대학교대학원
의과학과
이경화

인공지능 기반 종양학 임상 의사결정 지원 시스템: 다기관 확장을 위한 기술적 문제 및 과제

지도교수 김남국

이 논문을 공학박사 학위 논문으로 제출함

2022 년 2 월

울산대학교대학원
의과학과
이경화

이경화의 공학박사 학위 논문을 인준함

심사위원 서준범 (인)

심사위원 고범석 (인)

심사위원 김강모 (인)

심사위원 김남국 (인)

심사위원 김휘영 (인)

울산대학교대학원

2022 년 2 월

Abstract

Counseling patients in weighing their individual prognosis and selecting appropriate treatment strategies are essential aspect of care for patients with cancer. Although the common statistical models familiar to clinicians can determine prognostic factors and evaluate the relative risk of cases with specific prognostic factors, it is difficult to predict the individual prognosis of a patient using it. Recently, the medical field has experienced increased attempts to develop a clinical decision support system (CDSS) using artificial intelligence. However, developing personalized prognostic prediction models with various, complex information and ensuring its availability to other institutions with different patient groups and characteristics, remain challenges yet to be overcome. In this thesis, a machine learning-based two-stage model that can recommend initial treatment option and predict overall survival in patients with hepatocellular carcinoma (HCC) was developed. In particular, this model was verified using external datasets obtained from eight medical centers in South Korea, and the technical issues, challenges, and strategies for multi-institutional usability were discussed.

For the first phase of this thesis, the model to recommend one of six treatments used for the initial treatment of HCC using 20 pretreatment key variables was developed and validated employing multi-center datasets. The recommendation was made by considering the results of ensemble voting classifier that was created using five machine learning classifiers that offered the best performance after testing several models. In addition, the performance of individual training with the dataset of each institution was compared with those of external validation of the model trained with the internal dataset. Although individual training revealed better performance, results of external validation of the model exhibited acceptable performance with the setting of providing a second treatment option. Combining these experimental results, it was suggested that providing a second treatment option along with the first, with its level of confidence, were found to be effective in extending this model to multi-centers with different preferences and policies.

For the second phase of this thesis, a model for survival prediction following initial treatment was developed. Overall survival for each patient was predicted employing a random survival forest model using initial treatment information in addition to 20 key variables used in the model for treatment recommendation. Survival prediction from individual training for each center exhibited similar or worse performance than those obtained from external validation in contrast to the results of the model for treatment recommendation. Furthermore, an experiment was performed to stratify the risk of each treatment by simulating how the predicted survival changes according to the results of treatment recommendation in the first stage.

In the third phase of this thesis, specific scenarios demonstrating the applicability of this model in real clinical setting were presented. First, the possibility of employing this model as an alternative to a current staging system was investigated. The results of recommendation of this model and the Barcelona Clinic Liver Cancer (BCLC) staging system in group of BCLC C stage were compared. The agreement between the treatment recommended in this model and the treatment actually received was higher than the treatment recommended in the BCLC stage. Second, results of simulation for a case employing external datasets using two different models trained with dataset of two different centers were demonstrated. The result showed that for even patients with the same conditions the model can recommend different treatments and show different survival by reflecting the characteristics based on the dataset of each institution. These usage scenarios show examples of how this model can be extended and used in real clinical situations.

In conclusion, a machine learning-based two-stage model with 20 clinical variables to recommend appropriate initial treatment and sequentially predict overall survival for patients with HCC was developed. Furthermore, various experiments were conducted to apply this model to multiple centers in real clinical environment, and the results obtained were analyzed. This model is expected to provide practical utility to physicians and institutions with little experience in actual clinical settings.

Key words: Clinical decision support system, Machine learning, Hepatocellular carcinoma, Treatment recommendation, Survival prediction

Contents

List of Tables

List of Figures

1. Introduction

1.1. Motivation

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells [1]. In South Korea, cancer is the most common cause of death, accounting for nearly one of every four deaths [2]. Counseling patients considering their individual characteristics and selecting appropriate treatment strategies form a crucial part of the care for patients with cancer. The general statistical model is familiar to the clinician. However, although a statistical model can estimate the relative risk, it is difficult to predict the prognosis for an individual patient. Moreover, the quantity and quality of clinical data has been rapidly expanding, including electronic health records, enormous quantity of information from diverse medical images and examinations, genetic information, disease registries, and patient surveys. However, big data and digitalized information do not automatically lead to better patient care.

A clinical decision support system (CDSS) is concerned with improving healthcare delivery via enhancing medical decisions with targeted clinical knowledge, patient information, and other health information [3]. CDSS gathers and represents knowledge in a manner that facilitates simulations of human reasoning using computers to generate advice [4]. Therefore, high-quality CDSS is essential to exploit the full benefits of huge amounts of information. Recently, the biomedical field has witnessed increased use of machine learning, particularly in image diagnosis. However, development and evaluation for individualized prediction models with various, complex information using these methods remains a challenging problem. Despite extensive research regarding the use of machine learning and deep learning in medical fields has been published, several problems are yet to be solved. This thesis demonstrates the development of a model for CDSS and technical issues and challenges for multi-center extension.

Treatments for cancer patients have characteristics that continuously change with time. The growth of massive genetic and clinical databases, along with computing systems to exploit them result in the acceleration of the speed of treatment advances while shortening the cycle time for changes to treatment guidelines in oncology. In addition, these information management challenges have been occurring in a practice environment wherein the time available for tracking and accessing relevant information is minimal. In particular, in the case of hepatocellular carcinoma (HCC), rather than uniform standard treatment considering the stage, various treatments are used according to the various residual liver functions of the patient and whether the treatment perform well at each institution as well as the stage. In such a scenario, developing a model and ensuring its availability to multiple centers was the research motivation for this study.

1.2. Contributions

The main contributions of this thesis are summarized as follows. First, a machine learning-based model was developed for initial treatment recommendation in patients with HCC. In particular, the focus was on the modification and application of this model for use in a multi-center setting. Accuracy was increased by employing an ensemble voting machine compared with previous cascaded random forest model for the internal dataset. In addition, providing a second treatment option along with the first, with its level of confidence, were found to be effective in extending this model to multi-centers with different preferences and policies. Second, a model was developed for survival prediction following the initial treatment recommended by the model in the first phase. The two-stage model was simulated, and consequently the risk was stratified via predicting the survival considering the results of treatment recommendation. Third, we demonstrated several scenarios of this model in real clinical situations. The feasibility of this model functioning as an alternative to the current staging system was described as a first scenario. Furthermore, we simulated certain cases using two different models with same structure but separately trained with different dataset from two centers. The possible expansion of this model may aid both physicians and patients in real clinical setting.

1.3. Organization

The remainder of the thesis presents each of the above-listed contributions in further detail. Chapter 2 presents a brief overview of the background of this thesis. Chapter 3 describes the characteristics of internal dataset and a previously proposed CDSS for treatment recommendation and survival prediction after initial treatment in patients with HCC. In the previous study with dataset of a single center, cascaded random forest model was employed for treatment recommendation and random survival forest model for survival prediction. In addition, the process of feature selection of 20 pretreatment key variables as a model input and performance of previous models was described. Finally, data collection and characteristics of external datasets obtained from eight institutions in South Korea have been presented at the end of the chapter.

Chapter 4 presents in detail the various experiments performed and results obtained on the model for treatment recommendation. The ensemble voting machine has been applied and its performance was compared with a previous cascaded model. Various normalization and oversampling methods were employed to improve the model performance. Further, the results of individual training with dataset of each center were compared with those of external validation. The option for a second treatment in addition to first was also investigated for the application of this model to multi-center setting. Lastly, the model calibration and the results have been described in this chapter.

Chapter 5 explains the experiments performed and results obtained on the model for survival prediction. Different training modes for random survival forest model was evaluated for internal dataset. Further, using the modified model, the results of individual training with dataset of each center and those of external validation were compared. Furthermore, the simulation results of the proposed two-stage model and risk stratification via predicting the survival considering the results of treatment recommendation have been described.

In the Chapter 6, several usage scenarios of CDSS have been discussed. The results of recommendation between the proposed model and the Barcelona Clinic Liver Cancer (BCLC) staging system in patient group of BCLC C stage were compared and the possibilities of employing this model as an alternative of the current staging system were demonstrated. In addition, the concept of digital twin was presented with simulation of a case of one patient using two different models trained with datasets from two centers, and the expandability of this proposed model in real clinical setting has been described.

Finally, Chapter 7 presents a discussion of the limitations and challenges with regard to the proposed CDSS and a conclusion. Various issues and possible solutions focusing on the model deployment have also been described and consequently, future research topics, required for these issues and solutions have been presented.

2. Background

This chapter presents a brief overview of the background of this thesis. In Section 2.1, the basic information related to machine learning as well as issues and challenges associated with machine learning-based models on biomedical datasets have been described. In Section 2.2, various machine learning algorithms for multi-class classification have been presented, which were later employed in the model described in Chapter 4. In Section 2.3, a brief introduction of machine learning methods for time-to-event prediction has been discussed, which were used in the model discussed in Chapter 5. Finally, Section 2.4 presents commonly used metrics for evaluating performance of multi-class classification and time-to-event prediction.

2.1. Basic information on machine learning

2.1.1. Machine learning

According to Arthur Samuel, machine learning is defined as the field of study that enables computers to learn without being explicitly programmed [5]. Machine learning is used to teach machines the manner in which to handle the data more efficiently. With the abundance of datasets available, the demand for machine learning has increased, and many industries have applied machine learning to extract relevant data. Many studies focusing on ways to make machines learn by themselves without being explicitly programmed have been conducted. In addition, many mathematicians and programmers have applied several approaches to determine the solution of problems that contain huge data sets. Figure 2.1 shows commonly used algorithms in machine learning. Machine learning relies on different algorithms to solve data problems, and the type of algorithm employed depends on the type of problem to be solved, number of variables, and type of model best suited for the purpose

[6].

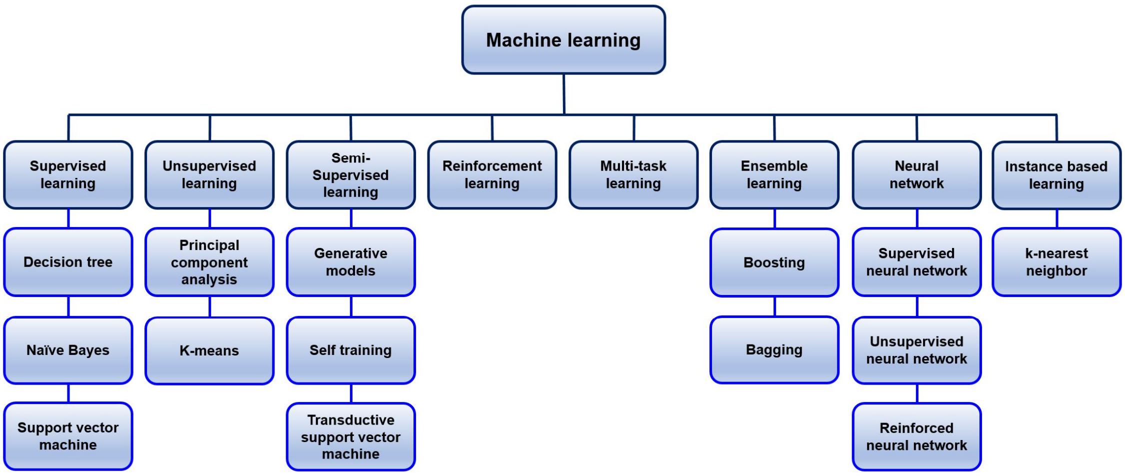


Figure 2.1. Commonly used algorithms in machine learning

2.1.2. Issues and challenges with biomedical dataset

In a majority–minority classification problem, class imbalance in the datasets can dramatically skew the performance of classifiers, which results in the introduction of a prediction bias for the majority class [7]. A dataset is imbalanced if the classification categories are not approximately equally represented, which is very common in case of a biomedical dataset. Often real-world datasets are predominately composed of "normal" examples with "abnormal" examples comprising a small portion. In addition, as a common occurrence, the cost of misclassifying an abnormal example as a normal example is often much higher than the cost of the reverse error. Moreover, when working with big data, the mitigation of class imbalance poses an even greater challenge because of the varied and complex structure of the relatively much larger datasets. Consequently, attempts have been made to handle imbalanced datasets in domains such as fraudulent telephone calls, telecommunications management, text classification, and detection of oil spills in satellite images [8-13]. Leevy et al. provided a large survey of published studies focusing on high-class imbalance (i.e., a majority-to-minority class ratio between 100:1 and 10,000:1) in big data to assess and address the adverse effects owing to class imbalance [7].

Their study discussed two techniques: data-level and algorithm-level methods. The data-level approach involves sampling and feature selection techniques. Further, sampling techniques consist of over-sampling and under-sampling solutions, wherein in case of the over-sampling process, instances from the minority class are added via replication to the given dataset, with the replication being done either randomly or using an intelligent algorithm. In contrast, during the under-sampling process, instances from the majority class are removed from the given dataset, with the removal following a random pattern. Chawla et al. proposed the synthetic minority over-sampling technique (SMOTE), wherein the over-sampling of minority class and under-sampling of majority class were combined, with the former involving the creation of synthetic minority class examples. It improved the classifier performance in receiver operating characteristics (ROC) space than that in case of only under-sampling the majority class [14]. However, in the Leevy et al.'s review, random over-sampling (ROS) was considered to exhibit a better classification accuracy than random under-sampling or SMOTE in most studies.

Feature selection methods may also aid in the selection of the most influential features that can yield unique knowledge for inter-class discrimination [15, 16]. Mladenic and Grobelnik et al. utilized a feature-subset selection approach developed for a Naive Bayes classifier on imbalanced text data from multiple domains [10]. They investigated 11 different feature scoring measures and determined that the odds ratio produced the best results. The authors also concluded that considering domain and algorithm characteristics significantly improves classification results. Zheng et al. also investigated feature selection for text categorization with imbalanced data [17]. Their approach selected the positive features and negative features separately using feature selection techniques including Information Gain, Chi Square, correlation coefficient, and odds ratio, and then explicitly combined. Further, they presented variations of the odds ratio and Information Gain metrics to especially address class imbalance. Their study used the Naïve Bayes and Regularized Logistic Regression as classifiers, and the proposed approach yielded good results. Yin et al. demonstrated that both decomposition-based and Hellinger's distance-based methods can outperform existing feature-selection methods for imbalanced data [18]. The higher the distance (i.e., lower affinity) value, the better the corresponding feature. Thus, the Hellinger's distance can be used to measure the prediction power of features to classify instances.

The algorithm-level approach includes cost-sensitive and hybrid/ensemble techniques. Cost-sensitive techniques are based on the general principal of assigning more weight to an instance or learner in the event of a misclassification. For example, a false negative prediction may be assigned a higher cost compared to a false positive prediction, given the hybrid/ensemble techniques are the class of interest. Further, cost-sensitive techniques include a fuzzy rule-based classification approach coupled with an online learner scheme. In contrast, the hybrid/ensemble techniques include a Bayesian optimization algorithm that maximizes Matthew's correlation coefficient by learning optimal weights for the positive and negative classes [19], coupled with an approach that combines ROS and support vector machines. In addition, ensemble methods can also be used as cost-sensitive methods, with the classification outcome being certain combinations of multiple classifiers built on the dataset; bagging and boosting are two common types of ensemble learners.

2.2. Multi-class classification

Supervised multiclass classification algorithms aim at assigning a class label for each input example. For a particular training data set of the form $(x_i,\ y_i)$, where $x_i \in \mathbb{R}^n$ is the ith example and $y_i \in \{1, \dots, K\}$ is the ith class label, the aim is to determine a learning model $\mathbb{H}$ such that $\mathbb{H}(x_i) = y_i$ for new unseen examples. Aly et al. [20] presented the different approaches employed to solve the problem of multi-class classification. The first approach relied on extending binary classification problems to address the multiclass case directly, through neural networks, decision trees, support vector machines, naive bayes, and k-nearest neighbors. The second approach decomposes the problem into several binary classification tasks, with several methods employed for this decomposition: one-versus-all, all-versus-all, error-correcting output coding, and generalized coding. Finally, the third one involved arranging the classes in a tree (typically a binary tree) and utilizing several binary classifiers at the nodes of the tree till a leaf node is reached. This section presents the algorithms in the first approach that are commonly used for multi-class classification.

Neural Networks. Multi-layer feedforward neural networks provide a natural extension to the multiclass problem [21]. Rather than having one neuron in the output layer, with binary output, N binary neurons can be acquired. The output codeword corresponding to each class can be chosen as one-per-class coding or distributed output coding. In case of one-per-class coding, each output neuron is designated the task of identifying a particular class, whose output code should be 1 at the particular neuron and 0 for the others. Therefore, $N = K$ neurons are required in the output layer, where K denotes the number of classes. Further, when testing an unknown example, the neuron providing the maximum output is considered the class label for that example. For instance, for a four-class problem the output codes can be 1000, 0100, 0010, and 0001. In contrast, in case of distributed output coding, each class is assigned a unique binary codeword from 0 to $2N - 1$, where N is the number of output neurons. Consequently, when testing an unknown example, the output codeword is compared to the codewords for the K classes, and the nearest codeword, according to certain distance measure, is considered the winning class. Usually, the Hamming distance is used, which is the number of different bits between the two codewords. For instance, for a four-class problem, and using N = 5-bit codewords, the coding can be as 00000, 00111, 11001, and 11110. Moreover, the hamming distance between each pair of classes is equal to 3, that is, each pair of codes differ in three bits. For an unknown example, if the codeword is 11101, its distance is computed using the four codewords shown above. Furthermore, the nearest codeword involves class 3 with a distance of 1, such that the class label assigned to that example is class 3.

Decision Trees. Decision trees are a powerful classification technique. Two widely known algorithms for building decision trees are Classification and Regression Trees [22] and ID3/C4.5 [23]. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. The split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [32]. The generalization error for forests converges to a limit when the number of trees in the forest increases to a large value, and it depends on the strength of the individual trees in the forest and the correlation between them. A random selection of features is used to split each node, which yields error rates that are comparable to Adaboost [33], but more robust with respect to noise. Further, internal estimates monitor error, strength, and correlation, which are used to indicate the response to increasing the number of features used in the splitting. In addition, they are used to measure variable importance. Tree boosting is a highly effective and widely used machine learning method [34]. Among the machine learning methods used in practice, gradient tree boosting [35] a technique that demonstrates favorable outcomes in many applications. Tree boosting has been proven to yield state-of-the-art results on many standard classification benchmarks. XGBoost is a scalable end-to-end tree boosting system, which has been used widely by data scientists for the purpose of achieving state-of-the-art results in case of many machine learning challenges. It was proposed for a novel sparsity-aware algorithm for sparse data and weighted quantile sketch to realize approximate tree learning.

***k*-Nearest Neighbors.** k-nearest neighbors (kNN) [24] is considered among the oldest non-parametric classification algorithms. To classify an unknown example, the distance (using certain distance measure e.g., Euclidean) from that example to every other training example is measured. Consequently, the k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation.

Naive Bayes. Naive Bayes [25] is a successful classifier based upon the principle of maximum a posteriori (MAP). For a particular problem with K classes $\{C_1, \ldots, C_K\}$ having prior probabilities $P(C_1), \ldots, P(C_K)$, the class label c can be assigned to an unknown example with features $x = (x_1, \ldots, x_N)$ such that $c = argmax_c P(C = c \parallel x_1, \ldots, x_N)$. In other words, the class with the maximum a posterior probability for the observed data is chosen. This aposterior probability can be formulated using Bayes theorem as follows: $P(C = c \parallel x\_1, \ldots, x\_N) = \frac{P(C=c)P(x_1,\ldots,x_N \parallel C=c)}{P(x_1,\ldots,x_N)}$. As the denominator is the same for all classes, it can be excluded from the comparison. Now, the class conditional probabilities of the features needs to be computed for the available classes. However, considering the dependencies between features, this can be a challenging task. The naive Bayes approach assumes class conditional independence, that is, $x_1, \ldots, x_N$ are independent for the particular class. Consequently, the numerator is simplified to $P(C = c)P(x_1 \parallel C = c) \ldots P(x_N \parallel C = c)$. Subsequently, the class c that maximizes this value over all the classes $c = 1, \ldots, K$, is chosen. As evident, this approach can be naturally extended to the case involving more than two classes and has been shown to perform well despite the underlying simplifying assumption of conditional independence.

Support Vector Machines. Support vector machines (SVM) are among the most robust and successful classification algorithms [26, 27], and are based upon the idea of maximizing the margin; that is, maximizing the minimum distance from the separating hyperplane to the nearest example. The basic SVM supports only binary classification, but extensions [28-30] with additional parameters and constraints added to the optimization problem have been proposed to handle the separation of the different classes, thereby rendering it suitable for multiclass classification. However, the formulation of [30] can result in a large optimization problem, which may be impractical for a large number of classes. In contrast, Crammer et al. [29] reported a better formulation with a more efficient implementation.

The decision function of SVM is an optimal hyperplane that serves to separate observations belonging to one class from another considering patterns of information regarding those observations called features [31]. Subsequently, the hyperplane can be used to determine the most probable label for unseen data. However, the features used to infer the hyperplane are not typically raw data; rather, they are most often derivative data resulting from interpolation during the feature selection stage. Moreover, the features are further referenced by coordinates based on their relationships to each other and form the support vectors. Similar to other forms of machine learning, working with SVM involves balancing two complementary aims: (1) maximizing the percentage of correct labels assigned to new examples by the classifier (i.e., optimizing its accuracy) and (2) ensuring that the classifier is generalizable to new data (i.e., optimizing its reproducibility). While the former is bound by the informativeness of the features used (i.e., feature importance), the latter is bound by the number of unique examples used to train the model.

2.3. Time-to-event prediction

There are several models that explain the time-to-event prediction as a classification problem [32]. In general, it is referred to as survival analysis, although survival analysis in a narrow sense specifically deals with survival versus death [33]. The outputs of survival prediction models vary because the modeling methods are diverse. However, the outputs can be categorized into two large types: time-independent and time-dependent. The time-independent model generates a single value as the output for each patient, regardless of the follow-up time. In contrast, the time-dependent model prediction model calculates the output value separately for each follow-up time for each patient. Therefore, one patient can have multiple model output values, one for each follow-up time. Furthermore, it is possible to convert one type of model output to the other under certain circumstances. The time-independent model outputs can be converted into time-dependent results. For example, the survival probability at time $t, S(t, X)$, can be calculated from the log-risk scores if the baseline survival probability at time t is available. The other way round, multiple time-dependent model outputs per patient may be reduced to a single time-independent value for analytical purposes.

In this section, recent machine learning algorithms of time-independent and time-dependent models were introduced, including the random survival forest, which is used in this thesis for survival prediction and was thus primarily focused upon. Certain examples of time-independent model are prediction models such as DeepSurv that use the log-risk score estimated by the Cox proportional hazards model denoted as $\sum_{i=1}^{p} \beta_i X_i$, and risk scores created using the rounded integer values of the regression coefficients (β) divided by the reference value (generally, the smallest β in the regression model). Random survival forest and Nnet-survival model fall under the gambit of time-independent models. Random survival forest, estimates the cumulative hazard function at time t, denoted as $\Lambda(t, X)$, while Nnet-survival provides the predicted $S(t, X)$ at multiple specified time points as the model output.

DeepSurv. Katzman et al. proposed a modern Cox proportional hazards deep neural network, henceforth referred to as DeepSurv, which is a deep learning model that uses the log-risk function of the Cox proportional hazards model as the final output function [34]. DeepSurv is a deep feed-forward neural network that predicts the effects of a patient's covariates on their hazard rate parameterized by the weights of the network θ. The baseline data of the patient x, is used as the input to the network. The hidden layers of the network comprise a fully connected layer of nodes, followed by a dropout layer. The output of the network $\hat{h}_\theta(x)$ is a single node with a linear activation which estimates the log-risk function in the Cox model.

$$\lambda(t|x) = \lambda_0(t) \cdot e^{h(x)} \qquad \textbf{(2. 1)}$$

The network was trained by setting the objective function to be the average negative log partial likelihood of (2.2) with regularization.

$$L_c(\beta) = \prod_{i:E_i=1} \frac{\exp(\hat{h}_\beta(x_i))}{\sum_{j\in\Re(T_i)} \exp(\hat{h}_\beta(x_j))} \qquad \textbf{(2. 2)}$$

where $N_{E=1}$ is the number of patients with an observable event and λ is the ℓ_2 regularization parameter. Further, the authors used gradient descent optimization to determine the weights of the network which minimize (2.3).

$$\ell(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}_\theta(x_i) - log \sum_{j\in\Re(T_i)} \exp\left(\hat{h}_\theta(x_j)\right) + \lambda \cdot \|\theta\|_2{}^2 \qquad \textbf{(2. 3)}$$

Nnet-survival. Gensheimer et al. described Nnet-survival, a discrete-time survival model that is theoretically justified, naturally deals with non-proportional hazards, and can be trained rapidly by mini-batch gradient descent. Their proposed model uses naturally incorporated time-varying baseline hazard rate and non-proportional hazards provided each time interval output node is fully connected to the neurons of the last hidden layer. Further, a loss function was used, which is the negative of the log likelihood function of a statistical survival model.

$$\boldsymbol{\sum_{i=1}^{d_j} \ln\left(h_j^i\right) + \sum_{i=d_j+1}^{r_j} \ln\left(1 - h_j^i\right)} \qquad \textbf{(2. 4)}$$

Random survival forest. Ishwaran et al. introduced the random survival forests, an ensemble tree method for analysis of right-censored survival data, which is an extension of random forests to right-censored survival data [35, 36]. A high-level description of the algorithm is as follows:

1. B bootstrap samples are drawn from the original data. Each bootstrap sample excludes on average 37% of the data, referred to as out-of-bag data (OOB data).
2. A survival tree for each bootstrap sample is grown. Thereafter, at each node of the tree, p candidate variables are selected randomly. Subsequently, the node is split using the candidate variable that maximizes survival difference between daughter nodes.
3. The tree is grown to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths.
4. A cumulative hazard function (CHF) for each tree is calculated, and subsequently the average is estimated to obtain the ensemble CHF.

5. Using OOB data, the prediction error for the ensemble CHF is calculated.

6. The central elements to the random survival forest algorithm involve growing a survival tree and constructing the ensemble CHF. A good split for a node maximizes survival difference between daughters, which causes the tree to push dissimilar cases apart. Eventually, with increase in the number of nodes and separation of dissimilar cases, each node in the tree becomes homogeneous and is populated by cases with similar survival. Further, the survival tree eventually reaches a saturation point when no new daughters can be formed because of the criterion that each node must contain a minimum of $d_0 > 0$ unique deaths. Thus, the most extreme nodes in a saturated tree are called terminal nodes. $(T_{1,h}, \delta_{1,h}), \dots, (T_{n(h),h}, \delta_{n(h),h})$ are the survival times and the 0–1 censoring information for individuals (cases) in a terminal node $h \in \mathcal{T}$. $d_{l,h}$ and $Y_{l,h}$ are defined as the number of deaths and individuals at risk at time $t_{l,h}$. Thus, the CHF estimate for h is the Nelson–Aalen estimator.

$$\widehat{H}_h(t) = \sum_{t_{l,h} \le t} \frac{d_{l,h}}{Y_{l,h}} \qquad \textbf{(2. 5)}$$

Each case i has a d-dimensional covariate x_i. The CHF for i is the Nelson–Aalen estimator for x_i's terminal node:

$$H\,(t|x_i) = \widehat{H}_h(t), \;\; if\; x_i \in h \qquad \textbf{(2. 6)}$$

The CHF (2.6) is derived from a single tree. However, to compute an ensemble CHF, both an OOB and bootstrap over B survival trees estimate are required. Thus, $I_{i,b} = 1$ is defined provided i is an OOB case for b; otherwise, $I_{i,b} = 0$. (2.7) is an average over bootstrap samples wherein i is OOB.

$$H_e^{**}\,(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^{*}\,(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}, \;\; if\; x_i \in h \qquad \textbf{(2. 7)}$$

The bootstrap ensemble CHF for i, where all survival trees are used and not just those where

i is OOB, is expressed in (2.8). Thus, Random survival forest uses both (2.7) and (2.8) to define a predicted outcome.

$$H_e^*\ (t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*\ (t|x_i) \qquad \textbf{(2. 8)}$$

Under general conditions, conservation of events implies that the sum of the estimated CHF over observed time (both censored and uncensored) equals the total number of deaths. This is applicable to a wide collection of estimators, including the Nelson–Aalen estimator. Thus, $\sum_{i=1}^{n(h)} \widehat{H}_h(T_{i,h}) = \sum_{i=1}^{n(h)} \delta_{i,h}$ for each terminal node $h \in \mathcal{T}$. In other words, the total number of deaths is conserved within h.

2.4. Evaluation metrics

2.4.1. Multi-class classification

Following metrics are the most common and basic measures for classification problems.

- **Accuracy**: This is the most common and simplest measure for evaluating a classifier. It is simply defined as the degree of right predictions of a model.

$$Accuracy\ (y,\ \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1\ (\hat{y_i} = y_i)$$

- **Precision**: Precision, also referred to as positive predictive value, is the number of true positive results divided by the number of all positive results, including those not identified correctly.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: Recall, also referred to as sensitivity, is the number of true positive results divided by the number of all samples that should have been identified as positive.

$$Recall = \frac{TP}{TP + FN}$$

- **F-score ($\mathbf{F_1}$ score)**: This is the harmonic mean of precision and recall of the test. The highest possible value of an F-score is 1.0, which indicates perfect precision and recall, while the lowest possible value is 0 in case either the precision or the recall is zero.

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}\,(FP + FN)}$$

- **Cohen's kappa score**: This measures the agreement between two labels from a classifier and a ground truth, which classify N items into C mutually exclusive categories. The kappa score is a number between -1 and 1. In general, a score above 0.8 is considered as good agreement whereas zero or lower implies no agreement (practically random labels). Further, p_o is the relative observed agreement among labels by a classifier and a ground truth, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of randomly seeing each category.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

- **Matthew's correlation coefficient (MCC)**: This considers true and false positives and negatives and is generally regarded as a balanced measure, which can be used even if the classes are of varied sizes. The MCC is in essence a correlation coefficient value between -1 and +1, where a value of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. In addition, in the multi-class case, the MCC can be defined in terms of a confusion matrix C for K classes. To simplify the definition, consider the following intermediate variables:

$t_k = \sum_i^K C_{ik}$ the number of times class k truly occurred,

$p_k = \sum_i^K C_{ki}$ the number of times class k was predicted,

$c = \sum_k^K C_{kk}$ the total number of samples correctly predicted,

$s = \sum_i^K \sum_j^K C_{ij}$ the total number of samples.

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

When more than two labels exist, the value of the MCC no longer ranges between -1 and +1. Instead, the minimum value lies in the range of -1 and 0 based on the number and distribution of ground true labels. The maximum value is always +1.

In the evaluation for multi-class classification, there are three average calculation methods that are employed to evaluate the accuracy of classification as shown in Table 2.1.

- y the set of predicted (sample, label) pairs
- $\hat{y}$ the set of true (sample, label) pairs
- L the set or labels
- $y_l, \hat{y}_l$ the subset of $y, \hat{y}$ with label l
- $P(A, B) = \frac{|A \cap B|}{|A|}$
- $R(A, B) = \frac{|A \cap B|}{|B|}$
- $F_1(A, B) = 2 \cdot \frac{P(A,B) \times R(A,B)}{P(A,B) + R(A,B)}$

Table 2.1. Definition of evaluation metrics

Average	Precision	Recall	F-score
Micro	$P(y,\ \hat{y})$	$R(y, \hat{y})$	$F_1(y, \hat{y})$
Macro	$\frac{1}{\lvert L\rvert}\sum_{l \in L} P(y_l, \hat{y}_l)$	$\frac{1}{\lvert L\rvert}\sum_{l \in L} R(y_l, \hat{y}_l)$	$\frac{1}{\lvert L\rvert}\sum_{l \in L} F_1(y_l, \hat{y}_l)$
Weighted	$\frac{1}{\sum_{l \in L}\lvert\hat{y}_l\rvert}\sum_{l \in L}\lvert\hat{y}_l\rvert P(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L}\lvert\hat{y}_l\rvert}\sum_{l \in L}\lvert\hat{y}_l\rvert R(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L}\lvert\hat{y}_l\rvert}\sum_{l \in L}\lvert\hat{y}_l\rvert F_1(y_l, \hat{y}_l)$

Micro-average is the calculation of metrics globally by counting the total true positives, false negatives, and false positives. In contrast, macro-average involves calculating metrics for each label and determining their unweighted mean, and label imbalance is not considered. Furthermore, weighted-average is the calculation of metrics for each label, and then determining their average weighted by support (the number of true instances for each label). This changes the ‘macro’ method to incorporate label imbalance. Moreover, if all labels are included, “micro” averaging in a multi-class setting results in precision and recall, which are all identical to accuracy. In addition, “weighted” averaging may produce an F-score that is not between precision and recall.

Ferri et al. demonstrated that to measure the quality of models that might be affected by classes with a very low percentage of elements, the area under the ROC curve-based measures and macro-averages is not useful. This is because the global measure is heavily influenced by a poor assessment of an infrequent class [37], which is because a consequence of these measures assigning equal value to all classes independent of their frequency. In contrast, accuracy, mean F-measure, Kappa statistic, log loss, calibration loss, and mean squared error assign a relevance to each class, which is proportional to its frequency.

2.4.2. Time-to-event prediction

Time-to-event analysis refers to the analysis of the length of time until the occurrence of the event of interest [33], as explained earlier and denoted as S(t, X), where X = $(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p})$ indicates the patient characteristics that are input to the model. For example, Nnet-survival provides the predicted S(t, X) at multiple specified time points as the model output [38]. Another example is the random survival forest, which estimates the cumulative hazard.

- **C-index**: The most frequently used evaluation metric of survival models is the concordance index (c index, c statistic). It is a measure of rank correlation between predicted risk scores $\hat{f}$ and observed time points y that is closely related to Kendall's τ, and is defined as the ratio of correctly ordered (concordant) pairs to comparable pairs. Two samples i and j are comparable if the sample with lower observed time y experiences an event, that is, if $y_j > y_i$ and $\delta_i = 1$, where δ_i is a binary event indicator. In addition, a comparable pair (i, j) is concordant if the estimated risk $\hat{f}$ by a survival model is higher for subjects with lower survival time, that is, $\widehat{f_i} > \widehat{f_j} \land y_j > y_i$, otherwise the pair is discordant. Harrell's C index discards the pairs that are incomparable because of censoring when computing the index value. Although easy to interpret and compute, Harrell's concordance index has been shown exhibit excessively optimistic characteristics with increasing amount of censoring [39].

- **Time-dependent Area under the ROC (AUC)**: The area under the ROC curve is a popular performance measure for binary classification task. In the medical domain, it is often used to determine the degree to which the estimated risk scores can separate diseased patients (cases) from healthy patients (controls). For a particular predicted risk score $\hat{f}$, the ROC curve compares the false positive rate (1 - specificity) against the true positive rate (sensitivity) for each possible value of $\hat{f}$. When extending the ROC curve to continuous outcomes, in particular survival time, the disease status of a patient is typically not fixed and changes over time: at the time of enrollment a subject is usually healthy but may be diseased at a later time point. Consequently, the sensitivity and specificity must be considered as time-dependent measures.

We consider cumulative cases and dynamic controls at a given time point t, which results in time-dependent cumulative/dynamic ROC at time t. Cumulative cases all indicate individuals who experienced an event prior to or at time t $(t_i \leq t)$, whereas dynamic controls are those with $t_i > t$. Computing the area under the cumulative/dynamic ROC at time t aids in determining the degree to which a model can distinguish subjects who fail by a given time $(t_i \leq t)$ from subjects who fail after this time $(t_i > t)$. Hence, it is considered most relevant if one desires to predict the occurrence of an event in a period up to time t rather than at a specific time point t. Thus, for a particular estimator of the i-th individual's risk score $\hat{f}(x_i)$, the cumulative/dynamic AUC at time t is defined as

$$\widehat{\mathrm{AUC}}(\mathrm{t}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I\left(y_j > t\right) I(y_i \leq t) \omega_i I(\hat{f}\left(x_j\right) \leq \hat{f}(x_i))}{(\sum_{i=1}^{n} I(y_i > t))(\sum_{i=1}^{n} I(y_i \leq t))\, \omega_i)}$$

where ω_i are inverse probability of censoring weights. The function also provides a single summary measure that refers to the mean of the $\mathrm{AUC(t)}$ over the time range (τ_1, τ_2).

$$\overline{\mathrm{AUC}}(\tau_1, \tau_2) = \frac{1}{\hat{S}(\tau_1) - \hat{S}(\tau_2)} \int_{\tau_1}^{\tau_2} \widehat{\mathrm{AUC}}(\mathrm{t})\, d\hat{S}(t)$$

- **Integrated Brier Score (IBS)** : The time-dependent Brier score is the mean squared error at time point t [40].

$$BS^c(t) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq t \wedge \delta_i = 1)\ \frac{\left(0 - \hat{\pi}(t|x_i)\right)^2}{\hat{G}(y_i)} + I(y_i > t) \frac{\left(1 - \hat{\pi}(t|x_i)\right)^2}{\hat{G}(t)}$$

where $\hat{\pi}(t|x)$ is the predicted probability of remaining event-free up to time point t for a feature vector x, and $1/\hat{G}(t)$ is an inverse probability of censoring weight, estimated using the Kaplan-Meier estimator. The IBS provides an overall calculation of the model performance at all available times $t_1 \leq t \ \leq t_{max}$. The integrated time-dependent Brier score over the interval $[t_1; t_{max}]$ is defined as

$$\mathrm{IBS} = \int_{t_1}^{t_{max}} BS^c(t)\, dw(t)$$

where the weighting function is $w(t) = t/t_{max}$. The integral is estimated via the trapezoidal rule.

3. Machine learning based CDSS with 20 clinical variables in patients with hepatocellular carcinoma

In this Chapter, first, the motivation of the CDSS for patients with HCC has been demonstrated. In addition, characteristics of internal dataset and previous two-stage model including the model for treatment recommendation and survival prediction after initial treatment are described. In a previous study with dataset of a single center, we employed the cascaded random forest and random survival forest models for treatment recommendation and survival prediction, respectively. In addition, the process of feature selection of 20 pretreatment key variables as a model input and performance of previous models has been demonstrated. Moreover, at the end of the chapter, data collection and characteristics of external datasets from eight institutions in South Korea have been presented.

3.1. Motivation

Primary liver cancer is the sixth most commonly diagnosed cancer and the third leading cause of cancer death worldwide in 2020 [41]. HCC comprises 75-85 % of primary liver cancer. The American Association for the Study of Liver Diseases and the European Association for the Study of the Liver currently endorse the BCLC staging system as a primary prognostic model and an allocating tool of HCC treatment [42, 43]. However, there exists a significant discrepancy in the initial treatment choice for HCC between the recommendations obtained from the BCLC system and real clinical practice [44, 45]. This is partially because treatment decision for HCC is highly multifactorial, wherein physicians must consider the HCC stage, baseline liver function, and performance status. Moreover, other factors such as location and distribution of tumor, presence of intermediate nodule, comorbidities, socio-economic status, availability of potential living related-donors, and the invasiveness and feasibility of each treatment option play critical roles in determining the clinical outcomes of patients with HCC. Consequently, such complex nature of HCC treatment decision has hindered large-sized clinical studies, because conventional statistical methods cannot aptly control multiple variables and factors.

Recent attempts on applying the AI technique to clinical practice have focused on using AI to develop CDSS [38, 46-48]. In our previous study, we developed CDSS that can evaluate multiple pretreatment variables to recommend optimal treatment options for HCC and also predict the overall survival of patients after treatment [49]. To evaluate the performance and expandability of this model, we collected the external dataset from eight other institutions in South Korea and investigated the performance.

3.2. Related work

3.2.1. Data collection for internal dataset

In our previous study, we retrospectively reviewed hospital records of 1,650 consecutive patients who were newly diagnosed with HCC at Asan Medical Center (Seoul, Korea) between January and October 2010. Patients who had a treatment history of HCC (N = 356), received HCC treatment at other hospitals (N = 138), had a metastatic liver cancer (N = 71), had secondary malignancies that might affect survival (N = 36), had combined hepatocellular cholangiocarcinoma (N = 21), and cases involving incidentally detected HCC after transplantation (N = 7) were excluded from the study. Consequently, the study cohort included 1,021 patients with HCC. All enrolled patients were diagnosed with HCC through a liver protocol computed tomography, or magnetic resonance imaging, or liver biopsy according to the current guidelines of the American Association for the Study of Liver

Diseases [50]. Thereafter, patients were randomly allocated to the derivation or validation set at a ratio of 4:1.

We used our institutional database to collect information regarding the initial treatment option, initial treatment response, and overall survival of all patients. Pre-treatment demographic, clinical, and imaging variables, treatment information, and survival status of all the 1,021 patents were retrospectively collected from the database of our center. Subsequently, the following demographic factors were assessed: age, sex, Eastern Cooperative Oncology Group (ECOG) score, etiology of liver disease, presence of potential liver-related donor, body mass index (BMI), occupation, resident area, educational attainment of patient, maximum tumor size, tumor number, tumor type (infiltrative or nodular), tumor enhancement pattern, tumor distribution, portal vein invasion, hepatic vein or inferior vena cava invasion, bile duct invasion, extrahepatic metastases, presence of dysplastic nodule, radiofrequency ablation (RFA) feasibility, presence of cirrhosis, Child–Pugh class, presence of varix, laboratory findings including alpha-feto protein (AFP) level, within or above the Milan criteria, initial treatment option, initial treatment response, and overall survival. Further, RFA feasibility was defined as a size or location of the tumor to facilitate the successful receival of percutaneous RFA without significant complications, evaluated by a single hepatologist, G.H.C. However, any tumor located adjacent to a large vessel, bile duct, hepatic hilum, liver capsule, or extrahepatic organ was classified as an RFA non-feasible lesion. Overall, the survival was defined as the time form date of imaging diagnosis of HCC to the date of death due to any cause.

Table 3.1. Initially assembled 61 pretreatment variables

Patient variables (***N*** **= 30)**	**Laboratory variables** (***N*** **= 13)**	**Tumor variables** (***N*** **= 18)**
Epidemiology	WBC count	Enhancement pattern
Sex	Haemoglobin	Tumor type
Age	Platelet count	Tumor number
Performance status (ECOG)	PT (INR)	Maximal tumor diameter
Body mass index	Creatinine	Tumor distribution
Aetiology	Estimated glomerular filtration rate	RFA feasibility*
Alcohol history		Presence of dysplastic nodule
Amount of alcohol intake	Albumin	Presence of portal vein invasion
Smoking history	AST	Location of portal vein invasion
HBsAg	ALT	Presence of hepatic vein invasion
HBeAg	Total bilirubin	Presence of IVC invasion
HBeAb	AFP	Presence of bile duct invasion
HBV DNA	PIVKA-II	Presence of metastasis
History of HBV Treatment	ICG test	Presence of clinically significant metastasis
HCV Ab		
HCV RNA		Location of metastasis
History of HCV treatment		BCLC stage
Liver cirrhosis-related		Milan criteria
Child-Pugh class		Asan criteria
Varix		
Ascites		
Hepatic encephalopathy		
Presence of splenomegaly		
Accompanying comorbidities		
Hypertension		
Diabetes mellitus		
Dialysis		
Heart disease		
Pulmonary disease		
Socio-economic status		
Marriage		
Potential donor		
Occupation		
Education		
Residence area		

Abbreviations: AFP, alpha-fetoprotein; ALT, alanine transaminase; AST, aspartate transaminase; BCLC, Barcelona clinic liver cancer; ECOG, Eastern Cooperative Oncology Group; HBsAg, hepatitis B surface antigen; HBeAg, hepatitis B envelope antigen; HbeAb, hepatitis B envelope antibody; HBV, hepatitis B virus; HCV, hepatitis C virus; ICG, indocyanine green; INR, international normalized ratio; IVC: inferior vena cava; PIVKA-II, protein induced by vitamin K absence or antagonist II; PT, prothrombin time; RFA, radiofrequency ablation; WBC, white blood cell.

*RFA feasibility was defined as a size or location of the tumor to receive percutaneous RFA successfully without significant complication.

3.2.2. Feature selection: 20 key variables

Among the 61 initial pretreatment variables, 20 key variables (Table 3.2) were selected based on the importance scores calculated using the automated classifier and survival prediction models in the derivation set. Specifically, 14 variables were patient-related factors (age, BMI, Child–Pugh class, presence of varix, presence of ascites, ECOG score, hemoglobin level, platelet count, albumin level, prothrombin time, alanine aminotransferase [ALT] level, total bilirubin level, creatinine level, and AFP level), and the remaining 6 were tumor-related factors (tumor number, maximal tumor size, tumor distribution, presence of portal vein invasion, presence of metastasis, and RFA feasibility). Furthermore, treatment options were classified as follows: transplantation, surgical resection, RFA or percutaneous ethanol injection therapy (PEIT), trans-arterial chemoembolization (TACE), TACE combined with external beam radiotherapy (EBRT), sorafenib treatment, supportive care, and other therapies, such as combined therapy (e.g., surgical resection with intraoperative RFA, TACE combined with sorafenib), palliative resection, intra-arterial cytotoxic chemotherapy, clinical trials, and EBRT alone.

Table 3.2. 20 Pretreatment key variables

Patient related factors (14)	
Age	Value
Body mass index, kg/m^2	Value
ECOG Performance status	0, 1, 2, 3, 4
Child-Pugh score	5 – 14
Varix	Absence / Presence
Ascites	Absence / Controlled uncontrolled
AFP, *ng/mL*	Value
Hemoglobin, *g/dL*	Value
Platelet count, *x10^9/mm^3*	Value
ALT, *U/L*	Value
Total bilirubin, *mg/dL*	Value
Albumin, *mg/dL*	Value
Prothrombin time, *INR*	Value
Creatinine, *mg/dL*	Value
Tumor related factors (6)	
Tumor number	1, 2, 3, 4 or more
Maximum tumor size, cm	Value
Distribution	Single segmental / Unilobal / Bilobal
Portal vein invasion	Absence / Unilateral / Main portal or both portal vein
Metastasis	Absence / Presence
RFA feasibility*	Feasible / Non-feasible

Abbreviations: AFP, alpha-fetoprotein; ALT, alanine aminotransferase; ECOG, Eastern Cooperative Oncology Group; RFA, radiofrequency ablation

* RFA feasibility is defined as a size or location of the tumor to receive percutaneous RFA successfully without significant complication

3.2.3. Model training and tunning

Using 20 key variables, the random forest and random survival forest methods were trained and evaluated again to recommend treatment options and predict overall survival in both the derivation and validation sets. Consequently, the primary outcomes were accuracies of treatment recommendation and survival prediction. The index date is defined as the date that patients undergo their first liver protocol computed tomography or magnetic resonance imaging, and the follow-up period for each patient is estimated from this date to the date of death or the last follow-up date. However, owing to the large differences in survival between treatments, training a machine learning-based model of treatment recommendation and survival prediction in an integrated way is challenging. Therefore, separate and training were conducted for the treatment recommendation and survival prediction models.

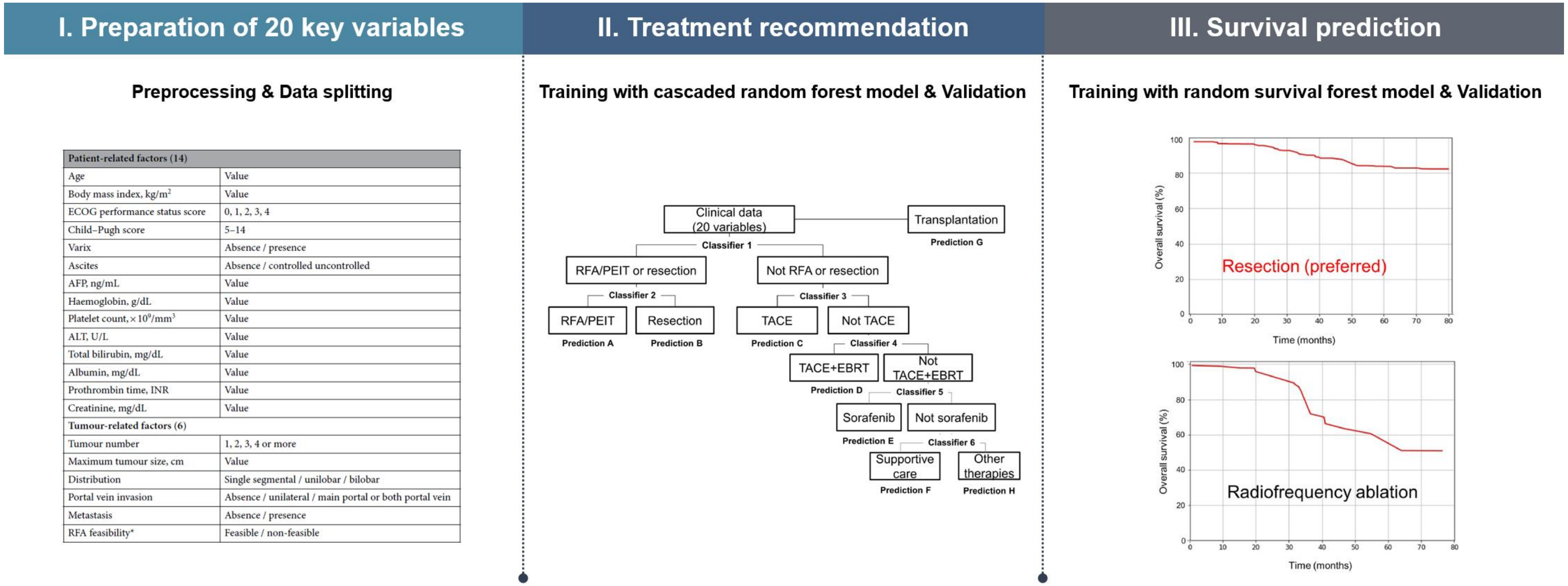
Treatment recommendation models were hierarchically designed using six classifiers similar to treatment planning in clinical practice. A supervised learning methodology was adopted to prefer curative modalities using a classifier method. Further, the transplantation option was not considered in the treatment decision algorithm because of the medical environment of severe shortage of deceased liver donor. However, although not included in the classifier model, transplantation was suggested as an option, provided the Millan criteria was satisfied. Moreover, because factors affecting the prognosis were different for each treatment, separate survival prediction models were developed for each treatment. In addition, the proposed CDSS system was operated by sequentially using treatment recommendation and survival prediction models.

The treatment recommendation and survival prediction model was developed employing the random forest model. Random forest, a representative ensemble method, is widely used because it is powerful and relatively lighter than other ensemble methods [51, 52]. It constructs several tree-type base models and forms an ensemble through a technique referred to as bootstrap aggregating or bagging. Further, Gini impurity and log-rank test were used as the splitting rules for random forests, in case of treatment recommendation and survival prediction models, respectively. In addition, other possible combinations of hyperparameters of models were investigated via a grid search using GridSearchCV library in Scikit-learn package.

Figure 3.1 shows the schematic diagram for the construction of the CDSS for HCC. The model comprised six multi-step classifiers and seven survival prediction sub-models. The input variables (N = 20) were processed with the algorithm for treatment recommendation with multi-step classifiers. The model for HCC was designed with preference for curative modalities (transplantation, resection, RFA or PEIT). Upon the selection of a treatment option, the model demonstrates the predicted survival curve for each patient. Additionally, if another treatment option is available, the model can suggest another predicted survival curve after the alternative treatment. (Table 3.3) Therefore, the model can predict different survival curves of the same patient with different treatments, which is expected to aid clinicians make treatment decisions in actual clinical setting.

Table 3.3. Rules for alternative treatment options

Prediction	Alternative treatment options
RFA	TACE
Op	1) RFA feasibility = feasible → RFA 2) Portal vein invasion = Absence → TACE 3) TACE+RT
TACE	No alternative option
TACE+RT	Sorafenib, None
Sorafenib	None
None	No alternative option



I. Preparation of 20 key variables

Preprocessing & Data splitting

Patient-related factors (14)	
Age	Value
Body mass index, kg/m^2	Value
ECOG performance status score	0, 1, 2, 3, 4
Child–Pugh score	5–14
Varix	Absence / presence
Ascites	Absence / controlled uncontrolled
AFP, ng/mL	Value
Haemoglobin, g/dL	Value
Platelet count, × 10^9/mm^3	Value
ALT, U/L	Value
Total bilirubin, mg/dL	Value
Albumin, mg/dL	Value
Prothrombin time, INR	Value
Creatinine, mg/dL	Value
Tumour-related factors (6)	
Tumour number	1, 2, 3, 4 or more
Maximum tumour size, cm	Value
Distribution	Single segmental / unilobar / bilobar
Portal vein invasion	Absence / unilateral / main portal or both portal vein
Metastasis	Absence / presence
RFA feasibility*	Feasible / non-feasible

II. Treatment recommendation

Training with cascaded random forest model & Validation

III. Survival prediction

Training with random survival forest model & Validation

Figure 3.1. Overall architecture of CDSS

3.2.4. Evaluation metrics

Baseline characteristics of the patients were compared using the chi-square and Mann–Whitney U tests for categorical and continuous variables, respectively. Survival distributions were compared using the Kaplan–Meier method with a log-rank test. Patients in the follow-up program who were not confirmed deceased were recorded as censored. In the initial phase of model development, a univariate Cox proportional hazards model was fitted to the treatment decision and survival endpoints and the selection of variables was realized by employing a two-step variable selection approach. The first step involved fitting a random forest model to compute a variable importance score, and thereafter the second step was to compute a relative selection frequency based on a bootstrap resampling method [53, 54]. Moreover, for the validation data sets, per-patient based analysis was performed from probability values using accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for each classifier. The accuracy was defined as the percentage of correctly classified instances and calculated as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}),$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively. Each survival prediction model was validated using bootstrapping to correct for optimistic bias. Further, time-dependent concordance (C)-index was used to evaluate predicted survival times, which were ranked based on the observed survival times. All P-values were two-sided, and $P < 0.05$ was considered significant. Further, the outcome of implicit feature selection of the random forest was visualized using the Gini importance. SPSS version 21 (SPSS, Inc., Chicago, IL), open-source Scikit-learn package in python version 0.19.1 [55], and random Forest SRC package in R version 3.4.1 (R Core Team, Vienna, Austria) [35] were utilized for statistical analyses.

3.2.5. Results

We trained our CDSS system using the derivation set (N = 813) and validated it in the validation set (N = 208). Two sets were divided via stratified random splits. Consequently, the same derivation and validation sets were used for both treatment recommendation and survival prediction models, respectively. Thus, a total of 460 and 128 patients died during the median follow-up periods of 37.8 (interquartile range [IQR], 8.3–84.7) and 48.6 (IQR, 8.3–83.1) months, respectively. The baseline demographics of patients are summarized in Table 3.4. Of the total 1,021 patients (mean age, 56.9 years), 81.8% were male, and 77.0% had positive hepatitis B virus surface antigen. Moreover, 76.3% of patients were classified with Child–Pugh class A, and 75.1% had an ECOG score of 0. Regarding tumor-related factors, 41.7% of patients had multiple tumors, and the median maximal tumor diameter was 4.0 cm (IQR 2.3–8.5). Portal vein invasion and distant metastasis were confirmed in 22.8% and 12.2% of patients, respectively. BCLC stages 0, A, B, C, and D were observed in 13.4%, 26.0%, 18.0%, 36.6%, and 6.3% of patients, respectively. As an initial treatment, transplantation was performed in 4.5%, resection in 32.9%, RFA or PEIT in 7.5%, TACE in 31.5%, TACE combined with EBRT in 6.6%, sorafenib treatment in 3.0%, and supportive care in 10.1% of patients. Additionally, 3.8% of patients underwent other therapies; nine patients underwent resection combined with intraoperative RFA, nine underwent palliative resection, eight underwent EBRT to liver, six underwent TACE combined with sorafenib or cytotoxic chemotherapy, and four underwent intra-arterial cytotoxic chemotherapy. Moreover, three patients were enrolled in clinical trials and underwent systemic therapy. There was no significant difference between the derivation and validation set with respect to patient-, tumor-, or treatment-related variables.

Table 3.4. Baseline characteristics of the patients, tumors, and initial treatment options

Characteristics		**All patients** **(n = 1021)**
Age, yr		56.9 ± 10.5
Gender	Male	835 (81.8)
	Female	186 (18.2)
ECOG Performance status	0	858 (84.0)
	1 or 2	123 (12.0)
	3 or 4	40 (3.9)
Etiology of liver disease	HBV	786 (77.0)
	HCV	71 (7.0)
	Others	164 (16.0)
Heavy alcohol consumption	Yes	168 (16.5)
Ascites	Present	173 (17.0)
Varices	Present	312 (30.6)
Child-Pugh class	A	779 (76.3)
	B	205 (20.1)
	C	37 (3.6)
Body mass index, kg/m^2		24.0 (22.1–26.0)
Tumor number	1	595 (58.3)
	2–3	217 (21.2)
	≥4	209 (20.5)
Maximal tumor size, cm		4.0 (2.3–8.5)
Distribution	Single segmental	475 (46.5)
	Unilobal	245 (24.0)
	Bilobal	300 (29.4)
Distant metastasis	Present	125 (12.2)
Vascular invasion	Unilateral	150 (14.7)
	Main or bilateral	83 (8.1)
RFA feasibility[†]	Feasible[†]	226 (22.1)
BCLC stage	0	131 (12.8)
	A	284 (27.8)
	B	228 (22.3)
	C	314 (30.8)
	D	64 (6.3)
Laboratory findings	AFP, *ng/mL*	42.1 (6.7–838.2)
	Hemoglobin, *g/dL*	13.5 (12.2–14.6)
	Platelet count, *x10^9/mm^3*	143 (97–197)
	ALT, *U/L*	37 (25–53)
	Total bilirubin, *mg/dL*	1.0 (0.7–1.4)
	Albumin, *mg/dL*	3.6 (3.2–4.0)
	Prothrombin time, *INR*	1.07 (1.01–1.17)
	Creatinine, *mg/dL*	0.8 (0.7–0.9)
Initial treatment	Transplantation	46 (4.5)
	Resection	336 (32.9)
	RFA or PEIT	77 (7.5)
	TACE	322 (31.5)
	TACE combined EBRT	67 (6.6)
	Sorafenib	31 (3.0)
	Supportive care	103 (10.1)
	Other therapies	39 (3.8)

Abbreviations: AFP, alpha-fetoprotein; ALT, alanine aminotransferase; BCLC, Barcelona clinic liver cancer; EBRT, external beam radiotherapy; ECOG, Eastern Cooperative Oncology Group; HBV, hepatitis B virus; HCV, hepatitis C virus; INR, international normalized ratio; PEIT, percutaneous ethanol injection; RFA, radiofrequency ablation; TACE, transarterial chemoembolization

*Variables are presented as mean±standard deviation or median (IQR)

[†] RFA feasibility is defined as a size or location of the tumor to receive percutaneous RFA successfully without significant complication

Table 3.5 lists the accuracy of the six classifier models trained from the derivation set. The recommended treatment from the model was compared with the treatment used in real clinical practice in the validation set. Overall, the proposed CDSS classifier model for HCC was well generalized and exhibited good performance, and its standard deviations were higher in the lower branches of the treatment (e.g., sorafenib treatment, supportive care, other therapies) owing to the number of patients being relatively smaller. The accuracies of classifiers 1, 2, 3, 4, and 5 were 81.0% (curative treatments versus not curative treatments), 88.4% (resection versus RFA/PEIT), 76.8% (TACE vs. or not TACE), 76.6% (TACE + EBRT versus not TACE + EBRT), 80.0% (sorafenib treatment versus not sorafenib treatment), and 80.1% (supportive care versus other therapies), respectively.

Table 3.5. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for 6 classifier model in validation set

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Classifier 1 (RFA/PEIT or resection vs. Not RFA/PEIT or resection)	81.0 ± 2.6	77.4 ± 4.1	83.7 ± 3.3	77.8 ± 3.6	83.5 ± 2.5
Classifier 2 (RFA/PEIT vs. Resection)	88.4 ± 3.1	56.2 ± 11.6	95.8 ± 2.7	76.8 ± 12.1	90.6 ± 2.3
Classifier 3 (TACE vs. Not TACE)	76.8 ± 2.9	82.3 ± 4.1	69.3 ± 5.5	78.3 ± 4.0	74.6 ± 4.9
Classifier 4 (TACE+EBRT vs. Not TACE+EBRT)	76.6 ± 4.7	43.9 ± 12.6	89.4 ± 3.9	61.6 ± 10.8	80.4 ± 4.3
Classifier 5 (Sorafenib vs. Not sorafenib)	80.0 ± 4.2	12.3 ± 13.3	95.0 ± 4.0	44.0 ± 37.7	83.1 ± 3.0
Classifier 6 (Supportive care vs. Others)	80.1 ± 6.3	53.0 ± 17.6	90.4 ± 5.2	67.7 ± 15.8	83.7 ± 5.6

Abbreviations: EBRT, external beam radiotherapy; NPV, Negative predictive value; PEIT, percutaneous ethanol injection; PPV, Positive predictive value, RFA, radiofrequency ablation; TACE, transarterial chemoembolization

Figure 3.2 shows predicted survival curves of each recommended treatment in the validation set. The 'Ground truth curves' represent the Kaplan–Meier survival curve of patients in the validation set in real clinical practice. Further, the C-index values for the derived models of overall survival for RFA/PEIT, resection, TACE, TACE+EBRT, sorafenib treatment, supportive care, transplantation, and other therapies were 0.725 (95% CI, 0.708–0.741), 0.695 (95% CI, 0.680–0.709), 0.803 (95% CI, 0.796–0.809), 0.676 (95% CI, 0.658–0.694), 0.684 (95% CI, 0.648–0.720), 0.710 (95% CI, 0.689–0.730), 0.959 (95% CI, 0.949–0.969), and 0.850 (95% CI, 0.835–0.884), respectively.

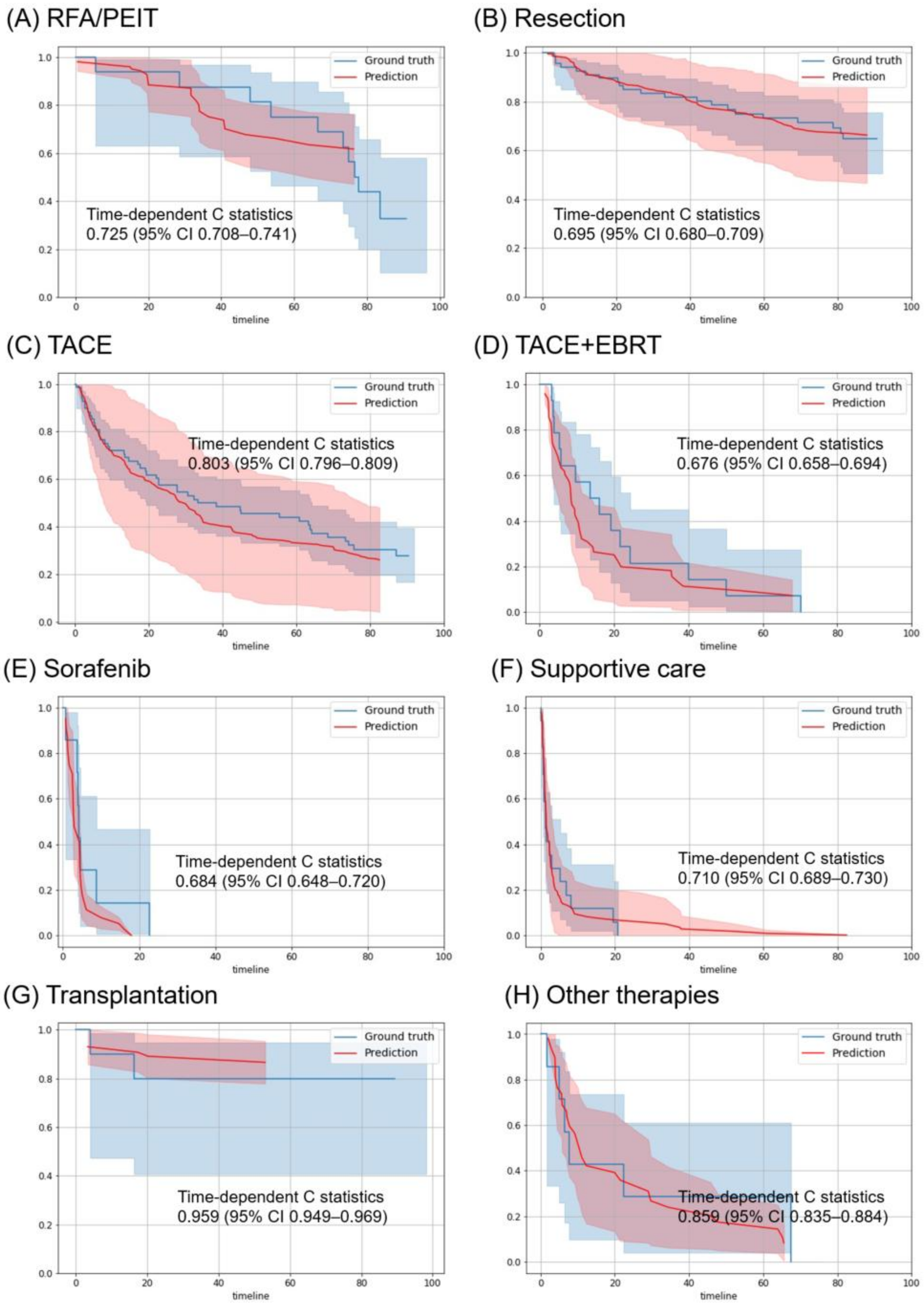


Figure 3.2. True and predicted overall survival according to the initial treatment in the validation set (A) RFA/PEIT. (B) Resection. (C) TACE. (D) TACE combined with EBRT. (D) Sorafenib. (E) Supportive care. (F) Transplantation.

3.3. Data collection for external dataset

The data of 2,685 consecutive patients who were newly diagnosed with HCC were collected from eight institutions, namely Korea University Guro Hospital (KUGH), Seoul National University Bundang Hospital (SNUBH), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Catholic Medical Center (CMC), Severance Hospital (SH), Chung-ang University Hospital (CUH), and Inha University Hospital (IUH), in South Korea between January 2010 and December 2012. Thereafter, the information on 20 key variables used in previous study, initial treatment option, and survival information were investigated. The overall survival was identically defined as the time form date of imaging diagnosis of HCC to the date of death owing to any cause. The initial treatment options were categorized into eight groups similar to that of an internal dataset: RFA/PEIT, surgical resection, TACE, TACE combined with EBRT, sorafenib treatment, supportive care, transplantation, and other therapies. However, owing to significant heterogeneity among centers in the group of other therapies, the option for other therapies were eliminated from choices of treatment recommendation. In addition, the transplantation option was also removed in the recommendation options for a reason similar to that in the previous study.

All enrolled patients were diagnosed with HCC using liver protocol computed tomography, or magnetic resonance imaging, or liver biopsy following the current guidelines of the American Association for the Study of Liver Diseases. Further, all external datasets were allocated to external validation set. As an initial treatment, RFA or PEIT was performed in 6.8%–21.6 %, resection in 3.7%–35.8%, TACE in 34.8%–64.3%, TACE combined with EBRT in 0%–24.4%, sorafenib treatment in 0%–7.4%, supportive care in 3.1%–16.7%, transplantation in 0%–12.8%, and other therapies in 0%–24.3% of patients in each center (Table 3.6 and Figure 3.3).

Table 3.6. Distribution of initial treatment in each center

Center	KUGH	SNUBH	SMC	SNUH	CMC	SH	AMC	CUH	IUH
RFA/PEIT	25 (18.1)	21 (10.9)	64 (14.6)	45 (20.1)	35 (21.6)	10 (6.8)	78 (8.3)	16 (9.4)	24 (8.7)
Resection	24 (17.4)	27 (14.0)	70 (15.9)	38 (17.0)	6 (3.7)	41 (27.7)	335 (35.8)	31 (18.1)	52 (18.9)
TACE	60 (43.5)	107 (55.4)	157 (35.8)	129 (57.6)	105 (64.8)	78 (52.7)	325 (34.8)	110 (64.3)	143 (52.0)
TACE+EBRT	0 (0.0)	6 (3.1)	107 (24.4)	0 (0.0)	1 (0.6)	0 (0.0)	65 (7.0)	0 (0.0)	0 (0.0)
Sorafenib	7 (5.1)	7 (3.6)	18 (4.1)	5 (2.2)	0 (0.0)	11 (7.4)	30 (3.2)	2 (1.2)	10 (3.6)
Supportive care	22 (15.9)	25 (13.0)	23 (5.2)	7 (3.1)	15 (9.3)	8 (5.4)	102 (10.9)	12 (7.0)	46 (16.7)
Transplantation	0 (0.0)	1 (0.5)	25 (5.7)	4 (1.8)	2 (1.2)	19 (12.8)	46 (4.9)	1 (0.6)	0 (0.0)
Other therapies	15 (10.9)	4 (2.1)	12 (2.7)	0 (0.0)	33 (20.4)	36 (24.3)	40 (4.3)	8 (4.7)	3 (1.1)
Total	**138**	**193**	**439**	**224**	**162**	**148**	**935**	**171**	**275**

Abbreviations: KUGH, Korea university guro hospital; SNUBH, Seoul national university bundang hospital; SMC, Samsung medical center; SNUH, Seoul national university hospital; CMC, Catholic medical center; SH, Severance hospital; AMC, Asan medical center; CUH, Chung-ang university hospital; IUH, Inha university hospital

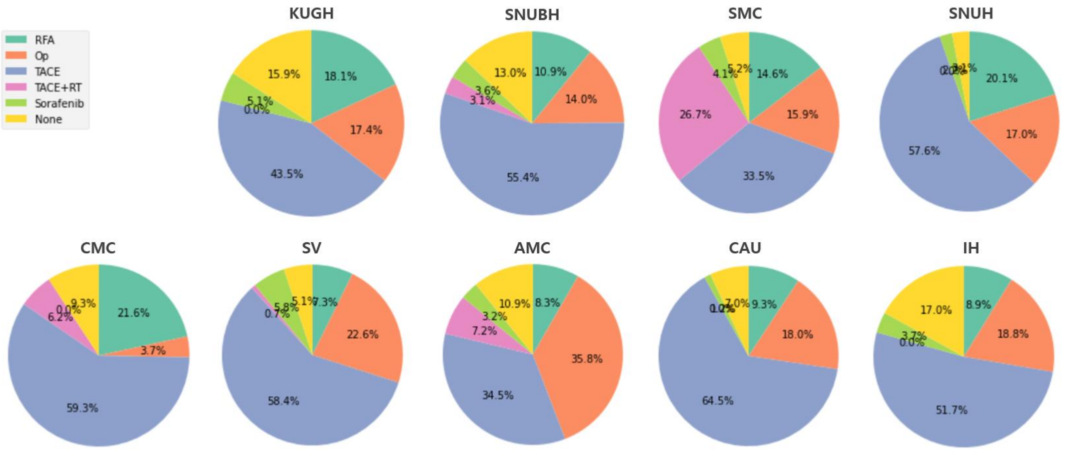


Figure 3.3. Distribution of initial treatment in each center.

4. Model for treatment recommendation

In this Chapter, various experiments related to the model for treatment recommendation aimed at improving performance and adjusting to multi-center dataset have been discussed. The use of ensemble voting machine and then comparing its performance with previous cascaded model are first described. Thereafter, various normalization and oversampling methods were applied to improve the model performance. Furthermore, individual training for each center and the option for a second treatment in addition to first treatment one was investigated to increase the accuracy in multi-center setting. Finally, the calibration of model and the results were described in this chapter.

Thus, the contribution of this work is following: 1) Accuracy was increased owing to use of ensemble voting machine compared with previous cascaded random forest model for internal dataset. 2) individual training for each center exhibited better performance than those of external validation, and for the option for second treatment in addition to first treatment option is suitable in multi-center setting.

4.1. Basic setup and notation

All experiments for internal dataset and individual training for external dataset were trained and validated with five-fold cross validation stratified by treatment. Further, experiments for external validation were performed with the model that was trained using the entire internal dataset. To evaluate the model and classify multiple treatments, the following metrics, which have been described in Chapter 2, were used: accuracy, macro-average of recall, weighted-average of precision, weighted-average F1 score, Kappa score, MCC

4.2. Experiments

4.2.1. Ensemble voting model

Nineteen different machine learning algorithms were evaluated for this task: logistic regression, decision tree classifier, extra-trees classifier, random forest classifier, Adaboost classifier, gradient boosting classifier, histogram-based gradient boosting classifier, Xgboost, light gbm, catboost, gaussian naive Bayes, naive Bayes classifier for multivariate Bernoulli models, gaussian process classification, linear discriminant analysis, quadratic discriminant analysis, C-support vector machine, multi-layer perceptron classifier, k-nearest neighbors classifier, and k-means clustering. Scikit-Learn's version 0.23.2, Xgboost's version 1.5.0, catboost's version 1.0.1, and lightgbm's version 3.2.1 were used to construct models. Moreover, all classifiers were trained and evaluated by employing a stratified five-fold cross-validation. Following each classifier being sorted based on mean accuracy, the voting classifier was trained using three, five, and seven top-performing classifiers and compared with each other including the top-performing classifier only. Finally, the performance of voting classifier compared to five top-performing classifiers were evaluated.

The performance of seven top-performing classifiers sorted based on accuracy is presented in Table 4.1. C-support vector machine with linear kernel demonstrated the highest mean accuracy of 65.45 and mean recall of 51.09, which was followed by the gaussian process classifier, random forest classifier, extra-trees classifier, histogram-based gradient boosting classifier, light gradient boosting machine, and multi-layer perceptron classifier.

Table 4.1. Top 7 classifiers sorted by accuracy for internal dataset

Model	Accuracy	Recall	Prec.	F1	Kappa	MCC
Linear SVM	65.45 (3.49)	51.09 (4.26)	64.79 (3.71)	64.32 (3.20)	51.41 (4.79)	51.75 (4.93)
GPC	64.71 (2.66)	48.92 (3.24)	63.05 (2.50)	63.22 (2.49)	50.02 (3.73)	50.41 (3.80)

RF	64.49 (5.17)	47.54 (5.18)	62.97 (6.04)	63.01 (5.19)	49.48 (7.27)	49.92 (7.53)
ET	64.49 (3.18)	49.83 (5.00)	63.29 (3.41)	63.37 (3.21)	50.07 (4.77)	50.37 (4.85)
HGBC	64.28 (2.99)	49.07 (5.62)	64.02 (3.58)	63.22 (2.80)	49.79 (4.32)	50.20 (4.54)
Light GBM	63.74 (3.05)	49.62 (4.26)	62.81 (2.71)	62.73 (2.65)	49.38 (4.17)	49.73 (4.41)
MLP	63.64 (3.04)	48.72 (2.88)	62.44 (2.10)	62.46 (2.58)	48.63 (4.31)	48.88 (4.35)

Abbreviations: Linear SVM, C-support vector machine with linear kernel; GPC, gaussian process classifier; RF, random forest classifier; ET, extra-trees classifier; HGBC, histogram-based gradient boosting classifier; Light GBM, light gradient boosting machine; MLP, multi-layer perceptron classifier

In case of three to seven top-performing classifiers, the C-support vector machine with linear kernel demonstrated inferior performance compared to voting classifier. Significant differences were not observed among voting classifiers with different numbers of top-performing classifiers (Table 4.2).

Table 4.2. Performance in terms of the number of classifiers composing voting classifier

Dataset	No. of classifiers	Performance					
		Accuracy	Recall	Precision	F1	Kappa	MCC
Internal dataset	Top 1	65.45 (3.49)	51.09 (4.26)	64.79 (3.71)	64.32 (3.20)	51.41 (4.79)	51.75 (4.93)
	Top 3	67.06 (2.16)	49.64 (2.88)	64.89 (2.01)	65.34 (1.96)	53.01 (2.97)	53.39 (3.04)
	Top 5	67.27 (2.94)	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
	Top 7	67.27 (2.96)	53.04 (3.94)	65.93 (2.96)	66.05 (2.78)	54.12 (3.95)	54.46 (4.10)

We compared cascaded method with ensemble voting model in the internal and external datasets. The performance of each model is shown in Table 4.3. In the internal dataset, mean accuracy was increased using the ensemble voting model. However, mean accuracy was slightly decreased in the external dataset, although it was not significant.

Table 4.3. Performance of ensemble voting classifier vs. cascaded random forest model

Dataset	Model	Performance

		Accuracy	Recall	Precision	F1	Kappa	MCC
Internal dataset	Cascaded	63.42 (3.74)	51.27 (2.51)	63.02 (3.52)	62.91 (3.54)	49.51 (4.92)	49.64 (4.96)
	Ensemble	**67.27 (2.94)**	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
External dataset	Cascaded	55.34 (6.09)	41.68 (3.88)	64.14 (5.11)	56.82 (4.61)	36.30 (7.89)	38.33 (7.38)
	Ensemble	**54.33 (4.29)**	42.56 (4.79)	66.07 (5.06)	56.51 (3.04)	36.68 (5.97)	39.16 (5.37)

4.2.2. Modification of internal dataset

To improve the performance of the model, we applied various normalization and oversampling methods to the internal and external datasets. For the normalization test, min-max normalization, z-score normalization, and robust normalization method ware applied and evaluated. Nevertheless, for the oversampling test, ROS and SMOTE were applied and evaluated (Table 4.4. and 4.5).

Table 4.4. Performance of various normalization methods

Dataset	Normalization	Performance					
		Accuracy	Recall	Precision	F1	Kappa	MCC
Internal dataset	Base	67.27 (2.94)	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
	Minmax	67.81 (3.13)	52.91 (5.39)	66.32 (2.93)	66.55 (2.86)	54.78 (4.50)	55.11 (4.63)
	Z-score	66.84 (3.17)	51.09 (4.49)	65.48 (2.93)	65.49 (2.94)	53.16 (4.53)	53.57 (4.69)
	Robust	67.06 (3.18)	50.89 (5.25)	65.06 (3.22)	65.32 (3.08)	53.30 (4.65)	53.78 (4.82)
External dataset	Base	55.34 (6.09)	41.68 (3.88)	64.14 (5.11)	56.82 (4.61)	36.30 (7.89)	38.33 (7.38)
	Minmax	55.04 (5.11)	41.24 (3.56)	64.35 (4.40)	56.71 (3.60)	36.01 (6.91)	38.11 (6.41)
	Z-score	55.20 (5.84)	41.84 (4.12)	64.42 (4.29)	56.67 (4.15)	36.10 (8.01)	38.12 (7.52)
	Robust	55.14 (5.63)	41.54 (3.79)	63.98 (4.73)	56.61 (4.03)	36.03 (7.46)	38.10 (6.98)

Table 4.5. Performance on various oversampling methods

Dataset	Oversampling	Performance					
		Accuracy	Recall	Precision	F1	Kappa	MCC
Internal dataset	Base	67.27 (2.94)	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
	ROS	65.67 (2.56)	55.41 (2.58)	66.59 (2.71)	65.30 (2.28)	53.12 (3.27)	53.53 (3.44)
	SMOTE	64.71 (3.19)	56.18 (2.68)	66.64 (3.39)	64.78 (2.94)	52.41 (3.96)	52.81 (4.07)

External dataset	Base	55.34 (6.09)	41.68 (3.88)	64.14 (5.11)	56.82 (4.61)	36.30 (7.89)	38.33 (7.38)
	ROS	56.09 (5.75)	47.82 (7.67)	66.44 (4.54)	58.01 (5.05)	39.20 (7.66)	41.55 (6.93)
	SMOTE	56.72 (5.94)	49.46 (8.12)	68.98 (4.03)	58.77 (5.37)	40.65 (7.71)	43.10 (6.86)

4.2.3. Individual training for external dataset

We trained and evaluated individually using the dataset of each center with a stratified 5-fold cross-validation. The five top-performing classifiers were sorted based on mean accuracy after testing 19 machine learning classifiers for each dataset, and different voting classifiers were trained and evaluated. Moreover, the voting classifier composed of the five top-performing classifiers for internal dataset was also trained with each dataset and evaluated for comparison. Tables 4.6 and 4.7 presents the performance of individual training with top 5 classifiers for each center and internal dataset, respectively. Almost no difference was observed in the mean accuracy, recall, and F1 score between two experiments.

Table 4.6. Performance of individual training with top five classifiers for each center

Center	Performance					
	Accuracy	Recall	Precision	F1	Kappa	MCC
KUGH	66.03 (5.81)	52.87 (3.49)	65.45 (8.95)	63.19 (6.77)	51.52 (7.54)	52.98 (7.13)
SNUBH	70.49 (3.68)	42.03 (3.63)	64.79 (6.79)	65.28 (5.73)	48.90 (6.58)	51.18 (5.88)
SMC	65.83 (4.30)	49.72 (4.97)	61.45 (4.75)	62.71 (4.32)	54.13 (5.80)	54.71 (5.83)
SNUH	62.10 (8.85)	39.64 (9.98)	59.00 (9.16)	59.92 (9.27)	31.67 (18.35)	32.06 (18.46)
CMC	71.57 (5.74)	45.33 (11.16)	68.47 (6.87)	68.58 (6.47)	40.38 (13.88)	42.25 (12.96)
SH	58.09 (6.31)	43.99 (11.70)	54.69 (6.82)	55.26 (6.87)	28.10 (12.93)	28.89 (12.85)
CUH	74.86 (6.58)	53.87 (12.96)	72.57 (9.60)	72.05 (7.61)	47.76 (15.14)	49.50 (14.90)
IUH	63.27 (5.91)	45.06 (7.13)	62.01 (7.52)	60.86 (7.36)	40.01 (12.00)	41.02 (11.73)
Average	**66.53 (5.90)**	**46.56 (8.13)**	63.55 (7.56)	**63.48 (6.80)**	42.81 (11.53)	44.07 (11.22)

Table 4.7. Performance of individual training with top 5 classifiers for internal dataset

Center	Performance					
	Accuracy	Recall	Precision	F1	Kappa	MCC
KUGH	66.67 (8.64)	53.20 (6.09)	66.66 (8.35)	63.96 (7.56)	51.50 (11.98)	53.06 (12.56)
SNUBH	69.43 (2.94)	40.52 (4.01)	63.17 (6.97)	63.96 (5.12)	45.61 (8.03)	48.82 (5.83)
SMC	65.61 (3.30)	49.18 (2.62)	61.85 (3.08)	62.15 (2.98)	53.74 (4.22)	54.50 (4.16)
SNUH	63.41 (6.87)	37.91 (7.94)	59.84 (6.97)	60.77 (7.06)	32.48 (13.88)	33.25 (14.00)
CMC	70.95 (7.23)	44.19 (12.06)	68.02 (9.20)	67.71 (8.02)	37.54 (18.21)	39.36 (17.91)
SH	56.76 (4.36)	33.32 (10.84)	48.98 (7.35)	51.11 (6.63)	22.65 (12.36)	23.87 (12.40)
CUH	76.05 (9.47)	54.78 (17.70)	73.64 (11.77)	72.90 (10.96)	49.90 (20.96)	51.63 (20.32)
IUH	65.45 (8.13)	42.99 (10.09)	61.57 (13.18)	61.13 (11.10)	39.80 (18.46)	41.56 (18.15)
Average	**66.79 (6.37)**	**44.51 (8.92)**	62.97 (8.36)	**62.96 (7.43)**	41.65 (13.51)	43.26 (13.17)

In the comparison of individual training and external validation, the accuracy and recall of the former was higher than those of the latter as shown in Table 4.8. Further, mean accuracies of all centers with individual training were improved compared to those of external validation except for one center in SH. Figure 4.1 shows the mean accuracies according to number of patients, where no clear trend in accuracy was indicated with increase in the number of patients required to train the model.

Table 4.8. Performance of individual training vs. external validation

Center	**No. of patients**	**Individual training**		**External validation**	
		Accuracy	**Recall**	**Accuracy**	**Recall**
KUGH	138	66.03 (5.81)	52.87 (3.49)	61.59	44.77
SNUBH	193	70.49 (3.68)	42.03 (3.63)	56.99	38.78
SMC	439	65.83 (4.30)	49.72 (4.97)	52.62	44.09
SNUH	224	62.10 (8.85)	39.64 (9.98)	52.68	44.69
CMC	162	71.57 (5.74)	45.33 (11.16)	43.21	39.62
SH	148	58.09 (6.31)	43.99 (11.70)	60.14	38.05
CUH	171	74.86 (6.58)	53.87 (12.96)	63.16	47.65
IUH	275	63.27 (5.91)	45.06 (7.13)	52.36	35.8
Average	219	**66.53 (5.90)**	**46.56 (8.13)**	55.34 (6.09)	41.68 (3.88)

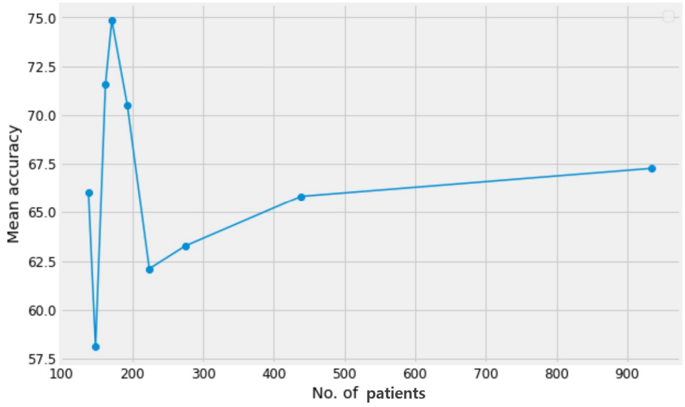


Figure 4.1. Mean accuracy according to number of patients

4.2.4. Suggestion of second recommendation

The first and second treatment options were derived from the treatment with the highest and second highest probabilities in voting classifier, respectively. All evaluation metrics were measured again by including the second treatment option as the correct answer. Table 4.9 shows the performance of previous cascaded random forest model and current ensemble voting machine in both internal and external datasets. Moreover, in the experiments for internal dataset, when only the first treatment option was accepted as the correct answer, the mean accuracies of cascaded random forest model and ensemble voting model were 63.42% and 67.27%, respectively. In contrast, in case the second treatment option was regarded as the correct answer, the mean accuracy of the ensemble voting model exhibited an increase of 87.27%, compared to 73.69% in the cascaded model. Thus, the results of external datasets show a bigger difference between two models. In addition, in case the second treatment option was regarded as the correct answer, the mean accuracy of the ensemble voting model increased significantly to 86.06%, while the accuracy of the cascaded model barely increased.

Table 4.9. Comparison of performance considering second options

Dataset	Model	Options	Performance					
			Accuracy	Recall	Precision	F1	Kappa	MCC
Internal dataset	Cascaded	1st	63.42 (3.74)	51.27 (2.51)	63.02 (3.52)	62.91 (3.54)	49.51 (4.92)	49.64 (4.96)
		2nd	**73.69 (3.16)**	**64.78 (0.88)**	75.20 (2.89)	73.70 (2.93)	63.94 (4.16)	64.36 (4.26)
	Ensemble	1st	67.27 (2.94)	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
		2nd	**87.27 (2.25)**	**71.24 (2.90)**	84.84 (2.19)	85.74 (2.17)	82.17 (3.17)	82.39 (3.20)
External dataset	Cascaded	1st	54.33 (4.29)	42.56 (4.79)	66.07 (5.06)	56.51 (3.04)	36.68 (5.97)	39.16 (5.37)
		2nd	**54.46 (4.33)**	**42.59 (4.80)**	66.20 (5.13)	56.68 (3.10)	36.83 (5.97)	39.28 (5.38)
	Ensemble	1st	55.34 (6.09)	41.68 (3.88)	64.14 (5.11)	56.82 (4.61)	36.30 (7.89)	38.33 (7.38)
		2nd	**86.06 (3.10)**	**64.49 (8.16)**	88.38 (3.77)	85.83 (3.24)	78.13 (4.50)	78.69 (4.36)

Table 4.10 demonstrates the performance of cascaded model, external validation with ensemble model, and individual training with ensemble model for the external datasets depending on whether the second option was included. When only the first treatment option was accepted as the correct answer, the mean accuracy in external validation was found to be higher than that in individual training with values of 66.53% and 55.34%, respectively. In contrast, the mean accuracy in external validation was 86.06%, which is better than 84.08% of that in individual training. In addition, the cascaded model did not show any increase in accuracy although the second treatment option was accepted.

Table 4.10. Performance of cascaded random forest vs. ensemble voting machine vs. individual training considering second options

Dataset	Options	Cascaded model		External validation		Individual training	
		Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
External dataset	1st	**54.33** (4.29)	42.56 (4.79)	**55.34** (6.09)	41.68 (3.88)	**66.53** (5.90)	46.56 (8.13)
	2nd	**54.46** (4.33)	42.59 (4.80)	**86.06** (3.10)	64.49 (8.16)	**84.08** (2.55)	59.63 (6.83)

4.2.5. Calibration of model

Finally, we calibrated the voting classifier. Figure 4.2 and Table 4.11 present results of pre- and post-calibration of the model. Owing to the model calibration, the standard deviation of accuracy of the external dataset was reduced, although the mean accuracy was decreased slightly. Figure 4.2 shows the calibration plot for each treatment in the one-fold of cross-validation folds in the internal dataset. The effect was minimal in TACE+EBRT, sorafenib, and supportive care; however, the calibration performance improved in the rest of the treatments.

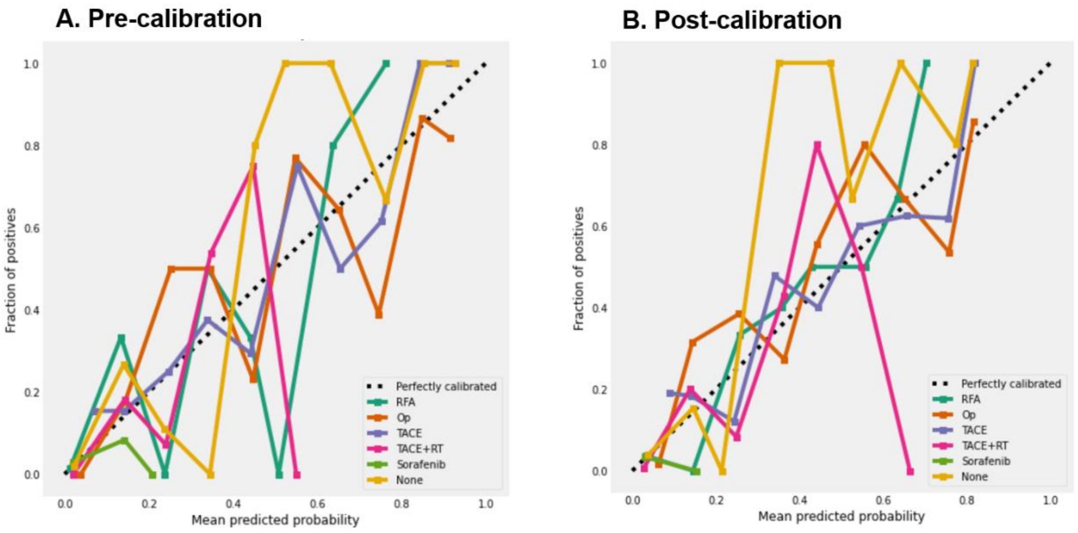


Figure 4.2. Calibration plot. A) Pre-calibration. B) Post-calibration.

Table 4.11. Performance of non-calibrated model vs. calibrated model

Calibration	Dataset	Options	Performance					
			Accuracy	Recall	Prec.	F1	Kappa	MCC
Pre-calibration	Internal dataset	1st	67.27 (2.94)	52.22 (4.67)	65.82 (2.68)	65.93 (2.70)	53.85 (4.21)	54.22 (4.34)
		2nd	87.27 (2.25)	71.24 (2.90)	84.84 (2.19)	85.74 (2.17)	82.17 (3.17)	82.39 (3.20)
	External dataset	1st	55.34 (6.09)	41.68 (3.88)	64.14 (5.11)	56.82 (4.61)	36.30 (7.89)	38.33 (7.38)
		2nd	86.06 (3.10)	64.49 (8.16)	88.38 (3.77)	85.83 (3.24)	78.13 (4.50)	78.69 (4.36)
Post-calibration	Internal	1st	**65.99 (2.68)**	50.35 (3.45)	64.45 (2.65)	64.60 (2.46)	51.97 (3.53)	52.32 (3.66)

	dataset	2nd	**86.52 (2.84)**	69.52 (4.24)	84.66 (3.31)	85.00 (2.92)	81.00 (4.11)	81.31 (4.06)
	External dataset	1st	**56.65 (4.98)**	45.28 (7.39)	66.17 (5.71)	58.60 (4.09)	38.33 (6.24)	40.50 (5.69)
		2nd	**85.29 (1.65)**	62.24 (7.32)	85.92 (3.59)	84.92 (2.44)	76.74 (2.89)	77.33 (2.82)

4.3. Discussion

In our experiments, ensemble voting machine exhibited better performance than previous cascaded random forest model. In the previous study [49], the model was intended to pretend according to the processes of decision making of physicians in real clinical situation. In most cases, decision for curative treatment versus non-curative treatment was preceded than decision for subdivided treatment choices. However, the hierarchical cascaded method presented worse performance compared to the ensemble voting method. One reason for this may be the errors that were generated at the higher level accumulated as they moved to the lower level. Meanwhile, the ensemble classifiers enable the compensation for the weakness of individual classifiers and use their combined knowledge to enhance its performance. Combining multiple outcomes in the ensembles have many reasons to achieve benefits of improved detection accuracy over a single base classifier. Dietterich et al. described the three major reasons involving statistical, computational, and representational reasons [56]. The statistical reason is that if the training data are insufficient for modelling the hypothesis space using one learning algorithm, then the aggregated knowledge of an ensemble may provide more correct outcomes. Computational reason is that the learning methods can be trapped in local optima, which may require high computational efforts for determining the global optima. Thus, it may be more appropriate for executing many local search methods from different initial points and aggregate them. Finally, the representational reason is that the optimal classification method may not be established, because the method may not be capable of modelling the hypothesis space of the problem accurately.

Kumar et al. presented the motivation and comprehensive review of intrusion detection systems based on ensembles in machine learning in their review paper [57]. In particular, they analyzed different ensemble methods in the field, considering different types of ensembles, and various approaches for integrating the predictions of individual classifiers for an ensemble classifier. In their study, it was shown that several studies results could be improved with help of ensemble classifier compared to that of an individual classifier as shown in our results. In general, ensemble classifier is generated on the basis of a set of individual classifiers, followed by selection of certain correct and diverse classifiers, and finally aggregating their outcomes. Furthermore, a different set of individual classifiers for each institution in the individual training did not significantly affect the performance compared to fixed set of individual classifiers for internal dataset. Moreover, the ensemble classifiers were not implemented based on weighted majority voting in our experiments. Therefore, the weighted majority voting method as well as a new method for constructing this type of ensemble classifiers can be explored in future research. The research can be focused on both developing a weighting scheme that defines the way to measure the reliability of each classifier, and the weight generation method that determines the values of weight coefficients used to measure the reliability of each classifier [57].

Experiments for various normalization and oversampling methods did not show any improvement over original dataset. Five machine learning classifiers composing ensemble voting machine were C-support vector machine with linear kernel, gaussian process classifier, random forest classifier, extra-trees classifier, and histogram-based gradient boosting classifier. Among them, the random forest classifier, extra-trees classifier, and histogram-based gradient boosting classifier are tree-based classifiers, wherein no difference in the results were observed regardless of whether they used unnormalized data or normalized data. Thus, it is likely that this was reflected in the results. In case of oversampling test, the mean accuracy in the internal dataset after applying ROS and SMOTE was slightly decreased, while that in external dataset it increased. Leevy et al. provided a large survey of published studies focusing on high-class imbalance (i.e., a majority-to-minority class ratio between 100:1 and 10,000:1) in big data to assist in addressing the adverse effects owing to class imbalance [7]. In their review, ROS demonstrated a better classification accuracy than RUS or SMOTE in most studies. However, none of the oversampling methods contributed significantly to the performance improvement with our datasets.

Figure 4.3 shows the distribution of features after reduction of features from 20 features to 2-dimension using t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA) after passing through the neural network. For visualization of 20 features, the entire internal dataset was trained with validation set of 20 % of whole dataset using an artificial neural network comprising three-layers (20-25-10-6) to classify to six classes. Further, 10 features from intermediated layer of neural network were reduced to two dimensions with t-SNA and PCA. In Figure 4.3, as evident, the six treatments are not clearly separated in both training and validation set. In addition, the features in case of incorrect prediction using the same visualization method were analyzed. The reduced features of four groups from twenty features are depicted on the two-dimensional plane in Figure 4.4. The four groups include a group of patients treated with resection in the training set (Train=Op), treated with RFA/PEIT in the training set (Train=RFA), recommended for treatment with resection in the test set who had actually undergone RFA/PEIT in real situation (Tx=RFA, Pred=Op), and recommended for treatment with resection in the test set and had undergone resection (Tx=RFA, Pred=Op). The group of incorrect prediction (“Tx=RFA, Pred=Op”) were mainly shown in the mixed area of “Train=RFA” and “Train=Op” groups as shown in Figure 4.4. Thus, this features analysis shows the reason for prediction of initial treatment for patients with HCC being challenging.

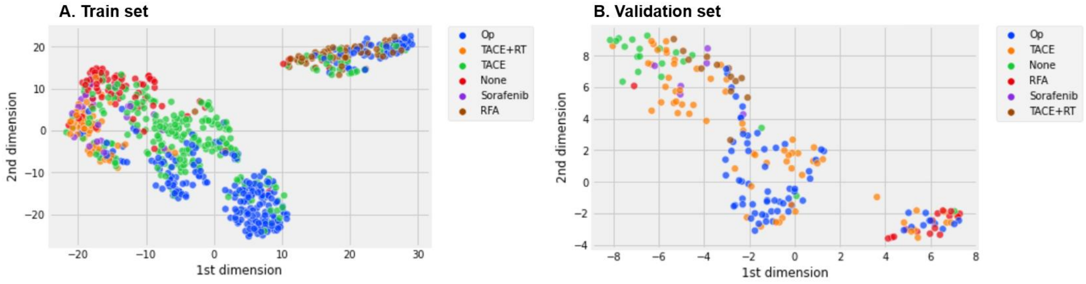


Figure 4.3. Visualization of features of six treatments. A) Features in train set. B) Features in validation set

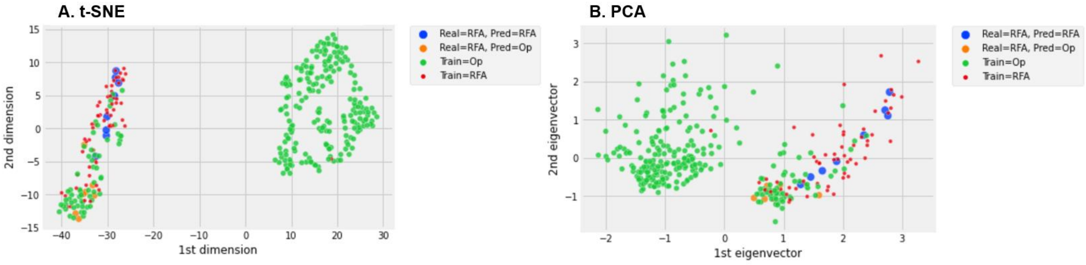
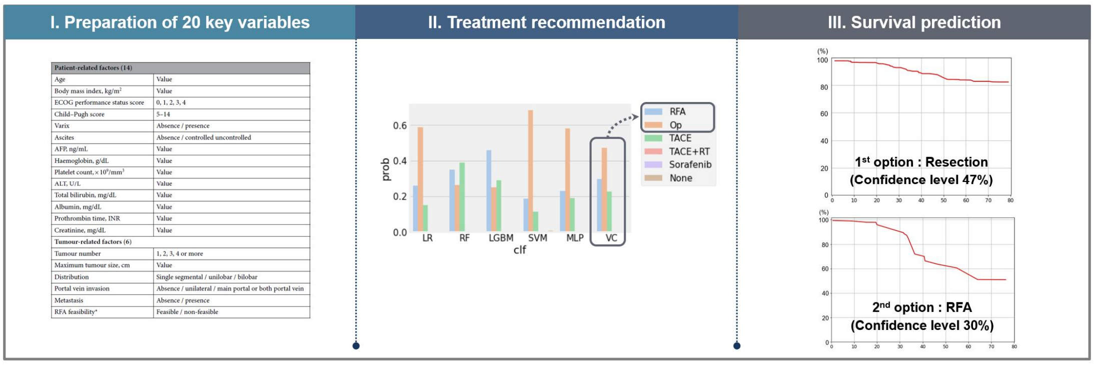


Figure 4.4. Feature analysis for cases of incorrect prediction. A) Results of feature reduction using t-SNE, B) Results of feature reduction using PCA

The results of external validation on datasets from eight centers of the model trained with the dataset of single center were not satisfactory. This may be primarily because each institution follows different patterns of treatments. This issue arises from the nature of results from machine learning being solely dependent on the training dataset. The European Commission also highlighted in their white paper on AI that retraining non-EU AI systems with European data may negatively impact accuracy without necessarily improving fairness if certain parameters that determine trade-offs between these two are specifically hard coded in the algorithm itself [58]. Moreover, they mentioned that algorithms may be fair even if the data used to train them capture human biases. Thus, to reflect more individual policies and preferences for each center, it is effective to train with their own sufficient dataset as shown in our results. However, building a robust model with small amount of data using machine learning is challenging. Although individual training with dataset of each center presented higher classification performance than those from external validation, they were highly imbalanced between treatments in certain institutions with high performance for individual training, which is prone to overfitting. In addition, the mean accuracy in external validation was better than that in individual training when the second treatment option was accepted as the correct answer. The results demonstrate that the trained model with more abundant dataset can predict more minor classes well and it may be reflected to improvement of performance. Furthermore, the probability scores obtained from voting classifier for first and second treatment options could be expressed with levels of confidence in addition to the results of survival prediction as depicted in Figure 4.5.



Patient-related factors (14)	
Age	Value
Body mass index, kg/m^2	Value
ECOG performance status score	0, 1, 2, 3, 4
Child–Pugh score	5–14
Varix	Absence / presence
Ascites	Absence / controlled uncontrolled
AFP, ng/mL	Value
Haemoglobin, g/dL	Value
Platelet count, × 10^9/mm^3	Value
ALT, U/L	Value
Total bilirubin, mg/dL	Value
Albumin, mg/dL	Value
Prothrombin time, INR	Value
Creatinine, mg/dL	Value
Tumour-related factors (6)	
Tumour number	1, 2, 3, 4 or more
Maximum tumour size, cm	Value
Distribution	Single segmental / unilobar / bilobar
Portal vein invasion	Absence / unilateral / main portal or both portal vein
Metastasis	Absence / presence
RFA feasibility*	Feasible / non-feasible

Figure 4.5. Modified model for treatment recommendation

5. Model for survival prediction

Chapter 5 presents the details of several experiments and results regarding the model for survival prediction. Different training mode with previous random survival forest model was evaluated for internal dataset. Further, using the modified model, results of individual training with dataset of each center and those of external validation were compared. Finally, the results of survival prediction according to recommended treatments were demonstrated.

All the proposed models were based on random survival forest [35, 36] architecture, which has been applied to previous model for survival prediction [49]. We begin by describing modification of training mode in previous model and the results of internal and external validation after modification of the model, followed with the results of individual training for each center. Finally, the various results of simulation of survival prediction according to treatment recommendation were demonstrated. Upon selecting the first and second treatment option using model to recommend initial treatment, the model for survival prediction demonstrates the predicted survival curve according to each option for each patient. Therefore, the model can predict different survival curves of the same patient with different treatments, which can aid clinicians in making treatment decisions in actual clinical setting.

5.1. Basic setup and notation

Similar to that in Chapter 4, all experiments for internal dataset and individual training for external dataset were trained and validated with five-fold cross validation stratified by treatment. Further, the experiments for external validation were conducted after fitting the model with the entire internal dataset. We used the following three metrics for evaluating the performance of survival prediction models: Harrell's C index, integrated time-dependent area under the ROC curve (iAUC), and integrated Brier score (IBS).

5.2. Experiments

5.2.1. Modification of training mode

In the previous model, training of random survival forest model was conducted with 20 key variables from each group who received specific treatment. We modified the training mode to training with 21 variables wherein the initial treatment information was included in addition to 20 key variables. The results of training with 20 and 21 variables are presented in Table 5.1. The implementation of modified training mode exhibit better performance in all evaluation metrics.

Table 5.1. Performance of survival prediction using 20 variables vs. 21 variables

Validation	20 variables			21 variables		
	C-index	**iAUC**	**IBS**	**C-index**	**iAUC**	**IBS**
Whole				**0.8381 (0.0276)**	**91.89 (2.08)**	**0.12 (0.01)**
RFA	0.7158 (0.1926)	72.98 (15.40)	0.23 (0.09)	0.7656 (0.1438)	77.49 (12.00)	0.17 (0.07)
Op	0.6681 (0.0401)	72.56 (6.05)	0.16 (0.04)	0.7049 (0.0421)	75.94 (6.09)	0.14 (0.03)
TACE	0.7796 (0.0397)	86.03 (2.85)	0.14 (0.01)	0.7842 (0.0338)	86.18 (2.42)	0.14 (0.01)
TACE+RT	0.6553 (0.0807)	67.57 (10.99)	0.21 (0.03)	0.6487 (0.0680)	72.63 (8.92)	0.20 (0.03)
Sorafenib	0.7419 (0.2060)	77.61 (22.02)	0.22 (0.04)	0.7562 (0.1505)	81.04 (15.81)	0.18 (0.03)
None	0.7352 (0.0385)	75.81 (2.98)	0.16 (0.02)	0.7566 (0.7360)	81.66 (79.16)	0.15 (0.16)
Average	**0.7160 (0.0996)**	**75.43 (10.05)**	**0.19 (0.04)**	**0.7360 (0.0783)**	**79.16 (8.37)**	**0.16 (0.03)**

Using the modified training mode, we evaluated the performance of survival prediction model for the internal and external datasets. As shown in Table 5.2, the results of internal validation are better than those of external validation test set. Furthermore, the performance of survival prediction for each treatment in the internal dataset is presented in Table 5.3. The iAUC ranges from 72.63 to 86.18 with the highest performance being demonstrated in TACE.

Table 5.2. Performance of survival prediction of internal and external datasets

Dataset	**Center**	**Performance**		
		C-index	**iAUC**	**IBS**
Internal	AMC	**0.8381 (0.0276)**	**91.89 (2.08)**	**0.12 (0.01)**
External	KUGH	0.7812	87.08	0.10
	SNUBH	0.7833	88.23	0.15
	SMC	0.7580	84.28	0.15
	SNUH	0.8053	89.48	0.14
	CMC	0.7458	84.78	0.18
	SH	0.8254	89.39	0.10
	CUH	0.7208	80.73	0.17
	IUH	0.7935	87.84	0.10
	Average	**0.7767 (0.0315)**	**86.48 (2.82)**	**0.14 (0.03)**

Table 5.3. Performance of survival prediction for each treatment in the internal dataset

Treatment	Performance		
	C-index	iAUC	IBS
RFA	0.7656 (0.1438)	77.49 (12.00)	0.17 (0.07)
Op	0.7049 (0.0421)	75.94 (6.09)	0.14 (0.03)
TACE	0.7842 (0.0338)	**86.18 (2.42)**	0.14 (0.01)
TACE+RT	0.6487 (0.0680)	**72.63 (8.92)**	0.20 (0.03)
Sorafenib	0.7562 (0.1505)	81.04 (15.81)	0.18 (0.03)
None	0.7566 (0.7360)	81.66 (9.16)	0.15 (0.06)
Average	0.7360 (0.0783)	79.16 (8.37)	0.16 (0.03)

5.2.2. Individual training for external dataset

We trained and evaluated individually with the dataset of each center employing a stratified 5-fold cross-validation. Table 5.4 presents the performance of individual training and external validation with the model trained with internal dataset. In contrast to the results of model for treatment recommendation, almost no difference was observed in the C-index, iAUC, and IBS between two experiments.

Table 5.4. Performance of survival prediction in the external validation and individual training

Center	External validation			Individual training		
	C-index	**iAUC**	**IBS**	**C-index**	**iAUC**	**IBS**
KUGH	0.7812	87.08	0.10	0.8069 (0.0546)	86.41 (2.64)	0.13 (0.03)
SNUBH	0.7833	88.23	0.15	0.7856 (0.0372)	87.67 (4.02)	0.15 (0.01)
SMC	0.7580	84.28	0.15	0.7715 (0.0166)	85.12 (2.31)	0.15 (0.01)
SNUH	0.8053	89.48	0.14	0.8213 (0.0351)	89.73 (3.04)	0.14 (0.01)
CMC	0.7458	84.78	0.18	0.7329 (0.0513)	83.07 (5.56)	0.17 (0.02)
SH	0.8254	89.39	0.10	0.8021 (0.0384)	87.31 (4.90)	0.13 (0.02)
CUH	0.7208	80.73	0.17	0.7896 (0.0308)	83.93 (4.76)	0.15 (0.01)
IUH	0.7935	87.84	0.10	0.7771 (0.0600)	86.12 (5.36)	0.14 (0.03)
Average	**0.7767 (0.0315)**	**86.48 (2.82)**	**0.14 (0.03)**	**0.7859 (0.0405)**	**86.17 (4.07)**	**0.14 (0.02)**

5.2.3. Survival prediction after treatment recommendation

We simulated our two-stage model sequentially constructed by the model for treatment recommendation and survival prediction. Subsequently, the risk was stratified by predicting survival considering the results of treatment recommendation. Figures 5.1 and 5.2 shows the predicted survival curve according to a recommended treatment. In Figure 5.1, the survival curve shows similar pattern between patient groups who were recommended for TACE and had undergone TACE or not. Meanwhile, the results of survival prediction presented different patterns when the group who were recommended for TACE but did not receive TACE in reality and thus were subdivided by each treatment actually received, as shown in Figure 5.2.

Although the treatment information were all the same with TACE, the survival curves were predicted differently by other features except for the treatment information.

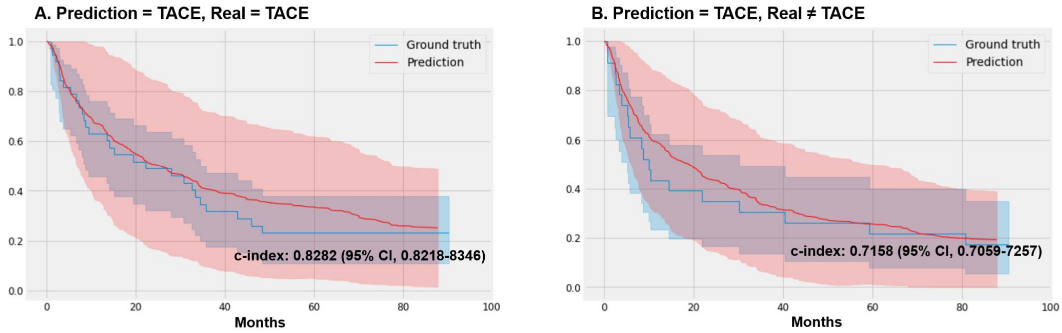


Figure 5.1. Predicted survival curve according to a recommended treatment. A) Results of a group of patients who were recommended for TACE, and undergone TACE. B) Results of a group of patients who were recommended for TACE but received a different treatment.

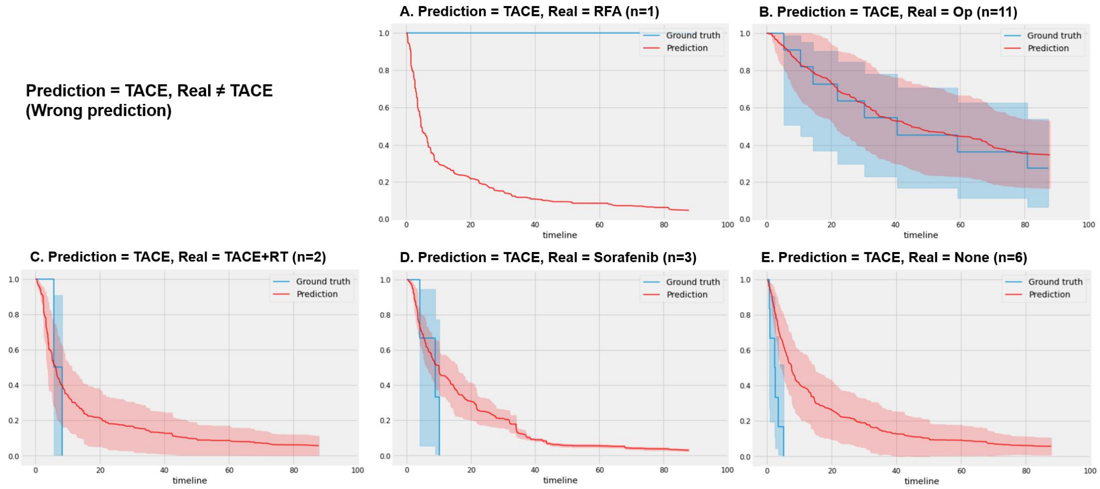


Figure 5.2. Predicted survival curve according to a recommended treatment. Results of a group of patients who were recommended for TACE but received A) RFA. B) Resection. C) TACE+EBRT. D) Sorafenib. E) Supportive care.

The impact of treatment information with one of the 21 features as input to survival prediction model is shown in Figure 5.3. Training with the different treatment information presented different survival curve in cases of TACE and resection, and TACE and supportive care. Further, the survival curve when the treatment feature was resection showed better survival than when it was TACE. However, the survival curve when treatment feature was supportive care demonstrated slightly worse survival compared to TACE. In particular, the predicted survival curve were more similar to the actual survival curve when the treatment information by the treatment recommendation model was inserted compared to when the actual treatment was included as treatment information.

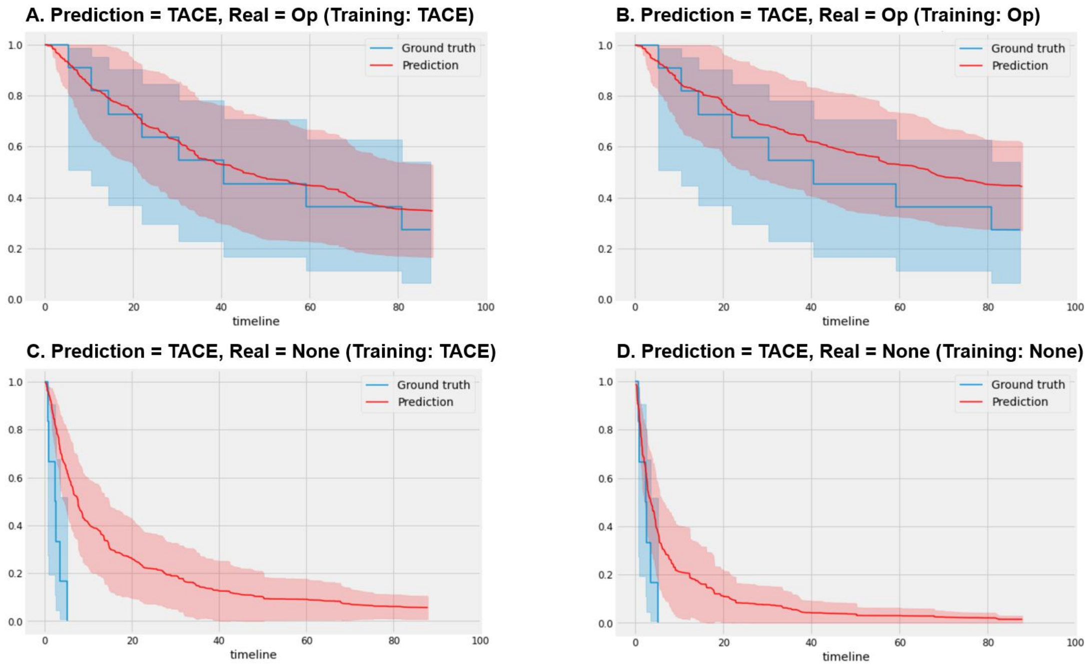


Figure 5.3. Predicted survival curve according to a recommended treatment. Results of a group of patients who were recommended for TACE but received resection with treatment information as A) TACE. B) Resection. Results of a group of patients who were recommended for TACE but received another treatment with treatment information as C) TACE. D) None. (=Supportive care)

Although not being very clear, iAUC showed a tendency to increase with increase in the number of patients, as depicted in Figure 5.4. However, this needs to be verified through further experiments.

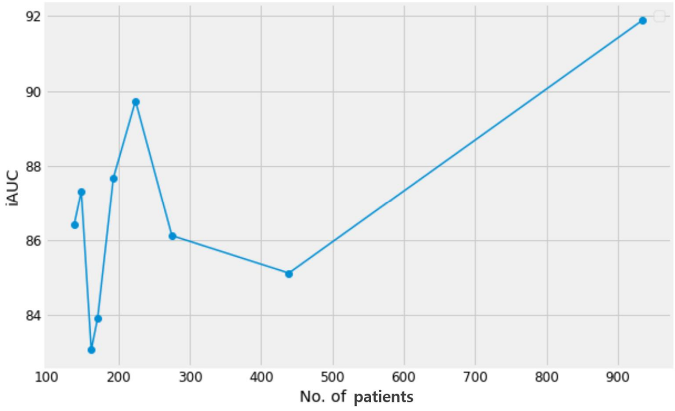


Figure 5.4. iAUC according to number of patients

5.3. Discussion

In our experiments on the model for survival prediction using random survival forest model, changing the training owing to the inclusion of initial treatment information in addition to 20 key variables from training with 20 key variables in each treatment group showed slightly higher performance. Further, the results of individual training with dataset of each center demonstrated similar or worse performance than those from external validation, which was different from the results of model for treatment recommendation. Finally, the two-stage model comprising treatment recommendation and subsequent survival prediction were simulated, wherein, the prognosis in each patient according to the results of treatment recommendation was stratified.

Previous training mode as training with 20 key variables in each treatment group demonstrated the noticeable drawback in the simulation study. In the group with fewer patients choosing the treatment, for example, Sorafenib group, the outcome was found to be always inferior, regardless of the other characteristics the patient possessed. However, after modifying the training mode by incorporating initial treatment information into features to be trained, the other characteristics as well as initial treatment were also considered for predicting survival, although the treatment feature showed the highest importance score in the random survival forest model.

In the experimental results of model for treatment recommendation in Chapter 4, individual training with dataset of each center presented higher performance than those from external validation when the second option was not considered. This may be because of the differences in patient groups, policies, and preferences in selecting treatments in other institutions could not be reflected only through training on the internal dataset. In contrast, individual training showed similar or worse performance than those from external validation in the experiment for survival prediction model. Moreover, the model performance tended to increase with increase in the number of datasets used for training in our experiments. Thus, from these results, it can be carefully inferred that the model trained with the internal dataset that has the largest amount of data is suitable for use because the model for survival prediction shows a similar trend regardless of institution, in contrast to the model for treatment recommendation.

In the simulation study of our two-stage model, the survival curves were predicted differently when the group who were recommended for TACE were subdivided based on each treatment they actually received despite the treatment information all being the same with TACE. The results predicted by the model were consistent with the real survival curves for each treatment. In another experiment on the impact of different treatment information, training with the different treatment information demonstrated different survival curve. In particular, the predicted survival curve were more similar to the actual survival curve when the treatment information by the treatment recommendation model was inserted compared to when the actual treatment was included as treatment information. Thus, this experimental result is an example demonstrates the reliability of the treatment recommendation model and the successive survival prediction model that agree well with reality. Moreover, these results show that the proposed two-stage model can reliably predict survival according to the type of treatment patients receive and it can be used as a risk stratification tool.

6. Various applications for CDSS

In this chapter, we demonstrated several scenarios of the proposed two-stage model in real clinical setting. The chances of utilizing this model as an alternative of current staging system were described as a first scenario. Further, the usage of staging system in oncology facilitated the prediction of prognosis of patients and aided in making a decision of selecting proper treatment based on certain stage. This was main purpose of the proposed model as well. We compared the recommendation from our model and BCLC stage in specific situation, that is, in patients with BCLC stage C. Second, the concept of digital twin has been discussed. Currently, digital twin technology is being developed and commercialized to optimize several manufacturing and aviation processes, while in the healthcare and medicine fields this technology is still in its early developmental stage [59-61]. We simulated certain cases using two different models with same structure but separately trained with different dataset from two centers. Furthermore, the possible expansion of this model in real clinical setting and issues of reliability were discussed in the latter in this chapter.

6.1. Treatment recommendation for BCLC stage

The BCLC staging system is a primary prognostic model and an allocating tool of HCC treatment. This stage recommends seven treatments including ablation, resection, transplant, chemoembolization, systemic therapy, and best supportive care according to five stages as shown in Figure 6.1. However, there exists a significant discrepancy in the initial treatment choice for HCC between the recommendations obtained from the BCLC system and that applied in real clinical practice.

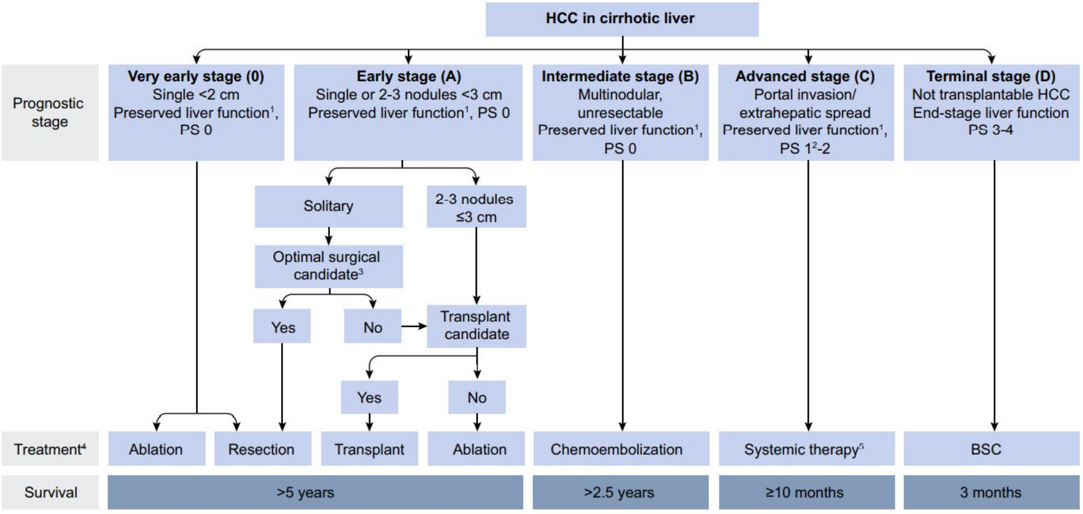


Figure 6.1. BCLC stages and treatment recommendation diagram [42]

The number of patients and their overall survival is demonstrated in Figure 6.2 using a Kaplan–Meier plot grouped with treatments that the patients have undergone in the internal dataset. In the internal dataset, there were 560 patients in BCLC stage C group. Although it was considered that the data were collected from January 2010 to October 2010 and BCLC guideline before 2010, a considerable number of patients treated with RFA/PEIT (also known as ablation), or resection, were contained in the group of BCLC stage C.

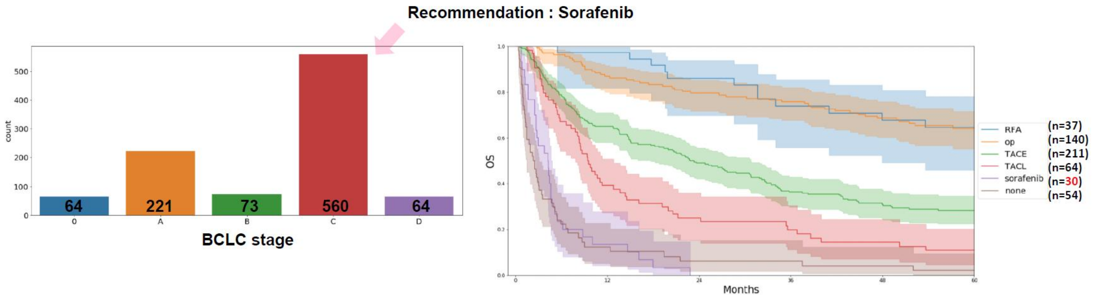
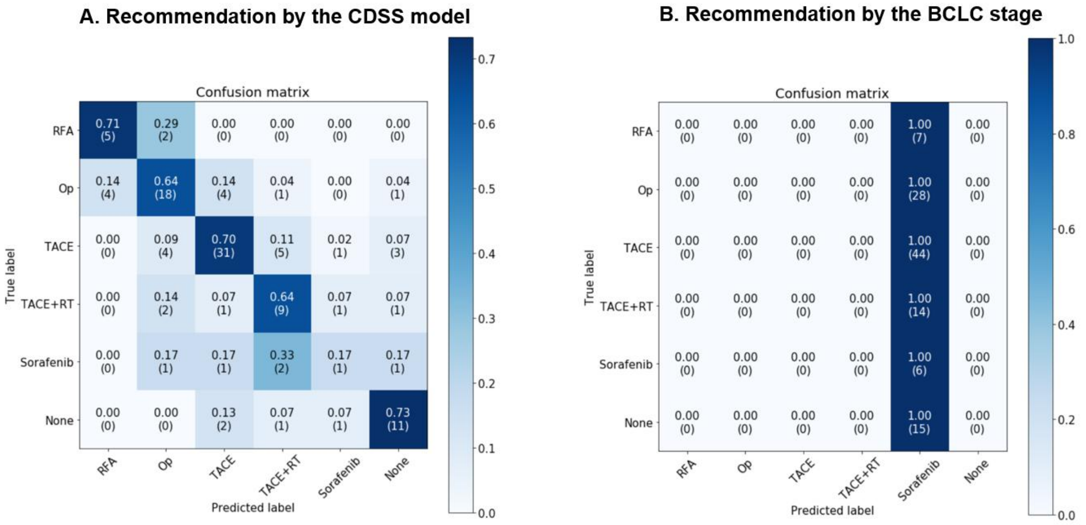


Figure 6.2. Number of patients and their overall survival in BCLC stage C group

Regarding the 560 patients in BCLC stage C group, we split patients 4:1 into a train and test group and thereafter evaluated the performance of the proposed model. Figure 6.3 demonstrates the results of recommendation of the model in the test group (N = 114).



	RFA	Op	TACE	TACE+RT	Sorafenib	None
RFA	0.71 (5)	0.29 (2)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Op	0.14 (4)	0.64 (18)	0.14 (4)	0.04 (1)	0.00 (0)	0.04 (1)
TACE	0.00 (0)	0.09 (4)	0.70 (31)	0.11 (5)	0.02 (1)	0.07 (3)
TACE+RT	0.00 (0)	0.14 (2)	0.07 (1)	0.64 (9)	0.07 (1)	0.07 (1)
Sorafenib	0.00 (0)	0.17 (1)	0.17 (1)	0.33 (2)	0.17 (1)	0.17 (1)
None	0.00 (0)	0.00 (0)	0.13 (2)	0.07 (1)	0.07 (1)	0.73 (11)

	RFA	Op	TACE	TACE+RT	Sorafenib	None
RFA	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (7)	0.00 (0)
Op	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (28)	0.00 (0)
TACE	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (44)	0.00 (0)
TACE+RT	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (14)	0.00 (0)
Sorafenib	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (6)	0.00 (0)
None	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (15)	0.00 (0)

Figure 6.3. Classification results in the test group for BCLC stage C group. A) Recommended by the ensemble voting machine, B) Recommended by the BCLC stage

Further performing feature analyses among groups with the same prediction can facilitate the configuration of the characters of each group using the importance of features from each classifier composing voting classifier. Certain features may be correlated with features used in BCLC staging system, while others may not because common staging system usually use less than ten features for convenience of usage. Machine learning-based model enables more accurate and sophisticated prediction for treatment and survival solely based on their own dataset. However, such a system cannot create a new solution that has not been included in the dataset. In other words, even when the institution is primarily following a pattern of treatment different from many other institutions, and is in contrast to the direction that the majority considers appropriate, this system cannot provide new options that are not included in the dataset. Despite realizing the decision-making process to the best of their knowledge, at times this model can be helpful, where physicians can refer to the manner in which other centers treat the patients with similar conditions and the results they have been achieving. In this context, we conducted a modelling using the concept of digital twin system and simulated the system in following topic.

6.2. Comparison between centers

At present, digital twin technology is being developed and commercialized to optimize several manufacturing and aviation processes, while in the healthcare and medicine fields this technology it is still in its early developmental stage [61-63]. It would be possible to use this model in same institution to choose initial treatment and predict prognosis following the current treatment system. In addition, it can be also used for performing comparisons between other institutions. The overall structure of proposed system is depicted in Figure 6.4. We used the dataset from two centers of AMC and SMC because these datasets contain the largest number of patients among all the datasets. We separately trained the model with different datasets and validated via cases from another center, which is not used for training in two centers at once.

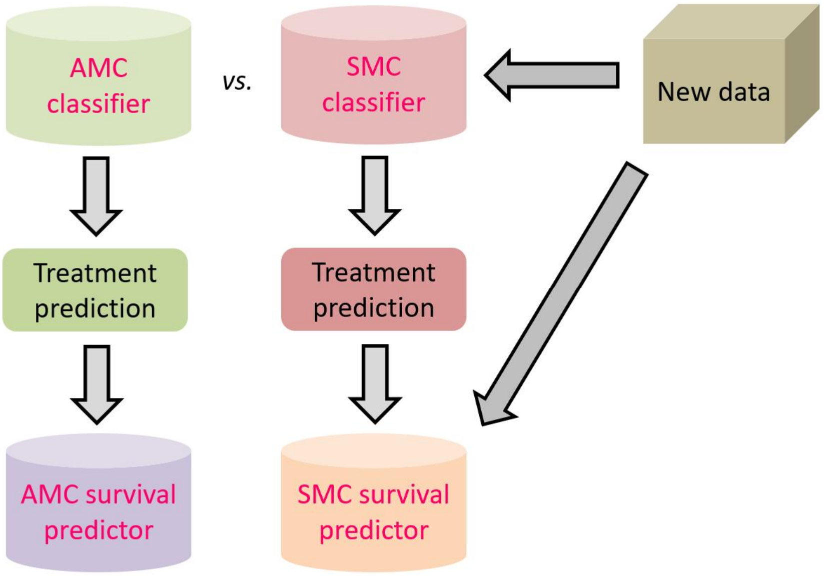


Figure 6.4. Conceptual diagram for center comparison

The result of survival prediction in case of one patient being treated with TACE in SNUBH is demonstrated in Figure 6.5. Considering the poor prognosis of the example case, it would not be possible to conduct resection in real clinical setting as twenty features are still not sufficient for making a definite decision and confirming it. Meanwhile, we can interpret very carefully that survival may be expected to increase even if the surgery might be possible in this patient. However, the final decision ultimately rests with the physicians and patients. The value of the proposed model will increase if physicians use these systems and review their own practices, and if these processes aid them in making the best decisions for cancer patients.

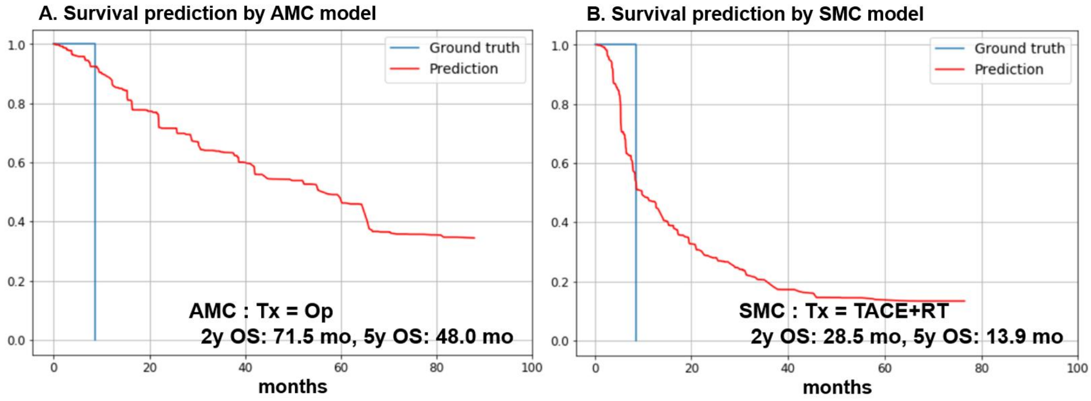


Figure 6.5. Results of survival prediction for one patient treated with TACE in SNUBH. A) Survival prediction after treatment recommended to resection by AMC model. B) Survival prediction after treatment recommended to TACE+EBRT by SMC model.

7. Discussion and conclusion

In this thesis, a two-stage model consisting of models for treatment recommendation and survival prediction after initial treatment was proposed for patients with HCC focusing on multi-center extension. The performance of treatment recommendation was improved using ensemble voting machine, and the model was applied in a multi-center setting owing to the suggestion of a second treatment option, in addition to first option with confidence levels. Although individual training with the dataset of each center revealed better performance in the model for treatment recommendation, results of external validation of the model trained by the internal dataset, exhibited acceptable performance with the setting of providing the second option. Meanwhile, results of individual training with the dataset of each center for survival prediction model were not observed to be superior to those of external validation of the model trained with the internal dataset. Further, the two-stage model was simulated, and consequently the risk was stratified via predicting the survival considering the results of treatment recommendation. Lastly, several scenarios of this model in real clinical situations were demonstrated. The feasibility of this model as an alternative to the current staging system and an experiment of virtual clinical simulation using this model trained with two different centers were presented as the possible expansion of this system.

Making treatment decision and predicting personalized prognosis for each patient is crucial in oncological management. However, there is no definite and unchanging gold standard in decision-making for oncological treatments. In particular, a staging system that can be used worldwide and the corresponding treatment recommendation system are lacking for HCC. Treatments for HCC have been developed in various ways to effectively control cancer while considering residual liver function. Each institution has individually investigated and applied various treatments that exploit the characteristics of each institution to improve the prognosis of HCC patients. In our study, the initial treatments in HCC patients around the same period showed considerably different distributions among institutions in South Korea. The diverse distribution of treatments by an institution reflects the preferences and policies for treatment selection of each institution, as well as the characteristics of the patient group. Therefore, individual training for each center is recommended if the goal of CDSS is to best reflect the characteristics of the center. However, often making a decision based on their own practice alone can be difficult for institutions with a small number of patients and inadequate experience with various treatments. In this case, the proposed modeling method can be employed as an alternative. According to our experimental results, the model trained by the internal dataset, containing the largest number of patients, showed acceptable performance with the setting of providing the second option along with the first option. However, a method to refer to the individual training version as well as the version of the big center together was introduced in Chapter 6. Moreover, in case of rare cancers, the decision-making may be more difficult due to scarce evidence. In addition, results of data-driven treatment recommendation and survival prediction using machine learning algorithm cannot solve this problem either. Thus, at centers with little experience that have not treated such precedents, it could be of great help in clinical practice to recommend the proper treatment and predict individual prognosis using experience of big centers with many patients. This information may also be helpful and provide certain intuitions to physicians particularly for in case of inexperienced one at the same center.

One of the weaknesses of this data-driven CDSS stems from their intrinsic data-restricted natures. Cancer treatments have characteristics that constantly change with time. New evidence is accumulated based on new research and treatments, and novel treatments are constantly replacing old treatments. In particular, the recent growth of massive genetic and clinical databases, along with efficient computing systems to facilitate them have accelerated the speed of treatment advances and shortened the cycle time for changes to treatment guidelines in oncology. While cancer treatment is changing rapidly, data-driven methods ultimately find answers in data based on the past. Until a proportion of newly developed treatment option is sufficient in the training dataset, the probability of recommending that treatment option can be very low. For example, in our case, Sorafenib treatment which is a protein kinase inhibitor with activity against many protein kinases, including vascular endothelial growth factor receptors, platelet-derived growth factor receptors, and RAF kinases [62, 63] has not been widely used in the period when dataset was collected. Consequently, it was rarely recommended by the proposed model as well as reliability of results of survival prediction in case with sorafenib was not high.

To overcome these weaknesses of the proposed model, physicians need to understand the limitations of this system when they use it, and developers must update new live data according to the requirement. Figure 7.1 shows the schematic of training and deployment of the model. For the model of treatment recommendation and survival prediction where there is no fixed gold standard and having characteristics that change over time, continual learning with new live data is required. In addition, monitoring whether results from the model correlate with those from physicians through periodic analysis is an essential part. If the concordance rate is decreased, modification for training dataset to reflect real clinical situation may be helpful through discussion between physicians and developers. Furthermore, application of the proposed model at a center as per the comparison introduced in Chapter 6 can present additional information by offering other practice options and their possible prognoses.

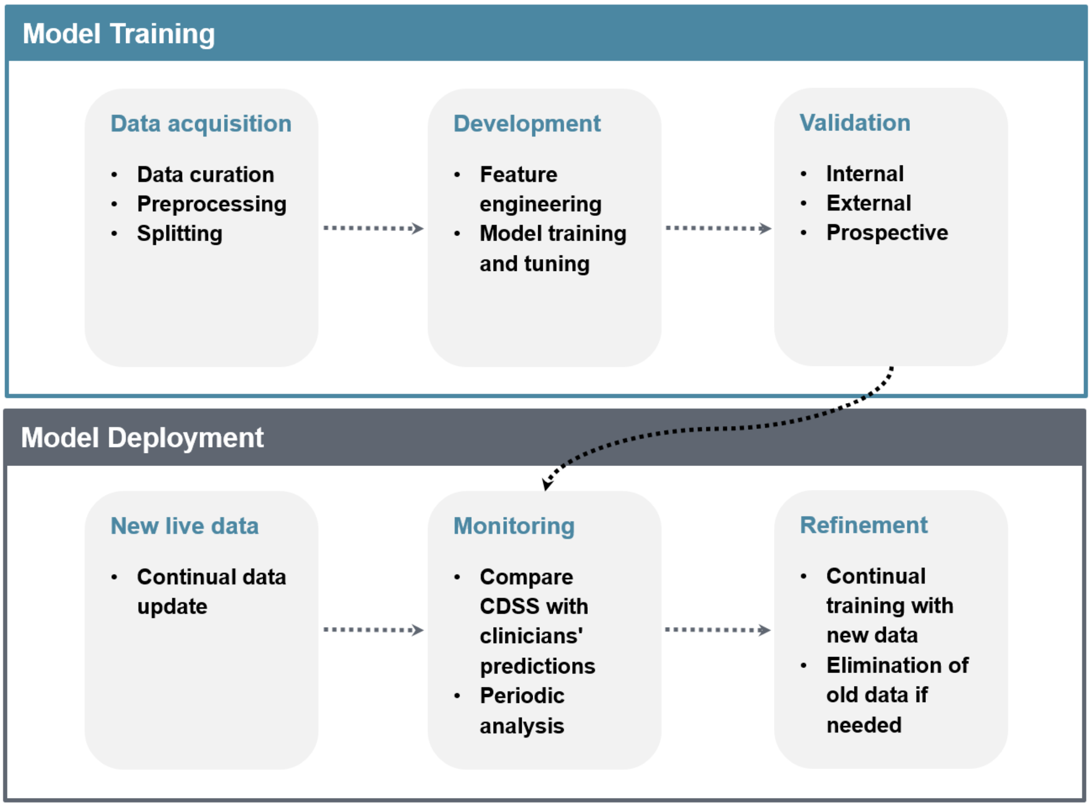
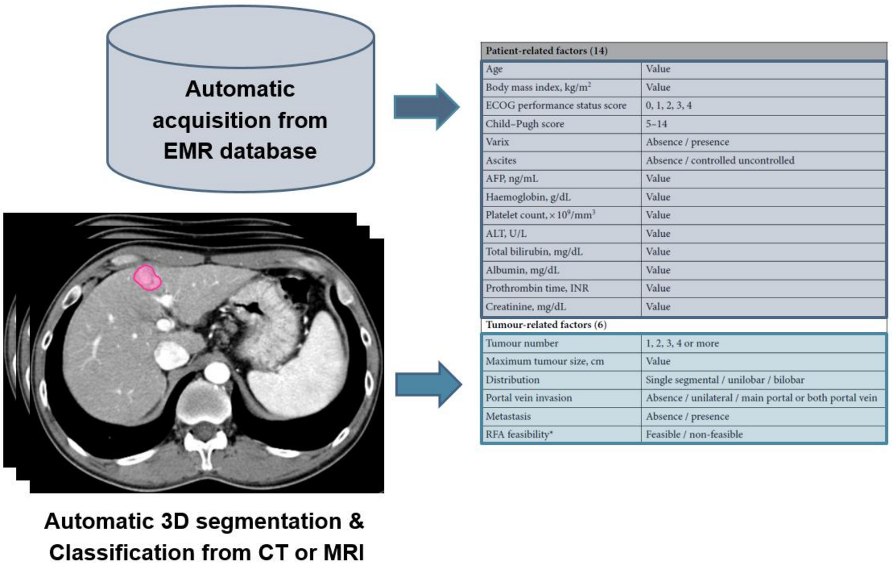


Figure 7.1. Issues on model training and deployment

Moreover, the factors that determine the treatment choices and prognosis of a patient are constantly changing in this environment wherein new treatments are constantly being developed. Thus, we selected twenty key features including patient-related 14 variables and 6 tumor-related factors based on the importance scores calculated using the automated classifier and survival prediction models among the 61 initial pretreatment variables. However, process of feature selection is very important for efficient training of machine learning model. Important features for changed and enlarged dataset with continual update of new data could be altered. Moreover, efficient feature selection is an additionally required important process in the continual learning. To facilitate automatic data update, automatic acquisition of data from electronic medical record (EMR) database and preprocessing for 14 patient-related factors and automatic segmentation and classification using deep learning technique for 6 tumor-related factors, shown in Figure 7.2, might be helpful in continual learning.

Furthermore, the efficient concatenation of deep features without hand-crafted feature reduction from medical images, such as computed tomography (CT) and clinical prognostic variables, would be helpful for treatment recommendation and survival prediction, as described in Figure 7.3.


Automatic acquisition from EMR database
Automatic 3D segmentation & Classification from CT or MRI

Patient-related factors (14)	
Age	Value
Body mass index, kg/m^2	Value
ECOG performance status score	0, 1, 2, 3, 4
Child–Pugh score	5–14
Varix	Absence / presence
Ascites	Absence / controlled uncontrolled
AFP, ng/mL	Value
Haemoglobin, g/dL	Value
Platelet count, × 10^9/mm^3	Value
ALT, U/L	Value
Total bilirubin, mg/dL	Value
Albumin, mg/dL	Value
Prothrombin time, INR	Value
Creatinine, mg/dL	Value
Tumour-related factors (6)	
Tumour number	1, 2, 3, 4 or more
Maximum tumour size, cm	Value
Distribution	Single segmental / unilobar / bilobar
Portal vein invasion	Absence / unilateral / main portal or both portal vein
Metastasis	Absence / presence
RFA feasibility*	Feasible / non-feasible

Figure 7.2. Overall framework for continual learning

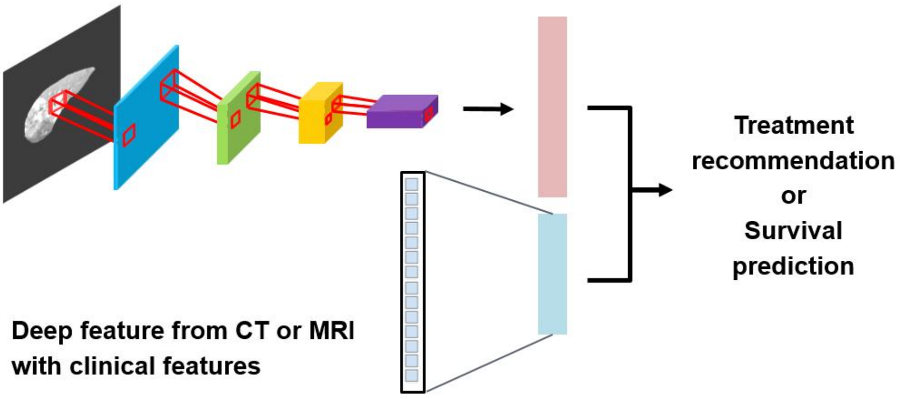

Treatment recommendation or Survival prediction
Deep feature from CT or MRI with clinical features

Figure 7.3. Probable model framework using deep features from medical images

Medicine is not just a science, but also a social and psychological subject. Wilkinson et al. mentioned two fundamental epistemological barriers to achieving individual-level predictions using machine learning [64]. First, machine learning approaches cannot identify cause and effect, because causal inference is fundamentally impossible to achieve without making assumptions. Second, majority of health states and events are so complex that we can only understand them probabilistically, and chance can never be predicted at the individual level. The authors asserted that many severe challenges in personalized medical care cannot be addressed through algorithmic complexity and thus require collaboration between traditional methodologists and experts in medical machine learning to avoid extensive research waste. In a similar context, any tool and guideline can only be used as a doctor's reference, and localization factors and individual elements should be considered for different patients, particularly for cancer patients with large heterogeneity. Further, the physical and mental state, economic situation, complications, and treatment preference of the patient as well medical reimbursement plan in different countries must be considered instead of providing advice based simply on existing knowledge. Furthermore, although our algorithm was validated from multicenter database in South Korea, the model might show less power when used in centers in other countries with different demographics (e.g., ethnicity, etiology, level of hospital facility, socio-economic status of the country, and even reimbursement policy), where the optimal treatment option would be different.

In conclusion, we developed a machine learning-based model for initial treatment recommendation and validated for multi-center datasets. We conducted various experiments to render this model usable in multi-center setting. Further, we improved the accuracy of model for treatment recommendation by using ensemble voting machine and demonstrated that the suggestion for a second treatment option in addition to first was suitable in multi-center setting. We simulated this two-stage model and stratified the risk by predicting survival according to the results of treatment recommendation. Moreover, we demonstrated several clinical scenarios of this model in real clinical setting and discussed issues on the model deployment with future research topics. This model for CDSS can be applied to provide practical utility to many physicians and patients of multiple centers.

References

1. Kalia M: **Biomarkers for personalized oncology: recent advances and future challenges**. *Metabolism* 2015, **64**(3):S16-S21.
2. Shin H-Y, Lee J-Y, Song J, Lee S, Lee J, Lim B, Kim H, Huh S: **Cause-of-death statistics in the Republic of Korea, 2014**. *Journal of the Korean Medical Association* 2016, **59**(3):221-232.
3. Osheroff JA, Teich JM, Levick D, Saldana L, Velasco FT, Sittig DF, Rogers KM, Jenders RA: **Improving outcomes with clinical decision support: an implementer's guide**: Himss Publishing; 2012.
4. Somashekhar SP, Sepulveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, Rauthan A, Arun Kumar N, Patil P, Rhee K *et al*: **Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board**. *Ann Oncol* 2018, **29**(2):418-423.
5. Samuel AL: **Some studies in machine learning using the game of checkers**. *IBM Journal of research and development* 1959, **3**(3):210-229.
6. Mahesh B: **Machine Learning Algorithms-A Review**. *International Journal of Science and Research (IJSR)[Internet]* 2020, **9**:381-386.
7. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N: **A survey on addressing high-class imbalance in big data**. *Journal of Big Data* 2018, **5**(1):1-30.
8. Kubat M, Matwin S: **Addressing the curse of imbalanced training sets: one-sided selection**. In: *Icml: 1997*: Citeseer; 1997: 179-186.
9. Lewis DD, Catlett J: **Heterogeneous uncertainty sampling for supervised learning**. In: *Machine learning proceedings 1994.* edn.: Elsevier; 1994: 148-156.
10. Grobelnik M: **Feature selection for unbalanced class distribution and naive bayes**. In: *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning: 1999*: Citeseer; 1999: 258-267.
11. Dumais S, Platt J, Heckerman D, Sahami M: **Inductive learning algorithms and representations for text categorization**. In: *Proceedings of the seventh international*

conference on Information and knowledge management: 1998; 1998: 148-155.

12. Ezawa KJ, Singh M, Norton SW: **Learning goal oriented Bayesian networks for telecommunications risk management**. In: *ICML: 1996*; 1996: 139-147.

13. Fawcett T, Provost FJ: **Combining data mining and machine learning for effective user profiling**. In: *KDD: 1996*; 1996: 8-13.

14. .

15. Wang H, Khoshgoftaar TM, Napolitano A: **An empirical investigation on wrapper-based feature selection for predicting software quality**. *International Journal of Software Engineering and Knowledge Engineering* 2015, **25**(01):93-114.

16. Malhotra R: **A systematic review of machine learning techniques for software fault prediction**. *Applied Soft Computing* 2015, **27**:504-518.

17. Zheng Z, Wu X, Srihari R: **Feature selection for text categorization on imbalanced data**. *ACM Sigkdd Explorations Newsletter* 2004, **6**(1):80-89.

18. Yin L, Ge Y, Xiao K, Wang X, Quan X: **Feature selection for high-dimensional imbalanced data**. *Neurocomputing* 2013, **105**:3-11.

19. Maurya A: **Bayesian optimization for predicting rare internal failures in manufacturing processes**. In: *2016 IEEE international conference on big data (big data): 2016*: IEEE; 2016: 2036-2045.

20. Aly M: **Survey on multiclass classification methods**. *Neural Netw* 2005, **19**:1-9.

21. Dietterich TG, Bakiri G: **Solving multiclass learning problems via error-correcting output codes**. *Journal of artificial intelligence research* 1994, **2**:263-286.

22. Loh WY: **Classification and regression trees**. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 2011, **1**(1):14-23.

23. Quinlan JR: **C4. 5: programs for machine learning**: Elsevier; 2014.

24. Bay SD: **Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets**. In: *ICML: 1998*: Citeseer; 1998: 37-45.

25. Rish I: **An empirical study of the naive Bayes classifier**. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence: 2001*; 2001: 41-46.

26. Cortes C, Vapnik V: **Support-vector networks**. *Machine learning* 1995, **20**(3):273-297.

27. Burges CJ: **A tutorial on support vector machines for pattern recognition**. *Data mining and knowledge discovery* 1998, **2**(2):121-167.

28. Lee Y, Lin Y, Wahba G: **Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data**. *Journal of the American Statistical Association* 2004, **99**(465):67-81.

29. Crammer K, Singer Y: **On the algorithmic implementation of multiclass kernel-based vector machines**. *Journal of machine learning research* 2001, **2**(Dec):265-292.

30. Bredensteiner EJ, Bennett KP: **Multicategory classification by support vector machines**. In: *Computational Optimization.* edn.: Springer; 1999: 53-79.

31. Pisner DA, Schnyer DM: **Support vector machine**. In: *Machine Learning.* edn.: Elsevier; 2020: 101-121.

32. Rau CS, Wu SC, Chuang JF, Huang CY, Liu HT, Chien PC, Hsieh CH: **Machine Learning Models of Survival Prediction in Trauma Patients**. *J Clin Med* 2019, **8**(6).

33. Park SY, Park JE, Kim H, Park SH: **Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches)**. *Korean J Radiol* 2021, **22**(10):1697-1707.

34. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y: **DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network**. *BMC medical research methodology* 2018, **18**(1):1-12.

35. **Random Forests for Survival, Regression, and Classification (RF-SRC), R Package Version 2.7.0** [https ://cran.r-proje ct.org/web/packa ges/rando mFore stSRC /citat ion.html]

36. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS: **Random survival forests**. *The annals of applied statistics* 2008, **2**(3):841-860.

37. Ferri C, Hernández-Orallo J, Modroiu R: **An experimental comparison of performance measures for classification**. *Pattern Recognition Letters* 2009, **30**(1):27-38.

38. Abajian A, Murali N, Savic LJ, Laage-Gaupp FM, Nezami N, Duncan JS, Schlachter T, Lin M, Geschwind JF, Chapiro J: **Predicting Treatment Response to Intra-**

arterial Therapies for Hepatocellular Carcinoma with the Use of Supervised Machine Learning-An Artificial Intelligence Concept. *J Vasc Interv Radiol* 2018, **29**(6):850-857 e851.

39. Harrell FE, Jr., Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors**. *Stat Med* 1996, **15**(4):361-387.

40. Graf E, Schmoor C, Sauerbrei W, Schumacher M: **Assessment and comparison of prognostic classification schemes for survival data**. *Stat Med* 1999, **18**(17-18):2529-2545.

41. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: **Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries**. *CA Cancer J Clin* 2021, **71**(3):209-249.

42. European Association for the Study of the Liver. Electronic address eee, European Association for the Study of the L: **EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma**. *J Hepatol* 2018, **69**(1):182-236.

43. Heimbach JK, Kulik LM, Finn RS, Sirlin CB, Abecassis MM, Roberts LR, Zhu AX, Murad MH, Marrero JA: **AASLD guidelines for the treatment of hepatocellular carcinoma**. *Hepatology* 2018, **67**(1):358-380.

44. Leoni S, Piscaglia F, Serio I, Terzi E, Pettinari I, Croci L, Marinelli S, Benevento F, Golfieri R, Bolondi L: **Adherence to AASLD guidelines for the treatment of hepatocellular carcinoma in clinical practice: experience of the Bologna Liver Oncology Group**. *Dig Liver Dis* 2014, **46**(6):549-555.

45. Park JW, Chen M, Colombo M, Roberts LR, Schwartz M, Chen PJ, Kudo M, Johnson P, Wagner S, Orsini LS *et al*: **Global patterns of hepatocellular carcinoma management from diagnosis to death: the BRIDGE Study**. *Liver Int* 2015, **35**(9):2155-2166.

46. Villanueva A, Hoshida Y, Toffanin S, Lachenmayer A, Alsinet C, Savic R, Cornella H, Llovet JM: **New strategies in hepatocellular carcinoma: genomic prognostic markers**. *Clin Cancer Res* 2010, **16**(19):4688-4694.

47. Cucchetti A, Piscaglia F, Grigioni AD, Ravaioli M, Cescon M, Zanello M, Grazi GL, Golfieri R, Grigioni WF, Pinna AD: **Preoperative prediction of hepatocellular carcinoma tumour grade and micro-vascular invasion by means of artificial neural network: a pilot study**. *J Hepatol* 2010, **52**(6):880-888.

48. Qiao G, Li J, Huang A, Yan Z, Lau WY, Shen F: **Artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma**. *J Gastroenterol Hepatol* 2014, **29**(12):2014-2020.

49. Choi GH, Yun J, Choi J, Lee D, Shim JH, Lee HC, Chung YH, Lee YS, Park B, Kim N *et al*: **Development of machine learning-based clinical decision support system for hepatocellular carcinoma**. *Sci Rep* 2020, **10**(1):14855.

50. Bruix J, Sherman M, American Association for the Study of Liver D: **Management of hepatocellular carcinoma: an update**. *Hepatology* 2011, **53**(3):1020-1022.

51. Wyner AJ, Olson M, Bleich J, Mease D: **Explaining the success of adaboost and random forests as interpolating classifiers**. *The Journal of Machine Learning Research* 2017, **18**(1):1558-1590.

52. Breiman L: **Random forests**. *Machine learning* 2001, **45**(1):5-32.

53. Carpenter J, Bithell J: **Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians**. *Statistics in medicine* 2000, **19**(9):1141-1164.

54. Bland JM, Altman DG: **Statistics notes: bootstrap resampling methods**. *bmj* 2015, **350**.

55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: **Scikit-learn: Machine learning in Python**. *the Journal of machine Learning research* 2011, **12**:2825-2830.

56. Dietterich TG: **Ensemble methods in machine learning**. In: *International workshop on multiple classifier systems: 2000*: Springer; 2000: 1-15.

57. Kumar G, Thakur K, Ayyagari MR: **MLEsIDSs: machine learning-based ensembles for intrusion detection systems—a review**. *The Journal of Supercomputing* 2020:1-34.

58. Cohen IG, Evgeniou T, Gerke S, Minssen T: **The European artificial intelligence**

strategy: implications and challenges for digital health. *Lancet Digit Health* 2020, **2**(7):e376-e379.

59. Barricelli BR, Casiraghi E, Fogli D: **A survey on digital twin: definitions, characteristics, applications, and design implications**. *IEEE access* 2019, **7**:167653-167671.

60. Jimenez JI, Jahankhani H, Kendzierskyj S: **Health care in the cyberspace: Medical cyber-physical system and digital twin challenges**. In: *Digital twin technologies and smart cities.* edn.: Springer; 2020: 79-92.

61. .

62. Keating GM, Santoro A: **Sorafenib: a review of its use in advanced hepatocellular carcinoma**. *Drugs* 2009, **69**(2):223-240.

63. Wilhelm SM, Adnane L, Newell P, Villanueva A, Llovet JM, Lynch M: **Preclinical overview of sorafenib, a multikinase inhibitor that targets both Raf and VEGF and PDGF receptor tyrosine kinase signaling**. *Mol Cancer Ther* 2008, **7**(10):3129-3140.

64. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, de Kamps M, Beam A, Konigorski S, Lippert C *et al*: **Time to reality check the promises of machine learning-powered precision medicine**. *Lancet Digit Health* 2020, **2**(12):e677-e680.

Abstract (In Korean)

암 환자를 치료함에 있어 환자의 개별 예후를 예측하고, 적절한 치료 전략을 선택하는 일은 매우 중요하다. 임상의에게 익숙한 일반적인 통계 모델은 표본에서의 관찰 결과를 통해 예후 인자를 판별하고 특정 예후 인자를 가진 경우의 상대적인 위험률 등을 평가할 수는 있지만, 이를 이용하여 환자의 개별 예후를 예측하기는 힘들다. 최근 의료 분야에서는 인공 지능 기술을 이용하여 임상 의사 결정 지원 시스템을 개발하려는 시도들이 활발하게 이뤄지고 있다. 그러나 이러한 예후 예측 모델을 개발하였다 하더라도 서로 다른 환자군 및 특성을 가진 다른 임상 기관에서도 사용할 수 있도록 일반화 시키는 것은 상당히 어려운 문제이다. 본 학위 논문에서는 단일기관에서 일정 기간 치료한 간세포 암종 (Hepatocellular carcinoma) 환자들의 데이터를 바탕으로 초기 치료를 권고하고, 이에 따른 생존율을 예측해 주는 기계 학습 기반의 2 단계 모델을 개발 및 검증하였다. 특히, 이 모델을 국내 8 개 의료기관에서 확보한 외부 데이터셋을 활용하여 검증하고 이를 다른 기관에서도 사용 가능하도록 수정 및 보완함으로써 다기관 활용성을 높일 수 있는 방안에 대해 논의하였다.

본 학위 논문의 첫 번째 단계에서는 치료 전 주요 변수 20 개를 이용하여 간세포암종의 초기 치료에 사용되는 6 가지 치료 중 하나를 권장하는 모델을 개발하였고, 이를 다기관 데이터셋을 사용하여 검증하였다. 우선 여러 기계 학습 모델을 테스트한 후 가장 좋은 성능을 보인 5 개의 기계 학습 분류기를 사용하여 최종적으로 앙상블 보팅 분류기 (Ensemble voting classifier)를 사용하여 치료를 추천하도록 하였고, 데이터셋에 대해 정규화나 샘플링 기법을 활용하여 분류의 성능이 향상되는 지 실험하였다. 또한, 단일 기관의 데이터셋으로 훈련한 모델을 다기관 데이터셋으로 검증하는 것에 그치지 않고 각 개별 기관의 데이터셋으로 훈련한 결과와 비교하였는데, 그 결과 개별 기관의 데이터셋으로 훈련한 모델의 결과가 더 높은 정확도를 보였다. 반면, 개별 기관의 데이터 수가 적어 예측의 신뢰도가 떨어지는 경우를 보완하기 위하여 타기관의 데이터로 훈련한 모델을 사용하기 위한 방법으로 첫 번째 추천 옵션과 두 번째 추천 옵션을 그 신뢰 수준과 함께 제공하는 방식을 사용하였을 때 가장 높은 정확도를 보였고, 이를 본 모델에서 다기관 확장성을 높일 수 있는 하나의 방안으로 제시하였다.

본 학위 논문의 두 번째 단계에서는 초기 치료 이후의 생존 예측 모델을 개발하였다. 치료 권고 모델에 사용된 20 개의 주요 변수에 초 치료 정보를 포함하여 총 21 개의

변수로 무작위 생존 숲 (Random survival forest) 모델을 사용하여 개별 환자의 생존율을 예측하였다. 이 모델에 대해서도 단일 기관의 데이터셋으로 훈련한 모델의 다기관 검증 결과와 각 개별 기관의 데이터셋으로 훈련한 결과를 비교하였는데, 그 결과 치료 권고 모델의 결과와는 대조적으로 다기관 검증 결과가 개별 훈련 후 검증시와 유사하거나 더 좋은 성능을 보였다. 또한, 첫 단계의 치료 권고 결과에 따라 생존율이 어떻게 달라지는 지 시뮬레이션 함으로써 2 단계 모델을 이용한 각 치료의 위험도를 층화시키는 실험도 수행하였다.

본 학위 논문의 세 번째 단계에서는 이러한 모델이 실제 임상 환경에서 어떻게 유용하게 사용될 수 있는 지 특정 시나리오들을 제시하였다. 첫 번째로, 본 모델이 현재 병기 시스템의 대안으로 사용될 수 있는 지 그 가능성을 살펴보는 의미로, 본 모델의 치료 추천 결과와 바르셀로나 임상 간암 병기(BCLC stage)의 치료 추천 결과를 BCLC C 병기의 환자들에서 실제 받은 치료와 비교하여 보았다. 그 결과 본 모델에서 권고한 치료와 실제 받은 치료의 일치도가 바르셀로나 임상 간암 병기에서 권고한 치료보다 높음을 확인할 수 있었다. 두 번째로는, 두 개의 서로 다른 센터의 데이터셋으로 훈련한 두 개의 모델에 대하여 다른 기관들의 데이터셋을 시뮬레이션해 보았다. 그 결과, 같은 조건을 가진 환자라도 각 기관의 데이터셋을 바탕으로 그 특성을 반영하여 서로 다른 치료를 권고하고, 다른 생존율을 보일 수 있다는 결과를 확인할 수 있었다. 이러한 활용 시나리오는 실제 임상 상황에서 본 모델을 확장하여 어떻게 사용할 수 있는 지에 대한 예시를 보여준다.

결론적으로 본 학위 논문에서는 20 개의 임상 변수를 이용하여 간세포 암종 환자에서 적절한 초기 치료를 권고하고 그에 따른 생존율을 예측해 주는 기계 학습 기반의 모델을 개발하였다. 뿐만 아니라 이 모델은 단일 기관이 아닌 국내의 여러 기관에서 사용할 수 있으려면 어떠한 점을 보완해야 하는 지 여러 가지 실험을 통해 검증하고 분석하였다. 본 CDSS 모델은 실제 임상 환경에서 경험이 적은 의사와 기관들에 실용적인 유용성을 제공해 줄 것으로 기대된다.

Acknowledgements

감사의 글

방사선 종양학과 전공의 4 년차 시절, 수련이 몇 개월 남지 않았던 어느 가을에 인공지능이 의료 분야에 대체 어떻게 쓰이고 있는지 궁금하여 한 달간 엿보기나 해 보자는 생각으로 김남국 교수님의 연구실 문을 두드리고, 파견 오게 되었을 때의 설레임이 생각납니다. 그리고, 한 달 간 모든 미팅들에 들어가며 내가 그동안 그렇게 공부해 보고 싶었던 분야가 이것이 맞는지 반문하고, 고민하고, 많은 분들과 상담한 끝에 큰 용기를 내어 늦은 나이에 공부를 시작하게 되었습니다.

저의 용기를 끊임없이 북돋아 주시고, 공학적으로 거의 백지 상태나 다름 없던 저에게 큰 기회를 주셨던 김남국 교수님께 감사하다는 말씀을 드리고 싶습니다. 전기전자 공학을 전공하고 관련 분야의 대기업에 근무하기는 하였지만 의학 분야로 새로이 들어와 근 10 여년을 정신 없이 새로운 지식들로 채우며, 사실 공학 지식은 의식의 저 너머로 넘어가 있었던 것 같습니다. 게다가, python, 머신러닝, 딥러닝은 학부 과정이나 근무 시절에도 전혀 다뤄보지 못했던 분야라 과연 나의 배경이 도움이 되는 지 어쩌는 지도 잘 느끼지 못하는 채로 처음 배우는 것과 같이 무작정 익혀나갔던 기억이 납니다. 새로운 지식을 습득하기에는 예전만큼 체력과 두뇌가 따라주지 못하는 것 같아 답답함을 느낄 때도 있었지만, 느린 속도로 한 걸음, 한 걸음 걷다 보니 아직도 많이 부족함을 느끼지만 처음보다는 이 분야와 많이 친숙해지고, 이것저것 해보고 싶은 것들도 많아짐을 느낍니다.

타과의 이제 갓 전문의가 된 의사와 함께 일하며 일반 공학자들에 비해 공학적인 실력은 많이 부족함에도 격려와 배려를 아끼지 않아주시어 이렇게 연구 결과를 낼 수 있도록 도와주신 임상과의 김강모, 최광현, 고범석 교수님께도 감사의 말씀 전합니다. 그리고 학위 심사 과정에서 날카롭고 유익한 조언으로 제 박사 논문이 좀 더 알차질 수 있도록 도와주신 서준범 교수님과 김휘영 교수님께도 감사드립니다.

이렇게 천천히 걸어오는 과정에서 제가 많은 질문을 여쭈어도 항상 차근차근 알려주신 윤지혜 교수님, 하나를 여쭤보면 항상 두 세가지 더 알려주시려 했던 김민규 박사님, 저와 똑같은 처지로 함께 헤매고 분투했던 장미소 선생님, 바로 전임 박사 선배님으로 연구실 박사 생활의 많은 부분들을 조언해 주신 함성원 박사님, 나도 저렇게 과학적으로 사고하는 박사가 되고 싶다고 느끼게 해 주신 황정은 박사님, 자주 뵙지 못했지만 항상 응원해 주셨던 이준구 교수님 등, 일일이 열거할 수 없이 많은 저희 연구실의 고마운 분들 덕분에 무사히 박사 생활을 마무리 할 수 있게 된 것 같습니다.

마지막으로, 제게 언제나 깊은 애정과 신뢰를 보내주고 격려해 주시는 오랜 친구들과 친지들, 항상 제게 큰 힘이 되어 주는 동생 주은이와 동호, 늦은 나이에 그토록 하기 싫다던 학문의 길을 뒤늦게 붙잡고 끙끙대는 모습에 답답하셨을 텐데도 꾹 참고 지원을 아끼지 않으시는 어머니, 그리고 제가 많은 암 환자들에게 도움이 되는 삶을 살고 싶다고 마음을 먹게 된 원인인 아버지께 깊은 감사의 마음을 전하며 글을 마칩니다.