Two-Stage Prediction of Comorbid Cancer Patient Survivability Based on Improved Infinite Feature Selection

PENG LIU, (Member, IEEE), AND SHUMIN FEI, (Member, IEEE)
Key Laboratory of Measurement and Control of CSE, School of Automation, Ministry of Education, Southeast University, Nanjing 210096, China
Corresponding author: Shumin Fei (smfei@seu.edu.cn)

ABSTRACT The modeling of comorbid cancer patients’ survivability has theoretical significance and practical needs. Cancer survivability prediction may provide guidance for clinical decision making and personalized medicine. The Surveillance, Epidemiology, and End Results (SEER) program provides large data sets suitable for analysis with machine learning methods. In this study, we consider survival prediction to be a two-stage problem. The first is to predict the five-year survivability of patients. For those whom the predicted outcome is ‘death’, the second stage predicts the remaining survival time. Male and female comorbid cancer cases (male-genital and urinary cancer for men and breast and female-genital cancer for women) were identified from the SEER database and labeled. In the classification stage, the dataset was processed with improved infinite feature selection (Inf-FS) and random undersampling-based data balancing. These two methods resolved the issues of biased data set and poor classification accuracy. In the lifespan prediction stage, unsupervised infinite feature selection (UnfFS) was applied. The results indicate that the proposed method is effective.

INDEX TERMS Cancer comorbidity, SEER, survival prediction, infinite feature selection, data balancing, unsupervised feature selection.

I. INTRODUCTION
Due to the promotion of cancer screening, developments in medical science, and improvements in supportive care, cancer prognosis has improved significantly. The 5-year cancer survival rate in 2016 was twice as high as that in 1950 [1]. Cancer survivors have a higher risk of developing a second primary cancer, and the rate was estimated to be 14% higher than the probability of developing a primary cancer among people who had never had cancer [2]. The increasing number of cancer survivors and the aged population are the main causes for the increasing number of multiple primary cancer (MPC) patients. The concurrence of multiple different cancers is also called cancer comorbidity.

Comorbidity was defined by researchers to represent coexisting noncancer diseases [3], while others used the term to represent the comorbidity of cancer and noncancer diseases. In [4], comorbidity represented coexisting cancers. An investigation of historical cancer cases reveals that certain types of cancers have higher correlations. Urinary and male genital cancers are the most common codeveloping cancers; breast and female genital cancers are the third most common cancer comorbidity [5].

Cancer survival prediction has been a popular research interest. Accurately predicting patients’ survivability may provide doctors with medical advice and help prescribe personalized medicines. Survivability is the possibility of a patient being alive for more than five years from the time of cancer diagnosis. It is a medical indicator for evaluating treating effects. Most cancer survivability studies aim to predict patients’ five-year survivability. These studies provide a limited amount of information for medical decisions. If a patient’s prediction is ‘death’, the survival time of the patient remains unknown. Survival time prediction should be studied to provide more precise information for medical decision making [6].

Cancer survival studies are challenging due to the lack of large-scale medical data that are available to the public. The Surveillance, Epidemiology, and End Results (SEER) program is an open source database providing de-identified, coded, and annotated information on cancer statistics in the United States [7], [8]. The scale of the data
is large enough to be analyzed with machine learning techniques [9].

This article aims to predict the survival time on a monthly measure. However, the prediction of survival time has been shown to be challenging since large generalization errors often occur when one-stage regression models are used [10], [11]. To resolve this issue, a two-stage prediction model is proposed. At the first stage, a classifier is adopted to predict whether the patients can survive for more than five years. At the second stage, a regression model is used to predict the survival time of patients who have been classified as not being able to survive for five years.

In the classification stage, two problems are encountered: the problem of biased dataset and poor classification performance. To demonstrate the problem of bias, a survival time histogram is adopted to balance the training dataset. To deal with poor classification performance, improved infinite feature selection (Iinf-FS) is proposed to cascade with the support vector machine classifier. The comparing feature selection methods for two-stage classifiers are: CHI2, feature selection via eigenvector centrality (ECFS), and mutual information based feature selection [12]. These feature selection methods are available publically [13]. In the regression stage, the above improvements cannot be applied because the predicted outcome is continuous. However, the error rate is high, and the training time is long without data preprocessing. Moreover, Iinf-FS cannot be used because it requires class labels. Thus, an unsupervised infinite feature selection (UinfFs) method is adopted instead. UinfFs does not require class information. The comparing unsupervised feature selection methods for the regression stage are: Nonnegative Spectral Analysis (NDFS) [14], L2 1-norm regularized discriminative feature selection (UDFS) [15], [16], and PCA-based contribution matrix related feature selection [17]. The main contributions of this article are as follows:

- Consider the survivability prediction problem to have two stages.
- Create two cancer comorbid datasets from the SEER database: breast and female genital cancers for women and urinary and male genital cancers for men.
- Apply Iinf-FS to feature selection in the classification stage.
- Apply random undersampling to balance the training data in the classification stage.
- Apply UinfFs in the regression stage.
- Compare the two-stage classification and regression model with the one-stage regression model.

One-stage prediction is also compared in this study. It has the classification accuracy 81.76% and regression RMSE value 22.566. The performance of the one-stage method in both classification and regression tasks are worse than the proposed two-stage framework’s performance. In the classification stage, the prediction accuracy of improved two-stage method is 86.42%, which is higher than that of the original linear support vector machine (Linear-SVM) (82.52%), and other state of the art improving methods. In the second stage, the RMSE of the improved random forests (RF) method is 13.267, better than the original RF method’s RMSE of 13.529 and other feature selection methods.

This remaining article is organized as follows. Section 2 introduces the related works on cancer comorbid, data balancing, feature selection, data cleaning, and the application of machine learning techniques to cancer survival prediction. Section 3 provides the detailed methods and experimental procedures. Section 4 presents the experimental results while section 5 presents the discussion of the results. Section 6 concludes the paper and presents possible future research directions.

II. RELATED WORK

Lynch et al. predicted the survival prognosis of lung cancer patients via supervised machine learning techniques [7]. Researchers used machine learning models to predict breast cancer patients’ survivability [18]–[20]. A Gaussian k-base Naive Bayes (NB) classifier system was proposed by Kaviarasi et al. to enhance the lung cancer classification accuracy of the NB classifier and linear regression [21]. Few scholars have focused their attention on patients with cancer comorbidity. Urinary and male genital cancers are the most common cancer comorbidities, and breast and female genital cancers are the second most common concurring cancers. The prediction of these cancer comorbidities is essential in cancer studies. Here, cancer survivability refers to the possibility of a patient being alive after 60 months since the diagnosis of the second cancer.

Machine learning techniques have been widely applied in medical fields [18]–[20]. Liu et al. applied an improved clustering algorithm for sample cutting to improve the training sample category representation capability. The experimental results indicated that the method improves the classification efficiency [22]. Ryu et al. compared machine learning techniques with statistical methods to predict survivability in spinal ependymoma patients and discovered that lower grade histology and higher extent of surgical resection were the key prognostic factors; they concluded that therapeutic factors are associated with improved overall survival. Machine learning methods generally perform better in prediction tasks, but the datasets are heterogeneous and complex, with numerous missing values [23]. Researchers studied several recently published papers on breast cancer survival prediction. Stage-specific prediction tasks were performed using machine learning methods. Stage-specific prediction models and joint models were created and compared, concluding that the data-driven knowledge obtained with machine learning methods must be subject to over-time validation before it could be clinically and professionally applied [24].

Zolbanin et al. used cancer data collected from the SEER database to create two comorbid datasets: one for breast and female genital cancers and another for prostate and
urinary cancers. Several popular machine learning techniques were then applied to the resultant datasets to build predictive models. The results showed that having more information about the comorbid conditions of patients improved the predictive power of the model [5]. In addition to Zolbanin’s study, Naghizadeh et al. also conducted research on comorbid cancer patients focusing on data preprocessing, including feature selection and data cleaning. The feature selection procedure was performed by applying the least angle regression, least absolute shrinkage and selection operator, and stepwise regression algorithms. The accuracy was increased, and the error rate was decreased [25]. Some researchers considered the survival prediction of colorectal cancer patients in the SEER database to be a two-stage process: the first stage was to predict survival, and the second was to predict the remaining life span of the patients whose predicted outcome was ‘death’. The first stage adopted a tree ensemble classification method that took into account the imbalanced data. The regression stage used a tree-based selective ensemble regression method called SRRT-SEM [6].

II. MATERIALS AND METHODS
SEER is an open-source database that provides deidentified, coded, and annotated information on cancers in the United States [26]. The database is large enough to provide plenty of cases to be analyzed with machine learning techniques. The cases were selected from the Incidence 0 SEER 18 Regs Research Data, November 2018 Sub database. The cancer diagnoses in the SEER cancer registries were verified clinically or microscopically by a recognized medical practitioner by law [27].

A. TWO-STAGE PREDICTION
Most cancer prognosis studies are limited to predicting whether a patient can live for a specific amount of time. The patient is then classified as ‘survived’ or ‘dead’. Since gastric cancer has a high mortality incidence, most cases would be classified as ‘dead’. The survival time for these patients remains unknown. Therefore, we propose a two-stage classification model consisting of a classification model that predicts the patient’s survivability, and a regression model that predicts the remaining lifespan of the patients whose predicted outcome is ‘dead’ [6].

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

In the classification stage, two problems are encountered. The first problem is that the biased training set would result in a biased classifier. The cases from the minority class would be misclassified as the majority class. To solve this problem, data balancing is required. Random undersampling is used to balance the datasets. The second problem is that the feature pool is large and the classification result is poor. Inf-FS is applied to the the cascade with a support vector machine classifier selecting a subset of features from the pool. The cascaded system has better classification performance than the original classifier.

The steps of the classification framework are as follows:
1) Collect the data on breast, urinary, male genital and female genital cancers from the SEER database.
2) Combine the data and rearrange the order of the data.
3) Divide the data into training, validating, and testing sets.
4) Select optimal features for modeling according to Inf-FS.
5) Apply random undersampling to balance the biased training set.
6) Apply the linear-SVM classifier to the prediction.
7) Evaluate the predicted outcomes with error criteria: accuracy, f-score, specificity and sensitivity.

The procedures of the regression framework are as follows:
1) Remove cases in the classification data whose survival months are greater than 60.
2) Divide the data into training, validating, and testing sets.
3) Select optimal features for modeling according to UnifFS.
4) Apply RF regressor to the prediction.
5) Evaluate the predicted outcomes with error criteria: root mean squared error(RMSE), mean absolute error(MAE), and $R^2$ score.

B. CREATING THE DATASETS
Breast, female genital, male urinary, and male genital cancer cases were extracted from the SEER data registries. Cases with ‘positive histology’, ‘complete dates’, and ‘active follow-up’ were chosen, while cases with ‘autopsy only’, ‘death certificate only’, ‘unknown cause of death’, and ‘unknown stage’ were excluded. Benign and in situ cancer patients were excluded since they have a higher likelihood of being cured and should be treated differently. There were 72,121 urinary cancer cases, 47,795 male genital cancer cases, 99,443 breast cancer cases, and 28,539 female genital cases. Searching for the same patient ID in urinary and male genital cancer datasets, 2,699 cases of male patients with multiple primary cancer were found. In the breast cancer and female genital cancer datasets, 1,275 cases of female patients had the same patient ID. Among all 3975 patients with cancer comorbidities, 1,062 lived more than 5 years after the diagnosis of the second primary cancer. The training and testing data were then separated at a ratio of 4:1. A total of 2917 patients who lived for more than 60 months were excluded from the regression stage dataset.

Figure 1 shows the distribution of the patients’ survival months. Approximately one-fourth of the patients lived for more than 60 months since the diagnosis of the first incidence of cancer.

Seer registry provides many attributes but only few are related to this research. Table 1 lists the attributes selected for this study. 25 features are selected in the base datasets.
C. IMPROVED INFINITE FEATURE SELECTION

1) THE IMPROVED SUPERVISED FEATURE SELECTION

Feature selection is the process of selecting a subset from the feature group. In the application of machine learning, there is always a large number of features in the feature pool. Some features may be irrelevant to the prediction label, and some may be dependent on each other. The irrelevant and redundant features may lead to a lengthy training time, over complicated models, and it is hard to be generalized. In this study, all the data are one-hot encoded. Forty-four features become 1297-dimensional features consisting of only zeros and ones. Regarding the dimension of the data, feature selection is necessary.

Given a training set \( F \) represented as \( F = \{f^{(1)}, \ldots, f^{(n)}\} \), a nondirected fully connected graph \( G = (V, E) \) can be built. The vertices \( V \) correspond to all features, while the edges \( E \) represent the pairwise relations between features. \( G \) can be represented as an adjacency matrix \( A \) whose elements \( a_{ij} \) represent the pairwise relationship between features [28]. The elements are called pairwise probabilistic energy terms, defined as:

\[
 a_{ij} = \alpha \sigma_{ij} + (1 - \alpha)c_{ij} \tag{1}
\]

where \( \alpha \in [0, 1] \),

\[
 \sigma_{ij} = \max((d(i) - d(j)) \tag{2}
\]

\[d(i) = (p(c^+) \sigma^+_i + p(c^-) \sigma^-_i) \]

\[
 \sigma^+_i \text{ is the normalized standard deviation of elements that belong to } f^{(i)} \text{ and the positive class, and } \sigma^-_i \text{ is the standard deviation of elements that belong to the same feature but the negative class. } p(c^+) \text{ and } p(c^-) \text{ are the corresponding probabilities of each class. A large dispersion between classes implies a large distance between samples in different classes. Therefore, to classify samples, we need to maximize the term } d(i). \]

The first term \( d(i) \) measures the dispersion between positive and negative classes. The second term is \( c_{ij} = 1 - |\text{corr}| \) with \( \text{corr} \) representing Spearman’s correlation coefficient between \( f^{(i)} \) and \( f^{(j)} \).

\[
 \sigma_{ij}, \alpha c_{ij} \in [0, 1], \text{ so the two terms are comparable in magnitude. } a_{ij} \text{ takes into account two features and measures their feature dispersion and their correlation. Defining } \gamma_{ij} \text{ denotes a path of length } l \text{ between nodes } i \text{ and } j \text{ through other features, we can then estimate the probability as}
\]

\[
P_{\gamma} = \prod_{k=0}^{l-1} a_{\nu_k, \nu_{k+1}} \tag{3}
\]

To account for the energy of all paths of the same length \( l \), \( R_{l(i, j)} \) is defined as the set of all paths of \( l \) between \( i \) and \( j \).

\[
 R_{l(i, j)} = \sum_{\gamma \in R_{l(i, j)}} P_{\gamma} \tag{4}
\]

which is equivalent to:

\[
 R_l = A^l, \tag{5}
\]

A single feature’s energy score can be estimated by summing all paths of all possible lengths of \( l \) and expanding \( l \) to infinity.

\[
 \hat{S} = \sum_{l=1}^{\infty} A^l \tag{6}
\]
TABLE 2. Regression stage feature selection.

| Feature              | Rank | RMSE  | Select | Feature              | Rank | RMSE  | Select |
|----------------------|------|-------|--------|----------------------|------|-------|--------|
| DerivedS19772        | 1    | 13.543| Yes    | Tumorsize1           | 21   | 13.522| No     |
| Grade1               | 2    | 13.65 | Yes    | PrimarySite1         | 22   | 13.47 | No     |
| M1                   | 3    | 13.579| Yes    | ScopeRegLNStat1      | 23   | 13.566| Yes    |
| Race1                | 4    | 13.53 | Yes    | HistologicType1      | 24   | 13.54 | Yes    |
| Histology1           | 5    | 13.576| Yes    | LYmphnodes1          | 25   | 13.61 | Yes    |
| N1                   | 6    | 13.533| Yes    | NumberofTumor        | 26   | 13.559| Yes    |
| Age1                 | 7    | 13.652| Yes    | PrimarySite2         | 27   | 13.53 | Yes    |
| Sequence1            | 8    | 13.5  | Yes    | Stage2               | 28   | 13.54 | Yes    |
| Sequence2            | 9    | 13.589| Yes    | T2                   | 29   | 13.6  | Yes    |
| Yearofdiagnosis1     | 10   | 14.552| Yes    | ScopeRegLNStat2      | 30   | 13.49 | No     |
| Stage1               | 11   | 13.498| Yes    | Tumorsize2           | 31   | 13.5  | No     |
| Marital1             | 12   | 13.51 | Yes    | Grade2               | 32   | 13.5  | No     |
| T1                   | 13   | 13.588| Yes    | N2                   | 33   | 13.55 | Yes    |
| State1               | 14   | 13.42 | No     | HistologicType2      | 34   | 13.581| Yes    |
| SurgPrimSite1        | 15   | 13.596| Yes    | SurgPrimSite2        | 35   | 13.39 | Yes    |
| RecordNumber2        | 16   | 13.523| No     | NumberofTumor2       | 36   | 13.54 | Yes    |
| Yearofdiagnosis2     | 17   | 14.14 | Yes    | First1               | 37   | 13.547| Yes    |
| Sex1                 | 18   | 13.53 | Yes    | DerivedS19771        | 38   | 13.61 | Yes    |
| RecordNumber1        | 19   | 13.539| Yes    | Histology2           | 39   | 13.554| Yes    |
| LYmphnodes2          | 20   | 13.548| Yes    | M2                   | 40   | 13.591| Yes    |

However, if $A$ is divergent, adding infinite $A^l$ together results in divergence. A regulatory factor is needed. It ensures divergence of equation (7).

$$\bar{S} = \sum_{l=1}^{\infty} r^l A^l$$  \quad (7)

Through matrix algebra it can be efficiently solved as:

$$\bar{S} = (I - rA)^{-1} - I$$  \quad (8)

The final scores for each feature term are calculated as follows:

$$\tilde{s}(i) = [\bar{S}e]_i$$  \quad (9)

where $e$ is an array of ones. [12]

This method not only considers the classification information of the features but also takes into account the positive/negative correlation between features and classes.

2) THE UNSUPERVISED INFINITE FEATURE SELECTION

In the regression stage, the prediction target is continuous that does not belong to specific classes. Thus, an unsupervised feature selection method is needed. Inf-FS from the previous subsection is a supervised method. Equation (10) is the same as Equation (1), it has two components.

$$a_{ij} = \alpha \sigma_{ij} + (1 - \alpha)c_{ij}$$  \quad (10)

$\sigma_{ij}$ represents the dispersion between features, $c_{ij}$ represents the correlation between features. Unsupervised feature selection process does not provide labels in the calculation. The correlation part requires no class labels but the dispersion part does. Thus, the dispersion part is removed to be unsupervised.

$$a_{ij} = c_{ij}$$  \quad (11)

The rest of the ranking calculation remains the same. Repeating equation (3) to equation (9), the scores of the features are calculated. Sorting the scores, the final ranking is obtained.

The calculation of the ranking does not include the prediction target. Thus the relevance of the reduced feature subset to the predicting label is not measured. An element-wise feature dropping method to measure the importance of each feature is adopted to resolve this issue. One feature component is excluded from the subset in each round. The root mean squared error (RMSE) of each round is calculated and compared with the original predictive error when all feature components are considered. If the RMSE decreases, the feature component can be dropped. The features ranks high in the Unsupervised-InfFS process and high RMSE values remain in the final feature subset. In this study, the top ten features are kept regardless of their dropping error. Other features can be dropped if the RMSE decreased when it was removed from the feature pool.

Table 2 presents the feature contribution score obtained, the RMSE if each feature is dropped from the feature groups, and the feature selection result.

D. RANDOM UNDERSAMPLING FOR BIASED CLASS BALANCING

As mentioned in section III.B, the classification data are biased such that only 1/4 of the cases are positive. When training a model with biased data, the classifier would show the bias. Different methods can be used to deal with biased data [29]. The two most common methods are undersampling and oversampling. In oversampling, the cases in the minority class are duplicated to balance the dataset. In undersampling, the bias in the dataset is controlled by removing samples from the majority class. Random undersampling is simple and has shown similar performance to other complex methods [29]. In this study, we use a random undersampling method to alleviate the bias in the dataset.
A random value between 0 and 1 is assigned to each training case. If the case belongs to the majority class and the randomly assigned value is greater than 0.3, this case is deleted from the training dataset.

**E. EXPERIMENTAL PROCEDURES**

A total of 3180 and 795 cases were included in the training and testing datasets, respectively. The dividing ratio was 4:1. When combining the features of the two cancers, some features were identical. Excluding identical features from the combined feature pool, 40 features were selected and transformed to one-hot encoded 1297-dimensional data consisting of zeros and ones only. In the classification stage, Iinf-FS reduced the data dimensionality, while the random undersampling process reduced the training cases. Linear SVM was adopted as the classifier. The compared feature selection methods in the classification stage are: CHI2 score based feature selection, ECFS, and mutual information based feature selection [13]. In the regression stage, patients who survived longer than 60 months were removed from the whole dataset. The unsupervised Inf-FS was applied because its unsupervised nature is compatible with the regression process. The compared unsupervised feature selection methods include: UDFS, NDFS [16], and PCA based feature selection [17]. These comparing methods are also utilized with element-wise feature dropping RMSE score. The top ten features are kept. Other features are dropped if the RMSE scores decrease when they are removed from the pool.

In each iteration, the training set trains the classifier and the accuracy score of the testing set is stored to be compared.

**F. ALGORITHMS APPLIED**

The classification utilized linear SVM as the classifier. It constructed a separating hyperplane between two classes and then classifies samples based on distance. The data has been one-hot encoded, consisting of zeros and ones only. Linear SVM satisfied the requirement for separating one-hot encoded data. It cascaded with Iinf-FS and random undersampling. The regression stage utilized RF as the regressor. As a commonly used bagging regressor, it was cascaded with contribution score based feature selection.

**G. PERFORMANCE METRICS**

The classification accuracy is quantified as recognition accuracy, specificity, and sensitivity. The formulas are as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (13)
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)
\]

**IV. RESULTS**

The performance for both stages are included in this section. Table 3 is the confusion matrix for the classification results. As mentioned in the previous sections, the improvement of the classification consists of improved infinite feature selection (Iinf-FS) and random undersampling-based data balancing. The one-stage regression-based classification performance is also included to be compared with.

**TABLE 3. Classification outcome vs. actual outcome.**

| Methods                 | Actual 0 | Actual 1 | Predicted |
|-------------------------|----------|----------|-----------|
| One-stage Regression    | 518      | 80       | 0         |
|                         | 65       | 132      | 1         |
| Two-stage RF original   | 505      | 78       | 0         |
|                         | 61       | 151      | 1         |
| + Inf-FS                | 511      | 72       | 0         |
|                         | 46       | 166      | 1         |
| +Inf-FS +Undersamp      | 503      | 80       | 28        |
|                         | 28       | 184      | 0         |

**TABLE 4. Performance metrics for classification stage.**

| Model                   | 2-stg SVM | CHI2 | ECFS | Mote | Inf-FS |
|-------------------------|-----------|------|------|------|--------|
| Acc                     | 0.8232    | 0.844| 0.8516| 0.8428| 0.8516 |
| Spec                    | 0.8662    | 0.8748| 0.8765| 0.8662| 0.8765 |
| Sens                    | 0.7123    | 0.7559| 0.783 | 0.7783| 0.783  |
| FeatN                   | N.A.      | 499  | 150  | 350  | 150    |
| Model                   | 1-stg     | +Under| +Under| +Under| +Under |
| Acc                     | 0.8176    | 0.8503| 0.8604| 0.8453| 0.8642 |
| Spec                    | 0.889     | 0.8679| 0.8525| 0.8419| 0.8628 |
| Sens                    | 0.6226    | 0.8019| 0.8821| 0.8349| 0.8679 |
| T.Samp                  | 3180      | 2944  | 2477 | 2724 | 2899   |
| Time                    | N.A.      | 7.55s | 12.58s| 8.204s| 10.48s |

Table 4 includes the classification performance metrics for our proposed two-stage linear SVM improved with Inf-FS and data balancing with comparing methods. The compared methods consist of a one-stage RF regression-based classifier, two-stage linear SVM improved with CHI2-based feature selection [30], two-stage linear SVM with ECFS [31], and two-stage linear SVM with mutural information-based feature selection [32]. Random undersampling-based data balancing is applied to all the two-stage classifiers. Except for the general performance metrics, including accuracy, specificity and sensitivity, the number of selected features for each feature selection method, the number of training samples after undersampling, and the data preprocessing time are recorded and compared.

Figure 2 contains the scatter plots for original two-stage random RF, UinfFS based RF, UDFS based, NDFS based, and PCA based feature selected RF for comparison. One-stage regression’s scatterplot is also included. The corresponding RMSE, MAE, and $R^2$ score values are in Table 5.

**V. DISCUSSION**

**A. CLASSIFICATION STAGE**

Table 3 is the confusion matrix for the one-stage regressor-based classifier, two-stage original SVM classifier, Inf-feature selected data with SVM classifier, and SVM with both improving procedures. The one-stage regressor-based classifier predicts patient survival time. The labels are then classified as 0 for patients whose prediction is less than
60 months and 1 for patients whose predicted survival time is greater than 5 years. From the confusion matrix, it can be seen that the one-stage regressor-based classifier is good at classifying the majority of patients. However, for the minority of patients whose label is 1, it misclassified 80 out of 212 patients to be 0. Slightly better than the one-stage model, the unimproved two-stage linear SVM classifier selected 518 out of 583 ‘dead’ cases correctly. However, it did not handle the ‘survived’ cases well, as 61 out of 212 positive patients were mistakenly classified as ‘dead’ patients. The biased result was caused by the bias in the training set. Applying Inf-FS improves the classification performance among both ‘survived’ and ‘dead’ patients. The number of misclassified cases decreased in both classes. The feature selection process improves the classification accuracy, while data balancing deals with biased datasets. Data balancing causes negative effects on cases with negative labels in the majority class. More negative patients are classified as positive. The classification performance of the minority class improves significantly when random undersampling is applied. These two improving procedures solved the two problems in the classification stage. The final model is unbiased and has the best accuracy score.

Table 4 confirms the above observation. The one-stage model’s accuracy is much weaker than that of the two-stage unimproved method. Both the accuracy score and sensitivity score are lower than those of the two-stage model. Inf-FS was compared with state-of-the-art feature selection methods. All these feature selection methods can improve the classification performance. Among them, ECFS and the proposed Inf-FS method had the best performance. After applying random undersampling-based data balancing to the training set, Inf-FS obtained the best accuracy score and sensitivity score. ECFS was the second best, but the execution time for ECFS was longer than that for Inf-FS.

Comparing the proposed method’s results with the results from [5], the classification accuracy (86.42% for all patients) is much better than the results from [5] (77.8% for breast-female genital and 73.48% for male genital-urinary).

| Model | 1-stage | 2-stage | UDPS | NDFS | PCA | Inf |
|-------|---------|---------|------|------|-----|-----|
| RMSE  | 22.566  | 13.539  | 13.478 | 13.38 | 13.39 | 13.255 |
| MAE   | 15.574  | 10.34   | 10.332 | 10.34 | 10.31 | 10.227 |
| R2    | 0.07199 | 0.319   | 0.329  | 0.324 | 0.324 | 0.342 |

**B. REGRESSION STAGE**

In the regression stage, the RMSE, MAE, and $R^2$ score are calculated to compare models. To match the results for the one-stage regressor, patients whose actual survival times were greater than 60 months were excluded. Table 5 shows that all the performance metrics for the one-stage regressor are worse than those of the two-stage method. In the previous subsection, the classification performance of the one-stage model is also weaker than that of the two-stage model. This result confirms the necessity for the proposed two-stage method. As shown in Table 5, the proposed UInfFS method
is compared with the state-of-the-art unsupervised feature selection method. The UnifFS two-stage regressor has the lowest RMSE, MAE and the highest R². This method is also easy to understand. It reduces the dimensionality of the data before the one-hot encoding procedure, so the deleted features and the remaining features can be named and visualized.

In Figure 2, the UnifFS method has the fewest outliers and shows linearity. The UDFS, NDFS methods have similar scatterplots but more outliers. The PCA based feature selection method has the weakest performance among comparing methods. The one-stage method looks messier than two-stage method. The RMSE values in Table 5 confirm our observations.

VI. CONCLUSION
Male genital-urinary and breast-female genital cancers are common MPCs. Predicting the survivability of MPC patients can help doctors, patients, and families. However, researchers have paid very little attention to MPC patient survival prediction. Thus, the survival prediction of MPC patients has become essential in cancer studies.

The proposed two-stage model provides even more information than survivability. The first stage predicts patients’ five-year survivability. If the prediction is ‘death’, which is common in cormbord cancer cases, the second stage predicts the remaining lifespan of the patient.

In this research, male genital-urinary and breast-female genital cases were identified and labeled to study survival prediction. In the classification stage, Unif-FS was carried out, and random undersampling-based data balancing was applied. In the regression stage, UnifFS was applied in the feature selection part. The results indicate that each proposed step improves the predictive power.

This study focused on genital MPC and proposed a novel feature selection. The feature selection can be further improved by studying interdisciplinary and intradisciplinary dispersions. In the future, we will keep studying feature selection methods that may improve current prediction performance. We may also study other MPCs such as second primary lung cancers.

REFERENCES
[1] N. Howlader, (Apr. 2019). Seer Cancer Statistics Review, 1975–2016. SEER Data Submission. Posted to the SEER Web Site. Accessed: Nov. 2018. [Online]. Available: https://seer.cancer.gov/csr/1975_2016/
[2] R. E. Curtis, New Malignancies Among Cancer Survivors: SEER Cancer Registries, 1973–2006, no. 2. Washington, DC, USA: US Department of Health and Human Services, National Institutes of Health, 2006.
[3] C. Diederichs, K. Berger, and D. B. Bartels. “The measurement of multiple chronic diseases–A systematic review on existing multimorbidity indices,” J. Gerontology Ser. A, Biol. Sci. Med. Sci., vol. 66A, no. 3, pp. 301–311, Mar. 2011.
[4] B. K. Edwards, A.-M. Noone, A. B. Mariotto, E. P. Simard, F. P. Boscoe, S. J. Henley, A. Jemal, H. Chou, R. N. Anderson, B. A. Kohler, C. R. Ehemann, and E. M. Ward. “Annual report to the nation on the status of cancer, 1975–2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer,” Cancer, vol. 120, no. 9, pp. 1290–1314, May 2014.
[5] H. M. Zolfbanin, D. Delen, and A. Hassan Zadeh, “Predicting overall survivability in comorbidity of cancers: A data mining approach,” Decis. Support Syst., vol. 47, pp. 151–161, Jun. 2015.
[6] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, “A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction,” Inf. Sci., vol. 474, pp. 106–124, Feb. 2019.
[7] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgeman, V. H. van Berkel, and H. B. Frieboes, “Prediction of lung cancer patient survival via supervised machine learning classification techniques,” Int. J. Med. Inform., vol. 108, pp. 1–8, Dec. 2017.
[8] NCI SEER Overview. (2015). Overview of the SEER Program. Surveillance Epidemiology and end Results. [Online]. Available: http://seer.cancer.gov/about/
[9] P. Liu, L. Li, C. Yu, and S. Fei, “Two staged prediction of gastric cancer patient’s survival via machine learning techniques,” in Proc. 7th Int. Conf. Intell. Appl., 2020, pp. 105–116, doi: 10.5121/ciit.2020.100308.
[10] B. Garzín, K. E. Emblem, K. Mouridsen, B. Nedregaard, P. Due-Tønnessen, T. Nome, J. K. Hald, A. Bjørnerud, A. K. Håberg, and Y. Kvinsland, “Multiparametric analysis of magnetic resonance images for glioma grading and patient survival time prediction,” Acta Radiologica, vol. 52, no. 9, pp. 1052–1060, Nov. 2011.
[11] I. H. E. A. T. Magome and A. Haga, “TH-E-BRF-05: Comparison of survival-time prediction models after radiotherapy for high-grade glioma patients based on clinical and DVH features,” Med. Phys., vol. 41, no. 33, p. 570, 2014.
[12] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, “Infinite latent feature selection: A probabilistic latent graph-based ranking approach,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1398–1406.
[13] Giorgio, (Jun. 24, 2020). Feature Selection Library. MATLAB Central File Exchange. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection
[14] L. J. E. A. Z. Li and Y. Yang, “Unsupervised feature selection using nonnegative spectral analysis,” in Proc. 26th AAAI Conf. Artif. Intell., Jul. 2012, pp. 1026–1032.
[15] Y. Yang, H. T. Shen, and Z. Ma, “|2,1-norm regularized discriminative feature selection for unsupervised,” in Proc. 2nd Int. Joint Conf. Artif. Intell., 2011, pp. 1–6.
[16] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” ACM Comput. Surv., vol. 50, no. 6, p. 94, 2016.
[17] F. Song, Z. Guo, and D. Mei, “Feature selection using principal component analysis,” in Proc. Int. Conf. Conf. Syst. Sci., Eng. Design Manuf. Informatica, Nov. 2010, pp. 27–30.
[18] Y.-Q. Liu, C. Wang, and L. Zhang, “Decision tree based predictive models for breast cancer survivability on imbalanced data,” in Proc. 3rd Int. Conf. Bioinf. Biomed. Eng., Jun. 2009, pp. 1–4.
[19] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, “Breast cancer survivability via adaboost algorithms,” in Proc. 2nd Australas. Workshop Health Data Knowl. Manage., vol. 80, 2008, pp. 55–64.
[20] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, “Robust predictive model for evaluating breast cancer survivability,” Eng. Appl. Artif. Intell., vol. 26, no. 9, pp. 2194–2205, Oct. 2013.
[21] R. Kavirarasi, “Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system,” J. Med. Syst., vol. 43, no. 7, p. 201, Jul. 2019.
[22] H. Liu, Z. Su, and S. Liu, “Improved chi text feature selection based on word frequency information,” Comput. Eng. Appl., vol. 49, no. 22, pp. 110–114, 2013.
[23] S. M. Ryu, S.-H. Lee, E.-S. Kim, and W. Eoh, “Predicting survival of patients with spinal epidymnoma using machine learning algorithms with the SEER database,” World Neurosurg., vol. 124, pp. e331–e339, Apr. 2019.
[24] R. Kleinlein and D. Riaño, “Persistence of data-driven knowledge to predict breast cancer survival,” Int. J. Med. Inform., vol. 124, pp. 303–311, Sep. 2019.
[25] M. Naghizadeh and N. Habibi, “A model to predict the survivability of cancer comorbidity through ensemble learning approach,” Expert Syst., vol. 36, no. 3, Jun. 2019, Art. no. e12392.
[26] N. M. Donin, L. Kwan, A. T. Lenis, A. Drakakis, and K. Channie, “Second primary lung cancer in united states cancer survivors, 1992–2008,” Cancer Causes Control, vol. 30, no. 5, pp. 465–475, May 2019.
[27] R. J. M. Adamo and L. Dickie, “SEER program coding and staging manual,” in U.S. Department of Health and Human Services National Institutes of Health National Cancer Institute. Bethesda, MD, USA: National Cancer Institute, 2018, Art. no. 20892.
[28] G. Roffo, S. Melzi, and M. Cristani, “Infinite feature selection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4202–4210.

P. Liu, S. Fei: Two-Stage Prediction of Comorbid Cancer Patient Survivability Based on Improved Infinite Feature Selection
[29] N. Japkowicz, “The class imbalance problem: Significance and strategies,” in Proc. Int. Conf. Artif. Intell., 2000, pp. 111–117.

[30] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, “Reliable Parkinson’s disease detection by analyzing handwritten drawings: Construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model,” IEEE Access, vol. 7, pp. 116480–116489, 2019.

[31] G. Roffo and S. Melzi, “Features selection via eigenvector centrality,” in Proc. New Frontiers Mining Complex Patterns (NFMCP), Oct. 2016, pp. 1–12.

[32] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

PENG LIU (Member, IEEE) was born in 1990. She received the bachelor’s degree from the Swanson School of Engineering, University of Pittsburgh, in 2012, and the master’s degree from the School of Electrical Engineering, University of Washington, in 2014. She is currently pursuing the Ph.D. degree with Southeast University. Her research interests include data mining, medical data analysis, medical image analysis, machine learning algorithms, and deep learning.

SHUMIN FEI (Member, IEEE) was born in 1961. He received the Ph.D. degree from Beihang University, Beijing, China, in 1995. From 1995 to 1997, he did Postdoctoral Research in the Research Institute of Automation, Southeast University, Nanjing, China, where he is currently a Professor. His research interests include the analysis and synthesis of nonlinear systems, robust control, adaptive control, and the analysis and synthesis of time delay systems. ** ** **