Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# PREDICTION OF COMORBID MALIGNANCY PATIENT SURVIVABILITY –EMPIRICAL PERSPECTIVE

**DR Y PADMA[1], DR NIDAMANURU SRINIVASA RAO[2], MR. PAVAN KUMAR KOLLURU[3], DR C ASHOK KUMAR MS. SHAIK SALMA BEGUM[5] DR. SURESH CHANDANAPALLI[6] KODEPOGU KOTESWARA RAO***

[1]Asst Professor, IT Dept, PVP Siddhartha Institute of Technology, Vijayawada,
[2]Associate Professor Dept of CSE, Narasimha Reddy Engineering College, Secunderabad
[3]Asst Professor, CSE Dept, VFSTR Deemed to be University, Guntur,
[4] Asst Professor, Dept Of Computing Technologies, School Of Computing, SRM Institute of Science and Technology Kattankulathur Chennai,
[5]Asst Professor, CSE Dept, SR Gudlavalleru Engineering Collge, Gudlavalleru,
[6] Professor, IT Dept, SR Gudlavalleru Engineering Collge, Gudlavalleru,
[*]Dept. of CSE, PVP Siddhartha Institute of Institute of Technology, Vijayawada, India,

E-mail: padmayenuga@pvpsiddhartha.ac.in, , rao75nidamanuriu@gmail.com , pavanwithu@gmail.com, ashokkuc@smrist.edu.in , shaiksalma.gec@gmail.com , sureshdani2004@gmail.com , koteswara2003@yahoo.co.in

## ABSTRACT

Modeling the survivability of comorbid cancer patients has both theoretical and practical implications. Cancer is one of the leading causes of death worldwide. Stomach, Liver, Thyroid, Lung and Skin Cancers are some of the most frequent cancers. The detection and prevention of these malignancies are crucial goals. According to recent discoveries, some people have cancer comorbidity. A number of studies have shown poorer survival among cancer patients with comorbidity. Several mechanisms may underlie this finding. The majority of studies found that cancer patients with comorbidity had a lower 5-year survival rate than those without, with hazard ratios ranging from 1.1 to 5.8. Only a few studies looked into the impact of specific chronic illnesses. Comorbidity does not appear to be linked to more aggressive cancers or other abnormalities in tumor biology in general. Another conclusion was that patients with comorbidity are less likely to obtain standard cancer therapies such surgery, chemotherapy, and radiation therapy, and their chances of completing a course of treatment are reduced. Predicting cancer survival may help with clinical decision-making and tailored therapy. Large data sets appropriate for machine learning analysis are available through the Surveillance, Epidemiology, and End Results (SEER) program. We regard survival prediction to be a two-stage problem in our study. The first is to forecast a patient's five-year survival rate. The second stage calculates the remaining survival time for individuals whose anticipated outcome is 'death.' The SEER database was used to identify and label male and female comorbid cancer cases (Stomach, Lung, Liver, Thyroid and Skin Cancers). The dataset was handled utilizing CHI2- based feature selection throughout the classification stage. These two solutions tackled the problems of a skewed data set.

**Keywords:** *CHI2,SEER, COMORBID, Survivability, Empirical Study*

## 1. INTRODUCTION:

Cancer prognosis has improved dramatically as a result of increased cancer screening, advances in medical knowledge, and improvements in supportive care. In 2016, the 5-year cancer survival rate was double that in 1950. Cancer survivors have a higher risk of having a secondary cancer, which is estimated to be 14% higher than the risk of developing a primary cancer in persons who have never had cancer. Multiple primary cancer (MPC) patients are on the rise as a result of an increasing number of cancer survivors and an ageing population. Cancer comorbidity refers to the presence of numerous cancers at the same time. [1-5]

Cancer survival prediction is a popular topic of study. Predicting patients' chances of survival accurately could help doctors give better medical advice and prescribe more tailored medications.

Survivability refers to a patient's ability to live for more than five years after being diagnosed with cancer. It's a medical metric for assessing treatment outcomes. The majority of cancer survival studies try to forecast patients' five-year survival rates. These studies only provide a small quantity of data to help doctors make decisions. If a patient's prognosis is 'death,' the patient's survival time is unclear. To provide more exact information for medical decision-making, survival time prediction should be investigated[6].

The paucity of large-scale medical data available to the public makes cancer survival research difficult. The SEER program (Surveillance, Epidemiology, and End Results) is an open-source database that provides de-identified, coded, and annotated data on cancer statistics in the United States. Machine learning techniques can be used to analyze the data because it is huge enough.

The goal of this essay is to forecast survival time on a monthly basis. When one-stage regression models are applied, however, substantial generalization errors frequently occur, making survival time prediction difficult. A two-stage prediction model is offered as a solution to this problem. A classifier is used in the first stage to estimate whether the patients would live for more than five years. A regression model is employed in the second stage to forecast the survival time of patients who have been identified as not having a five-year survival rate. CHI2 feature selection using eigenvector centrality (ECFS), and mutual information-based feature selection are the methodologies for comparing feature selection methods for two-stage classifiers. These methods for selecting features are open to the public. Because the anticipated outcome is continuous, the foregoing enhancements cannot be made during the regression stage. However, without data pre-treatment, the error rate is significant, and the training time is considerable. [7-9]

## 2. LITERATURE SURVEY

[10]Y. Wang, et'al proposed A tree ensemble based two-stage model for advanced-stage colorectal cancer survival prediction. The majority of existing data-driven cancer survival prediction studies use classification to predict whether a patient will live for more than five years. The prediction results obtained in this manner, however, are not precise enough to support medical decision-making. For example, in the five-year survivability classification, the exact outcome (survival time) of patients classified as negative (unable to survive more than five years) is unknown, which deserves more attention, particularly for high mortality cancers. Survival time prediction can be used to make more precise predictions, which is more difficult but also more meaningful for medical doctors. Traditional studies commonly use statistical tools to build prediction models based on survival-related factors such as palliative prognostic score, palliative performance index, and cancer, intra-hospital cancer mortality risk model and prognostic score However, keep in mind that the above statistically-based prediction models are for terminal cancer patients whose survival time is less than one month in order to provide proper support.[10]

The goal of this paper was to use machine learning methods to predict survival time on a monthly basis, which can aid in making effective treatment decisions. So far, it has been demonstrated that predicting survival times is extremely difficult because large generalization errors frequently occur when one-stage regression models are used. To address this challenge, we propose a two-stage model based on tree ensembles for cancer survival prediction, in which an effective classifier is used in the first stage to predict whether patients can survive for five years, and a novel regression tree ensemble is used in the second stage to predict the specific survival time for patients who are predicted to be unable to survive for five years.

[11]Kaviarasi, R et'al proposed Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system. Measurable classifier and great precision are a fundamental piece of the exploration in clinical information mining. Exact forecast of cellular breakdown in the lungs is a fundamental stage for pursuing powerful clinical choices. Subsequent to recognizing the cellular breakdown in the lungs, least degrees are accessible in the prescriptions for patient living on the planet. Hemoglobin level and TNM stage wise patient's endurance period must be fluctuated. Some gathering endurance period is insignificant and one more gathering endurance time is extended. This study is meant to foster a forecast model with new clinical factors to anticipate cellular breakdown in the lungs patients. It depends on modified eighth version investigation of TNM in cellular breakdown in the lungs. These new traits are gathered from SEER data sets, Indian malignant growth medical clinics and examination focuses. The gathered new traits are ordered utilizing regulated AI calculations of direct relapse, Naïve

Bayes classifier and proposed calculations of Gaussian K-Base NB classifier. Specifically, for TNM stage 1 gathering with typical hemoglobin level (NHBL), that gathering of cellular breakdown in the lungs patient personal satisfaction is profoundly improved. Which demonstrated by utilizing managed AI calculations. The proposed calculation grouped the data set as far as regarding growth size and HB level and the outcomes are affirmed in the R climate. The nonstop trait order technique to demonstrate first degree of TNM in cellular breakdown in the lungs patient alongside standard hemoglobin must be kept up with that individual's survivability rate is higher than the more modest degree of hemoglobin individual's endurance rate. The Gaussian K-Base NB classifier is more compelling than the current AI calculations for cellular breakdown in the lungs forecast model. The proposed order exactness has estimated utilizing ROC strategies.

[12]Ryu, Sung Mo, et al proposed Predicting survival of patients with spinal ependymoma using machine learning algorithms with the SEER database**.** The purpose of this study was to learn about the clinical and demographic factors that influence the overall survival (OS) of patients with spinal ependymoma and to predict the OS using machine learning (ML) algorithms. The Surveillance, Epidemiology, and End Results (SEER) registry was used to compile cases of spinal ependymoma diagnosed between 1973 and 2014. Statistical analyses were performed using the Kaplan-Meier method and the Cox proportional hazards regression model to identify the factors influencing survival. In addition, we used machine learning algorithms to predict the survival of patients with spinal ependymoma. Age 65 years, histologic subtype, extraneural metastasis, multiple lesions, surgery, radiation therapy, and gross total resection (GTR) were found to be independent predictors of OS in the multivariate analysis model. Our ML model predicted a 5-year OS of spinal ependymoma with an area under the receiver operating characteristic curve (AUC) of 0.74 (95 percent confidence interval [CI], 0.72-0.75) and a 10-year OS with an AUC of 0.81 (95 percent CI, 0.80-0.83). The stepwise logistic regression model performed worse, with an AUC of 0.71 (95 percent confidence interval, 0.70-0.72) for predicting a 5-year OS and an AUC of 0.75 (95 percent confidence interval, 0.73-0.77) for predicting a 10-year OS.SEER data confirmed that therapeutic factors such as surgery and GTR were associated with improved overall survival. ML techniques outperformed statistical methods in predicting OS; however, the dataset was heterogeneous and complex, with numerous missing values.

[13]David Riao, Ricardo, and Kleinlein suggested persistence of data-driven knowledge to forecast survival from breast cancer. By adapting machine learning prediction models to the stage of the cancer at the time of diagnosis, breast cancer survival rates can be increased. However, the predictive capability of these models as well as the importance of the clinical characteristics in that prediction may alter with time. figured out if the results about the performance of machine learning models and the effect of clinical factors in the prediction of breast cancer survival are temporary or permanent, and if temporary, how long the newly acquired information will be valid if it is.

On the application of machine learning techniques to predict breast cancer survival, there have been fifteen recent publications with pertinent conclusions. Several data-driven models were subsequently developed throughout time to estimate the five-year survival of breast cancer using the breast cancer data in the SEER database. Three different machine learning techniques were used. Step-specific models and joint models were taken into consideration for each stage. The predictive capability of the models and the significance of clinical indicators were submitted to a persistence study over time in order to establish the validity and long-term viability of these fifteen results. Only 53% of the judgments in the SEER cases from 1988 to 2009 were accurate, and only 75% of these across time.Relevant conclusions, such as the inability to increase survival prediction accuracy for the most frequent stages with more data or the significance of cancer grade in predicting breast cancer survival for patients with distant metastasis, were found to be false when subjected to a temporal analysis. Our study has found that before being used in clinical and professional settings, data-driven knowledge generated through machine learning techniques has to be evaluated over time.

A model developed by Narges Habibi, Majid, and Naghizadeh employs an ensemble learning method to predict the prognosis of cancer comorbidity. Cancer is one of the leading causes of death worldwide. Breast and vaginal cancer in women, as well as prostate cancer in men, are some of the most common malignancies. The early detection and prevention of these cancers are crucial goals. Conditions have a worse chance of survival than those with just one type of cancer. The significance of concurrent chronic illnesses during cancer therapy is assessed using a range of machine-learning approaches using SEER data. Use the gradient boosting ensemble technique for feature selection. According to recent investigations, some people have concurrent cancer. The accuracy of estimating cancer patient survival rates in patients with related illnesses is improved by modeling. This technique shows a significant improvement in prediction accuracy when compared to prior proposed models and suggests an increase in the survival rates for comorbid cancer. The forecasting of the survival rate in patients with cancer comorbidity is recommended using an ensemble-based technique. The initial stage in the strategy to locate the targeted comorbid patients was combining the necessary SEER data sets. The important input features are determined using ensemble methodologies after each record is classified as either living or dead, preprocessing (such as handling missing values), and balancing the resulting data set.. Several prediction methods are tested using a traintest split, and Gradient Boosting is finally chosen as the best predictor because to its improved performance. According to the findings of the studies done here, the suggested model performs better than the other approaches in terms of precision, error, sensitivity, and specificity when it comes to predicting survival in cancer comorbidity..

[14]J. A. Bartholomai Recently, results for patients with malignant growth have been assessed using a variety of AI techniques on significant datasets like the Surveillance, Epidemiology, and End Results (SEER) programme data set. et'al Supervised machine learning classification methods for predicting lung cancer patient survival. Particularly for cellular breakdown in the lungs, it is uncertain which procedures would produce more accurate data and which information credits should be employed to establish this data. This study uses a number of directed learning approaches, including as straight relapse, Decision Trees, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and a custom ensemble, to group patients with cellular breakdown in the lungs according to endurance. By using these strategies, credit for essential information will be given. The expectation is viewed as a nonstop objective rather than a classification as a first step in improving endurance forecast. The results show that the anticipated features match the actual qualities during low to direct endurance durations, which make up the majority of the data. The custom troupe performed the best, with an RMSE of 15.05 for Root Mean Square Error. In the custom group, GBM was the most effective model, with Decision Trees perhaps being useless since they provided too few discrete outputs. With an RMSE value of 15.32, the findings also show that GBM was the most reliable model among the five created separately. The SVM failed to match predictions despite having an RMSE of 15.82, The outcomes of the models are foreseeable when a conventional Cox relevant hazards model is utilized as a viewpoint approach. We believe that measuring patient endurance time with the explicit goal of illuminating patient consideration choices could be aided by applying these administered learning strategies to the SEER data set's information on cellular breakdown in the lungs, and that the demonstration of these procedures with this specific dataset may be comparable to that of conventional methods.

## 3. PROBLEM STATEMENT

A well-known research topic has been the anticipation of malignant development durability. The majority of illness survivability research focuses on trying to forecast patients' five-year survival rates. These tests provide a constrained amount of information for making clinical decisions. The patient's remaining components' endurance season is unknown if the patient's prognosis is "death." To provide more precise information for clinical decision-making, endurance time expectation should be investigated. With this project, the endurance time will be predicted on a month-to-month basis. The suggested forecasting model has two stages[15].

**Objective**

Comorbidity focused on illnesses that already coexisted. Examining actual disease cases reveals that some diseases have stronger correlations than others. A well-known scientific area has been the expectation of disease endurance. Accurately predicting a patient's chance of survival might help professionals with therapeutic advice and pharmaceutical recommendations. The likelihood that a patient will survive a long period after the diagnosis of their illness is known as survivability. It is a clinical marker for evaluating the effects of treatment. The majority of illness survivability research focuses on strategies to predict patients' five-year survivorship. These tests provide a constrained amount of information for making clinical decisions. To provide more precise information for clinical decision-making, endurance time projection should be taken into account.[16]

The focus of comorbidity was on diseases that previously coexisted. Some diseases have higher associations than others, as shown by an examination of real sickness cases. Expected illness endurance has been a well-known scientific field. The ability to accurately anticipate a patient's likelihood of survival might aid specialists in making therapeutic suggestions and medication recommendations. The probability that a patient will live a significant amount of time following the diagnosis of their condition is referred to as survivability. It serves as a clinical indicator for assessing the outcomes of treatment. The majority of research on sickness survivability concentrates on methods to forecast patients' five-year survival rates. Making clinical judgments using the information from these tests is limited. The forecast of endurance time should be taken into consideration to offer more exact information for clinical decision-making.[17]

The purpose of this article is to predict the endurance time on a month-to-month basis. However, the anticipation of endurance time has been shown to be challenging because when one-stage relapse models are used, significant speculative errors typically occur. A two-stage expectation approach is suggested to solve this problem. At the first step, classification, a classifier is used to determine if the patients would be able to survive for more than five years. A relapse model is used to predict the endurance season of patients who have been identified as having no choice to survive for a long period at the next stage, which is regression..

Poor classification performance is the problem that develops during the classification step. The issue of bias is shown using a survival time histogram in the section that follows, and the classification performance using SVM and Naive Bayes is determined. It is suggested that CHI2 feature selection be used in cascade with the support vector machine and nave bayes classifiers to improve classification performance. For two-stage classifiers, the feature selection approach CHI2 is used. The public is welcome to use this feature selection process. The aforementioned enhancements cannot be utilized at the regression step since the projected outcome is continuous.

However, the error rate is large and training takes a long period without data pretreatment. The suggested two-stage framework outperforms the one-stage strategy in both classification and regression tasks. The original linear support vector machine (Linear-SVM) and logistic regression have higher prediction accuracy than the naïve bayes classifier in the classification stage. In the second stage, the RMSE of the enhanced random forests (RF) approach is lower than the RMSE of the first-generation RF method and other feature selection techniques.[18]

The main goal of this study was to investigate the endurance issue from a different angle.Instead of the enduring rate on a time point of an associate following the finding in the conventional endurance examination, we tried to address the question of how long a specific patient would survive after the conclusion. It was demonstrated through a sequence of data from standard trials that the survival could be achieved using normal machine learning techniques.[19]

## 4. PROPOSED WORK:

The training and testing datasets each had 10985 instances. When the characteristics of the numerous primary malignancies were pooled, several characteristics were the same. After removing duplicate features from the merged feature pool, features were chosen and translated using Label Encoding, consisting entirely of zeros and ones. In the classification step, CHI2 feature selection decreased data dimensionality, while splitting the dataset decreased the number of training cases. The linear SVM classifier and the Nave Bayes classifier were employed as classifiers. The classification stage employs the CHI2 feature selection approach. During the regression step, patients who lived for more than 60 months were excluded from the total dataset.. The random forest Regressor was employed because of how well suited to the regression process it is by nature. The element-wise feature dropping RMSE scores are also compared using these techniques. The top 10 characteristics are kept. Their RMSE ratings drop as additional characteristics are taken out of the pool. Every iteration, the training set instructs the classifier, and the accuracy score of the testing set is recorded for comparison.

### 4.1 Data Preprocessing

Two types of preprocessing are used to balance and clean the data:

**1) Data balancing:**

The class imbalance problem, which is common in supervised learning methods, is characterized by a large discrepancy in sample counts between classes. Because learning algorithms are typically biased towards large classes and perform badly on smaller classes, unbalanced data sets are a problem. As a result, stratified sampling is employed in this work to balance samples prior to modeling. Making the necessary modifications and comprehending the distribution of your training data across the classes you wish to forecast are essential elements in creating a high-quality classification model. When trying to anticipate something infrequent, such infrequent fraudulent transactions or odd equipment breakdowns, imbalanced datasets are highly prone to happen. The distribution of the target classes should always be taken into account, regardless of the domain.[20-21]

**2) Data cleaning:**

There must be proper handling of missing values because the SEER data set includes certain fields with blank values. These fields can make it more difficult to create models during the learning phase and decrease prediction accuracy and processing speed. Features having more than 50% nonexistent values are not included in this scenario. The median values of the characteristics with fewer than 50% missing data are changed. Only a portion of the SEER variables and the variables that were excluded from the models are included, along with descriptions of those variables, due to the length of the entire list.

Data cleaning is the process of preparing data for analysis by removing or altering data that is inaccurate, lacking, irrelevant, duplicated, or formatted incorrectly. This information is typically not needed or useful when it comes to data analysis because it might slow down the procedure or lead to erroneous findings. There are several techniques for cleaning data, depending on how it is kept and the questions asked. Data cleaning involves finding ways to optimize a data set's correctness without necessarily losing information. It goes beyond just eliminating data to create place for new data. In addition to deleting data, data cleaning also involves addressing spelling and grammar problems, standardizing data sets, and resolving errors including empty fields, missing codes, and other types of errors and locating data points that are duplicates. Because it is essential to the analytical process and the identification of trustworthy solutions, data cleaning is regarded as a fundamental component of data science fundamentals.[22-23]

### 4.2 Approach- Two Stage Prediction

Biased datasets and subpar classification performance are two problems that come up during the classification step. The bias issue is shown using a survival time histogram as an illustration. It is calculated how well the support vector machine and naïve bayes classifier do at classifying data. Improved CHI2 feature selection is suggested to cascade with the Support Vector Machine, Logistic Regression, and Naive Bayes classifiers in order to overcome poor classification performance. For two-stage classifiers, the CHI2 feature selection approach is applied. The public is welcome to use these feature selecting techniques.

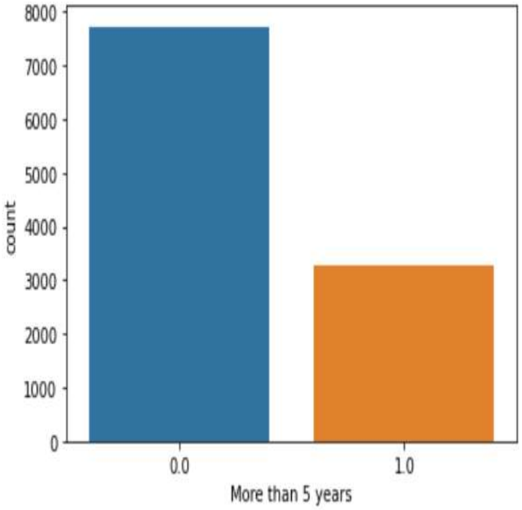

Fig 4.1: Biased data before over-sampling

The aforementioned improvements cannot be utilized during the regression stage because the predicted outcome is continuous. However, the error rate is high and training takes a long time without data preprocessing. Regression is carried out using a random forest Regressor.[24-25]

1. Assume there are two stages to the survival prediction issue.
2. Build cancer comorbid datasets using the SEER database.
3. Use CHI2 feature selection during the classification phase.
4. Apply SVM to the classification process.
5. Employ the random forest Regressor during the regression phase.
6. Compare and contrast the two-stage classification and regression model with the one-stage regression model.

The suggested two-stage framework outperforms the one-stage strategy in both classification and regression tasks. In the classification stage, the Naive Bayes classifier's prediction accuracy is inferior to that of the original Linear-SVM and Logistic Regression. In the second stage, the RMSE of the enhanced random forests (RF) approach is lower than the RMSE of the first-generation RF method and other feature selection techniques.[26-27]

### 4.3 Methodology

The majority of malignant growth projection studies are limited to determining whether a patient will live for a specific amount of time. Then, the patient is designated as "made due" or "dead." Most cases of liver malignant development would be considered "dead" because of the high fatality incidence. These patients' endurance duration is yet unknown. Then, we provide a two-stage order model that consists of a characterization model that forecasts the patient's likelihood of survival and a relapse model that forecasts the excess life expectancy of patients whose projected result is "dead." With the exception of the fundamental AI kinds, the two phases use similar methodologies. In the grouping step, straight SVM classifiers, Naive Bayes classifiers, and RF classifiers are employed to predict the endurance condition. Regressors are used to predict endurance months during the relapse period. Two problems are encountered throughout the ordering process. The main problem is that a one-sided classifier would result from a one-sided preparation set. Cases from the minority class would be incorrectly categorized as belonging to the larger group. Information adjustment is necessary to address this problem. The next problem is that the element pool is quite large and the characterization outcome is subpar. The fountain is subjected to CHI2 Feature Selection using a support vector machine classifier and a Nave Bayes classifier selecting a selection of pool highlights. The first classifier was not preferred by the flowing framework during grouping execution.[28]

The steps of the categorization framework are as follows:

1. consulting the SEER database for statistics on MPCs such liver, lung, stomach, thyroid, and skin malignancies.

2. Combine the data and change the order of the data.

3. Divide the data into training and testing sets.

4. To balance the dataset, employ SMOTE (Synthetic Minority Oversampling Technique).

5. Select the top characteristics for modelling using CHI2 Feature Selection.

6. Use the linear-SVM, Naive Bayes, and Logistic Regression classifiers for prediction.

7. Evaluate the outcomes that were predicted using error metrics like accuracy and f-score. The steps in the regression framework are as follows:

1. 1. Remove instances with a survival month of more than 60 from the categorization data.
2. 2. Separate the data into training and testing sets.
3. 3. Apply the RF Regressor to the forecast.
4. 4. To assess the accuracy of the predictions, consider the root mean squared error (RMSE), mean absolute error (MAE), and R2 score.

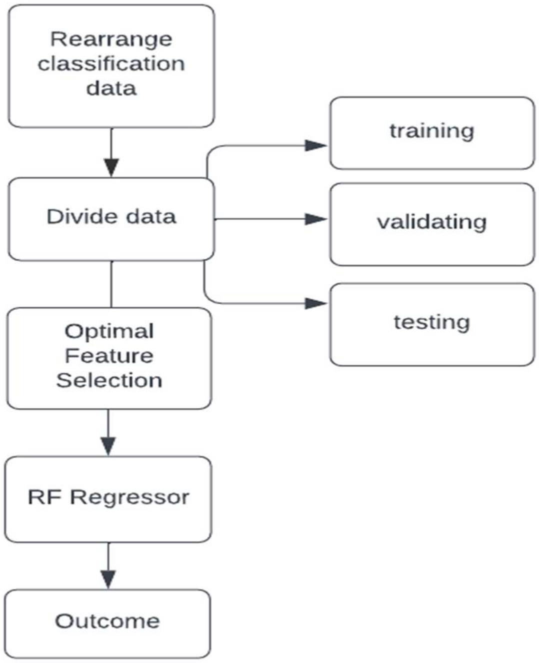

Fig 4.2 Methodology

Linear SVM and Naive Bayes were the classifiers employed in the classification. Prior to categorizing samples based on distance, it builds a dividing hyper plane between two classes. With just zeros and ones, the data has been one-hot encoded.

Linear SVM and Naive Bayes were able to separate one-hot encoded data as needed. It employed a chain of CHI2 feature selection and random under sampling. In the regression phase, the RF repressor was applied.

As a typical bagging Regressor, it was cascaded with feature selection based on contribution score.

### 4.4 Materials and Methods

Data about malignancies in the United States are deidentified, categorized, and annotated in the free and open-source database SEER. The database is big enough to provide machine learning algorithms a lot of examples to study. The clinical or microscopic confirmation of a cancer diagnosis in the SEER cancer registries was performed by a licensed medical professional.

The majority of cancer prognosis studies merely estimate how long a patient will live. The patient is then classified as having "survived" or "passed away." The majority of liver cancer patients would be considered "dead" because of the disease's high fatality rate. The lifespans of these patients remain unknown. We thus suggest a two-stage categorization methodology. It contains a classification model that forecasts the patient's likelihood of survival and a regression model that forecasts the life expectancy of patients whose forecasted result is "dead."

Both phases adhere to identical techniques, with the exception of the fundamental machine learning types. The survival condition is predicted using linear-SVM, Naive Bayes, and logistic regression classifiers in the classification stage, and the survival months are predicted using RF regressor and Decision Tree Regressor in the regression stage. During the categorization phase, two problems emerge. The first problem is that the biassed training set would produce a biassed classifier. Minority-group cases would be incorrectly categorised as belonging to the dominant group. To address this issue, data balancing is required. The second problem is the size of the feature pool, which leads to a subpar classification outcome.

A support vector machine classifier and a Nave Bayes classifier are used in a cascade with CHI2 Feature Selection to select a subset of features from the pool. In terms of classification performance, the cascaded system outperforms the original classifier.[29]

### 4.5 Algorithms Applied

Naive Bayes and linear SVM were used as the classifiers in the classification. It builds a separation hyper plane between two classes before categorizing samples according to their distance. One-hot encoding was used to create the data, which solely contains zeros and ones.

The need for segregating one-hot encoded data was met using linear SVM and Naive Bayes classifier. With the choice of CHI2 features, it cascaded. RF served as the Regressor in the regression step. It was cascaded with contribution score-based feature selection as a typical bagging Regressor.

**Linear SVM:**

When a dataset can be divided into two classes by a single straight line, it is said to be linearly separable, and the Linear SVM classifier is used to separate the dataset into its two groups. Depending on the dataset, we employ different machine learning techniques to forecast and categorized data. A linear model called the SVM, or Support Vector Machine, can be utilized to address classification and regression issues. It has several practical uses and may be applied to both linear and nonlinear situations. The basic idea behind SVM is simple: To categorized the data, the algorithm creates a line or a hyper plane. SVMs initially identify a line (or hyper plane) that divides the data from two classes. The SVM algorithm takes data as input and produces, if it is possible, a line that divides those classes.[30]

**Naïve Bayes:**

A collection of classification methods founded on Bayes' Theorem are referred to as "Naive Bayes classifiers." It is a group of algorithms that are all based on the idea that every pair of characteristics that is used to classify something is independent of the other. In applications such as sentiment analysis, spam filtering, recommendation systems, and others, naive Bayes algorithms are frequently employed. Although they are quick and easy to implement, their primary drawback is the requirement for independent predictors. The predictors are often dependent in real-world scenarios, which hinders the effectiveness of the classifier. The Naive Bayes algorithm is a supervised learning method that addresses classification issues by applying the Bayes theorem. With a sizable training dataset, it is primarily used for text classification. The Naive Bayes Classifier is a rapid and efficient classification technique that helps create machine learning models that can learn quickly and anticipate outcomes. Being a probabilistic classifier, it makes predictions based on the likelihood of an item. Popular Naive Bayes Algorithm uses include spam filtration, sentiment analysis, and article categorization.

**Random Forest:**

A well-known machine learning method from the supervised learning approach is Random Forest. It may be used to solve machine learning challenges including classification and regression. It is based on the idea of ensemble learning, which is a method that combines several classifiers to solve a challenging problem and enhance the performance of the model. A classifier called Random Forest uses a number of decision trees on different subsets of the provided dataset. It takes the average to increase the dataset's forecast accuracy," as the name suggests. The random forest uses the forecasts from each decision tree to anticipate the ultimate result based on the majority vote of predictions rather than depending just on one decision tree. The accuracy is higher and the risk of over fitting is lower the more trees there are in the forest. One of Decision Trees' biggest drawbacks, variation, is addressed with the Machine Learning method Random Forests.

Decision Trees are a greedy algorithm in spite of their versatility and simplicity. Instead of concentrating on how that split impacts the entire tree, it concentrates on optimizing for the present node split. A rapacious strategy expedites Decision however makes them vulnerable to over fitting. A high-variance learning model is produced as a result of an over fit tree being highly optimized for forecasting the values in the training dataset.[31]

**Logistic regression:**

We employ the logistic regression statistical modeling method when the result is binary. When the outcome variable is binary, logistic regression modeling may be used to predict the outcome whether the independent variables are continuous or categorical. Logistic regression is the method of estimating the likelihood of a discrete outcome from an input variable. The majority of logistic regression models feature a binary result that can be true or false, yes or no, or another value. Modeling situations with more than two discrete outcomes may be done using multinomial logistic regression. A helpful analysis technique for classification issues is logistic regression, which may be used to determine if a new sample belongs in a particular category. because of factors Logistic regression is a helpful analytical method for classification issues in cyber security, such attack detection. Logistic regression is an easier and more effective solution for issues involving binary and linear classification. It is a classification model with linearly separable classes that is straightforward to use and produces outstanding results. It is a classification method that is frequently used in business. The logistic regression model is a statistical technique for binary classification that can be extended to multiclass classification, just like the Adaline and Perceptron. Multiclass classification tasks can be handled by the highly optimized logistic regression implementation in Scikit-learn.

**Decision tree:**

The family of supervised machine learning algorithms includes the decision tree method. Both classification and regression issues may be solved with it. The objective of this approach is to build a model that predicts the value of a target variable. To do this, a decision tree is used, which represents the issue as a tree with characteristics represented on the core node of the tree and a leaf node that corresponds to a class label. The family of supervised learning algorithms includes the decision tree algorithm. In contrast to other supervised learning algorithms, the decision tree technique may also be utilized to address classification and regression issues. Building a training model is the purpose of employing a decision tree. That can learn straightforward decision rules from historical data and anticipate a target variable's class or value (training data). In decision trees, we start at the tree's base to forecast a record's class label. We contrast the root attribute and the record attribute's values. We follow the branch leading to that value's value based on the comparison and go on to the next node. The correctness of a tree is strongly influenced by the choice of strategic splits. Regression and classification trees have different decision criteria. Decision trees use a variety of algorithms to determine whether to divide a node into two or more sub-nodes. The homogeneity is increased by the creation of sub-nodes. of the resulting sub-nodes. In reference to the target variable, the node's purity rises, in other words. The decision tree divides the nodes according to all of the variables that are available, then it selects the split that results in the most homogeneous sub-nodes.[32]

**Oversampling and under sampling**:

A considerable skew in the class distribution can be seen in imbalanced datasets, such as 1:100 or 1:1000 samples in the minority class relative to the majority class. Many machine learning algorithms may be affected by this bias in the training dataset, and others may totally ignore the minority class. Minority forecasts are sometimes the most crucial, thus this is a concern. Randomly resampling the training dataset is one way to address class imbalance. Under sampling, or removing examples from the majority class, and oversampling, or duplicating examples from the minority class, are the two main techniques for randomly resampling an unbalanced dataset.

The two primary methods of random resampling are oversampling and under sampling for categorization that is unfair.

Duplicate samples in the minority class selected at random using oversampling.

Random Under sampling, randomly remove instances from the majority class.

The technique of randomly choosing instances from the minority class and substituting them in the training dataset is known as random oversampling. The act of randomly picking instances from the majority class and eliminating them from the training dataset is known as random under sampling. Both methods can be used repeatedly until the training dataset achieves the desired class distribution, such as an equal split across the classes.

They are referred to as "naive resampling" techniques since they employ neither heuristics nor assumptions about the data. Because of this, they are simple to use and quick to carry out It is perfect for really big and complicated datasets. Both methods may be used to classify issues with two classes (binary) or with many classes that include one or more majority or minority classes. Importantly, the training dataset is the sole one to which the class distribution modification is performed. The intention is to alter the models' fit. It is not necessary to resample the test or holdout datasets used to assess a model's performance. These simplistic techniques may work in general, but it also depends on the particulars of the dataset and models being used. The practice of random oversampling involves adding duplicates of minority class samples to the training dataset. Machine learning algorithms that are impacted by skewed distributions and when several variables are present may benefit from this strategy For a given class, duplicate examples might affect model fit. This may involve iteratively learning coefficients-based techniques like stochastic gradient descent-based artificial neural networks. Support vector machines and decision trees are two examples of models that might be affected.[32]

### . 4.6 Chi Square Feature Selection

The process of removing the most pertinent features from a dataset and then using machine learning algorithms to boost the performance of the model is known as feature selection, also known as attribute selection. Over fitting is more likely and training time is exponentially increased by a large number of irrelevant features.

**Chi-Square Feature Extraction:**

To extract categorical characteristics from a dataset, utilize the Chi-square test. The Chi-square test is performed between each feature and the target, and the features with the highest Chi-square scores are chosen. It determines whether the relationship between two categorical variables in the sample accurately reflects their relationship in the population.

$$\chi^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

A well-liked technique for choosing features from text data is the Chi-Square feature selection method. The 2 test in statistics is used to establish the independence of two events. Determine whether the occurrence of a specific term and the occurrence of a specific class are independent in feature selection.

Two distributions are compared using the Chi-Square test to see how comparable their relative variances are. Its null hypothesis is based on the supposition that the provided distributions are independent. Thus, by identifying which features are most reliant on the output class label, this test may be used to identify the optimal features for a certain dataset. Each feature in the dataset has its chi2 value determined, and the features are then sorted in decreasing order using the chi2 value. The more dependent the output label is on the feature and the more crucial the feature is in determining the output, the higher the chi 2 values.

The Chi-Square test's application in machine learning and its effects are extensively questioned. Because we will have several features in line and must choose the best ones to create the model, feature selection is a crucial issue in machine learning. By analyzing the relationship between the characteristics, the chi-square test assists in feature selection.

The chi-square test in statistics is used to examine if two occurrences are independent of one another. From the data of two variables, we can obtain the observed count O and the expected count E. The difference between the observed count O and the expected count E is calculated using the Chi-Square formula.

The observed count is close to the expected value when two features are independent therefore, the Chi-Square value is lower. expected count. A large value for the Chi-Square statistic suggests that the independence hypothesis is untrue. Simply put, the more dependent a feature is on the response, the higher the Chi-Square value, and the more suitable it is for model training.[32]

**Limitations:**

In table cells, Chi-Square is sensitive to low frequencies. In general, chi-square can produce false results when the expected value in a table cell is less than 5.

## 5. RESULTS

SMOTE (Synthetic Minority Oversampling Technique) oversampling and CHI2 feature selection make up the enhancement of the classification stage. The classification performance metrics for SVC, Gaussian Nave Bayes, and Logistic regression are listed in the table below. F1 score, Accuracy, and Confusion matrix are the performance metrics used for comparison. The R2 score, RMSE, and MAE are the performance indicators used in regression.

### 5.1 Classification Stage:

To transform text input into numerical data, we had employed label encoding. The issue of the class gap has been covered in the sections above. The "less than 5 years of survival" class of cases dominates the other class of cases, as can be shown in Fig. 1. One of the most popular approaches to address the issue of class imbalance is over-sampling. Given the small size of the dataset, we had taken into account the SMOTE oversampling method in this case. The dataset's size increased from 10985 to 15439 cases after SMOTE was applied.

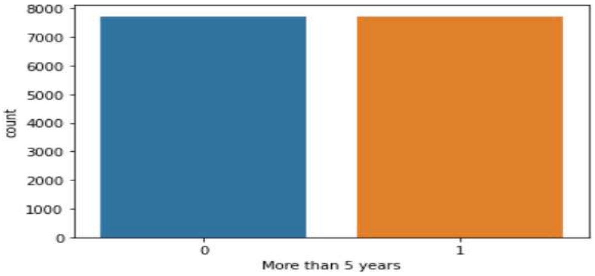

*Figure 5.1 Classification Stage*

Results after applying SMOTE: 0-cases with less than 5 years of survival, 1-cases with more than 5 years of survival.

Using CHI2-based feature selection, we had chosen the six top features out of a total of 16 characteristics. These six features were chosen by the CHI2-feature selection as the top ones:

1. Replace the age with one (1).
2. AJCC T, 6th edition derived (2004-2015).
3. AJCC N, 6th edition derived (2004-2015).
4. AJCC M, 6th ed. derived (2004-2015).
5. Labelled Primary Site.
6. AJCC Stage Group, 6th edition derived

The data were divided one to three. There were 11,759 test cases and 3860 train cases in the data, respectively. For patients whose expected survival time is less than 60 months and for patients whose expected survival time requires more than five years, the classifiers assigned labels of 0 and 1, respectively. With an F1 score of 0.788, the SVC has the highest among the three models and is more accurate than the other two.

The results of the Confusion matrix are listed below the table along with the accuracy and F1 score of the three models.

*Table1: Accuracy And F1 Score Of The Three Classification Models.*

| MODEL | ACCURACY | F1 SCORE |
|---|---|---|
| Gaussian Naïve Bayes | 74.63 | 0.769 |
| Logistic Regression | 77.74 | 0.763 |
| Support Vector Classifier | 78.54 | 0.788 |

*Table2: Results From The Confusion Matrices Of The Three Classification Models.*

| MODEL | PREDICTED 0 | PREDICTED 1 | ACTUAL |
|---|---|---|---|
| Gaussian Naïve Bayes | 1251 | 728 | 0 |
| | 251 | 1630 | 1 |
| Logistic Regression | 1613 | 366 | 0 |
| | 493 | 1388 | 1 |
| Support Vector Classifier | 1488 | 491 | 0 |
| | 337 | 1544 | 1 |

### 5.2 Regression Stage:

The classification stage's output was filtered to only include instances with anticipated labels of 0. (less than five years of survival time). Decision tree and random forest regression models are used. R2, RMSE, and MAE are the comparison metrics for these two models. The random forest regressor has the highest R2, the lowest RMSE, and MAE of the two.

| MODEL | R2 SCORE | RMSE | MAE |
|---|---|---|---|
| Random Forest Regressor | 0.42 | 32.03 | 21.60 |
| Decision Tree Regressor. | 0.41 | 32.29 | 21.69 |

## 6. CONCLUSION & FUTURE SCOPE:

The bulk of current survival analyses concentrate on the relationships between the characteristics and patients' chances of surviving five years. The specific question of how long a patient with concomitant cancer would live is still mostly unanswered. In this experiment, the patient-specific survival time of cancer patients with concomitant conditions was predicted. The customized query is split into two machine learning issues. The distinction between patients who will live longer than five years and those who won't is the first problem. The second step is to develop a regression model that forecasts the patient's five-year survival rate.

Cancers of the lung, liver, stomach, thyroid, and skin are among the most prevalent. It can be beneficial for doctors, patients, and families to predict the prognosis of cancer patients. The suggested two-stage approach not only predicts survival but also the number of months a patient will live. The first stage foretells whether or not a patient will survive for more than five years. The second stage estimates the patient's remaining months of life if the prediction is death. Scaling of features is used in the classification stage during feature selection. Use of the Random Forest Classifier is made during the regression phase.

Applying Feature Selection during the regression stage can further increase accuracy. Investigating multidisciplinary and intradisciplinary dispersions can help the feature selection process become even better. We will keep looking at feature selection techniques that might boost present prediction performance in the future. Studying second primary breast cancers are another MPC that may be investigated.

## 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, 1973–2000, no. 5. 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. Cho, R. N. Anderson, B. A. Kohler, C. R. Eheman, 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. Zolbanin, D. Delen, and A. Hassan Zadeh, ''Predicting overall survivability in comorbidity of cancers: A data mining approach,'' Decis. Support Syst., vol. 74, pp. 150–161, Jun. 2015.

[6] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, ''A tree ensemblebased 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. Balgemann, V. H. van Berkel, and H. B. Frieboes, ''Prediction of lung cancer patient survival via supervised machine learning classification techniques,'' Int. J. Med. Informat., 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. Artif. Intell. Appl., 2020, pp. 105–116, doi: 10.5121/csit.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. Kvinnsland, ''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-selectio%n-library

[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. Syst. Sci., Eng. Design Manuf. Informatization, 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. Kaviarasi, ``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 ependymoma 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. Informat., vol. 129, 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. Drakaki, and K. Chamie, ``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.

[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.