Improving Hepatocellular Carcinoma Survival Prediction with Artificial Intelligence Strategies

Lingfei Zhao

Abstract

Hepatocellular Carcinoma (HCC), which constitutes over 90 percent of liver cancers, is the sixth most frequently diagnosed cancer and the third-leading cause of cancer related deaths in the world. Its incidence rate has tripled since 1980 and is estimated to further increase 61.9% within two decades. Thus, the increasing prevalence of HCC emphasizes the need for more accurate methods of survival prediction in order to increase survival rate.

Two proposed methods of survival prediction for Hepatocellular Carcinoma patients were compared, both utilizing the missForest algorithm to impute missing values in the dataset. The first method incorporated random forests to classify the data into two categories: Positive (the patient will not survive past one year) and Negative (the patient will survive past one year), while the second method used a combination of principal component analysis and support vector machine (PCA-SVM) to make these predictions. Although the Random Forest model had a higher mean accuracy of 0.727 compared with the PCA-SVM model's mean accuracy of 0.712, the latter had a significantly higher mean True Positive Rate (0.630 compared with the 0.547 of the Random Forest model, $p < 0.05$). In other words, the PCA-SVM model was superior in correctly predicting that a patient was in critical condition and would not survive past one year. This model allows for more effective treatment to target the patient's individual needs, thus, the PCA-SVM was recommended as a more reliable method for prognosis evaluation.

1 Introduction

Currently, Hepatocellular Carcinoma (HCC) is the sixth most frequently diagnosed cancer and the third-leading cause of cancer-related deaths in the world (American Cancer Society, 2019). Constituting over 90% of liver cancers, HCC occurs predominantly in patients suffering from liver cirrhosis, with the most common risk factors including chronic viral hepatitis, alcohol, and nonalcoholic fatty liver disease. Each year, not only are over 800,000 people throughout the world diagnosed with the disease, liver cancer also accounts for over 700,000 deaths (American Cancer Society, 2019). Furthermore, the liver cancer incidence rate has tripled since 1980, and is further estimated to increase 61.9% by 2040, from 841,080 cases this past year to 1,361,836 in 2040 (Rawla et al., 2018). Thus, the increasing prevalence of HCC emphasizes the need for accurate survival prediction, which would assist in the application of effective treatments. Such a task includes analysis of a substantial amount of data, drawing patterns from the data, and using these conclusions to predict the survivability of patients, all generally accomplished through machine learning techniques (Santos et al., 2015).

While survival prediction is typically performed using computational methods such as machine learning, most existing models have limitations, particularly due to the size and complexity of the datasets. Firstly, in many cases, only small datasets are available, which provide insufficient information for some algorithms and thus limit data mining techniques. Regarding the second issue of data complexity, as many datasets include patient heterogeneity and/or missing data (i.e., incomplete variables or variables with missing values) that the algorithm fails to take into account, biased models may be produced (Santos et al., 2015).

Santos et al. (2015) proposed a new, more robust HCC survival prediction model using the technique of cluster analysis, which increases the accuracy of predictions by generating homogenous groups. Four different approaches were taken – the last two use the proposed methodology of cluster-based oversampling – keeping in consideration both the presence of patient heterogeneity as well as missing data: direct use of obtained dataset, oversampling the obtained dataset with the Synthetic Minority Over-sampling Technique (SMOTE) algorithm, generating datasets using the proposed cluster-based oversampling method and combining them in a representative dataset, and lastly, combining each previously generated dataset from the third approach with the representative dataset, forming an augmented dataset.

The majority of previous studies done on HCC used Neural Networks (NN), a set of algorithms

modeled loosely after the human brain that are trained to recognize patterns, and Logistic Regression (LR) models, which are used to describe data and the relationship between variables. Thus, Santos et al. (2015) tested all four approaches discussed in the previous paragraph with the same two algorithms. Through the process of data imputation, a value was substituted for each missing piece of data using a nearest neighbor approach, which selected from the available data the closest neighbor to each incomplete case. Next, the data was divided into naturally occurring clusters with similar features through the K-means algorithm, an unsupervised learning algorithm that finds a fixed number k of clusters in a set of data (Piech, 2013). In essence, the algorithm finds k centroids and assigns all data points to the nearest centroid thus forming a cluster, then updates each centroid by computing the mean of all data points within each cluster. This process repeats until a stopping criterion is met, i.e., when the maximum number of iterations is reached, or the points no longer change clusters. In order to equally represent each cluster and thus balance the database, Santos et al. (2015) incorporated the SMOTE algorithm, a method of oversampling that avoids the usual problem of overfitting. The SMOTE algorithm generates synthetic samples for smaller clusters based on their nearest neighbors, balancing all groups to the second largest cluster. Finally, for the second sampling phase, an augmented dataset was constructed by merging the oversampled datasets to better represent the problem.

It was found that when combined with the NN classifier, the two proposed approaches based on cluster oversampling, particularly the augmented sets approach, proved much more successful than the original two widely-used methods. Thus, in the context of HCC survival prediction, cluster-based oversampling with augmented datasets is a much more appropriate approach. Two directions of research arose from this: firstly, applications to other classification problems, whether they be medical or nonmedical, and secondly, adapting the algorithm for missing data imputation (Santos et al., 2015).

In contrast, a different HCC prognostic evaluation model dealt with the problem of tumor heterogeneity by focusing on molecular classification (Ke et al., 2018). Using data from 371 HCC patients, factors including gender, age, and TNM (tumor, node, metastasis) stage were observed to determine which genes significantly affected the overall survival rate of the patients. A clustering method was applied on these such genes in order to determine the molecular subtypes, which were hypothesized to provide more accurate and significant predictions. Both conditional logistic regression and binary logistic regression were used to generate the prognostic model, which was a linear combination of the subtype genes expression values after normalization. Ke et al. (2018)

confirmed through a series of tests the presence of two molecular subtypes in HCC that were associated with prognosis: molecular subtype 1 and molecular subtype 2, the first of which had a significantly longer survival time. Taking this into consideration, logistic regression analysis was performed with the inclusion of six genes in the final model, which ultimately proved effective at distinguishing between the two subtypes and further predicting the patients survival time, independent of the clinical conditions. The final prognostic model was determined to be

$$prognostic\ value = \frac{e^p}{1 + e^p} \tag{1}$$

where p is the sum of the normalized expression values of the six selected genes, and e represents Euler's number. Thus, it was concluded that the evaluation model was suitable for application in a clinical setting, replacing previous, widely-used methods. However, further research is still necessary to confirm its clinical significance (Ke et al., 2018).

Regarding the problem of missing value imputation, which is particularly important in data analysis, the most popular methods such as k nearest neighbor imputation (KNNimpute), multivariate imputation by chained equations (MICE), and the Missingness Pattern Alternating Lasso (MissPALasso) require a parametric model and make assumptions about data distribution. To avoid these issues, a new, nonparametric method based off of the random forest (RF) algorithm was introduced, named missForest. In addition to its ability to handle mixed-type data, RFs high performance and efficiency under difficult conditions such as higher dimensions, non-linear relations, or complex interactions between variables made it a fitting choice for data imputation (Stekhoven & Bühlmann, 2011).

Stekhoven and Bühlmann (2011) trained an RF on existing datasets and used it to predict missing variables, then repeated these steps iteratively until a stopping criterion was reached, i.e., when the difference between the two last-imputed data matrices increases. This model was tested on both continuous and categorical variables. For continuous variables, the performance was determined by calculating the normalized root mean squared error (NRMSE), defined by

$$NRMSE = \sqrt{\frac{mean((X^{true} - X^{imp})^2)}{varX^{true}}} \tag{2}$$

where X^{true} is the original complete matrix and X^{imp} is the imputed matrix. For categorical variables, the proportion of falsely classified entries (PFC) over the total number of missing values was used. The results of MissForest on continuous variable datasets were compared to those of

KNNimpute and MissPALasso, and the results for categorical and mixed-type (both continuous and categorical variables) datasets were compared with the MICE algorithm and KNNimpute with dummy coding. It was found that overall, MissForest outperformed the other methods, particularly for the categorical and mixed-type variables, although its computational efficiency was second to KNNimpute (Stekhoven & Bühlmann, 2011). Further research can be conducted on decreasing missForest runtimes or incorporating this method in HCC prognosis evaluation.

2 Problem

Using data retrieved from the UCI Machine Learning Repository (Dua & Graff, 2017), two models for Hepatocellular Carcinoma survival prediction were constructed in the programming language R, with the goal of determining the more optimal method.

1. How should the problem of missing values be approached?
2. How can random forests be incorporated to increase the accuracy of the model's predictions?
3. Does the use of principal component analysis along with support vector machine provide better results?

3 Methodology and Results

3.1 Missing Data Imputation

3.1.1 Random Forests

Before dealing with data imputation, the random forest algorithm must first be understood. Among the most powerful ensemble algorithms in machine learning, random forests are composed of many classification trees, also known as decision trees, each of which aims to split variables in such a way as to generate the optimal two-child subsets, forming branches or nodes at which the data is classified conditionally into two subsets (in the simplest form, "yes" and "no"), and at each of these new nodes, the data is split again. This continues on until a leaf node is formed and branching is no longer possible. Thus, each instance must start at the root node and travel along the branches until it ultimately reaches a leaf node, at which a "decision" is made (Kingsford & Salzberg, 2008).

A random forest grows many of these classification trees, each of which are grown by randomly sampling N cases with replacement from the original dataset. At each node of the tree, a constant number of variables m, where $m << M$ and M represents the number of input variables, is selected to split the node. This process repeats until the tree reaches the largest size possible (Breiman & Cutler, 2015). When all the trees have been grown, each instance is run down all the trees, receiving a class for each one. At the end, the final class is determined for that data point by averaging the classes of each tree, and the process repeats for the remainder of the data (Figure 1).

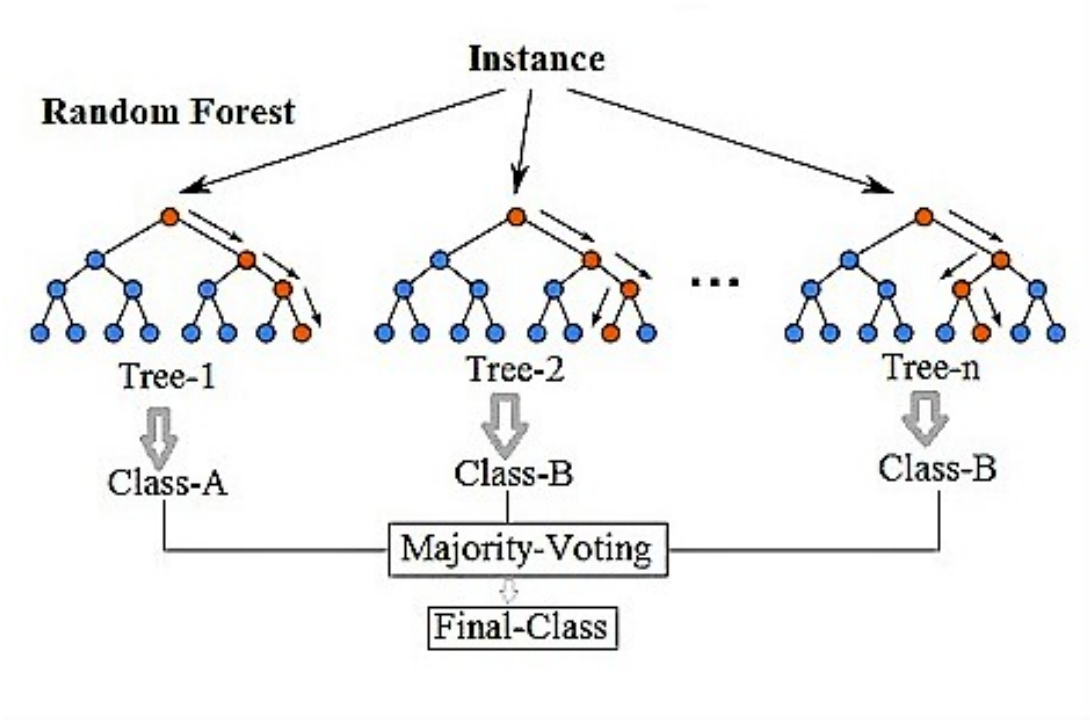

Figure 1 Simplified random forest. An instance is run through all the trees. At each node (represented by a circle), it is further classified until it reaches a leaf node, at which there no longer are branches for it to travel across. Each tree thus arrives at a classification, and the final class is determined through a process of "voting" across all trees (Reinstein, 2017).

3.1.2 Data Imputation

For the process of missing data imputation, the method applied was the missForest algorithm, proposed by Stekhoven and Bühlmann (2011). Based on random forest, MissForest is a nonparametric imputation method that can be applied to almost every type of data, with the only requirement that the rows of the dataset are pairwise independent. As missForest can cope with difficult conditions such as mixed-type data, higher dimensions, non-linear relations, or complex interactions between variables, it is a fitting choice for data imputation. MissForest essentially fits a random forest on

the observed part of each variable, and predicts the missing values. This iterative process repeats until a stopping criterion is reached, typically when the difference between the two last-imputed data matrices increases or the maximum number of iterations is reached.

The data, retrieved from the UCI Machine Learning Repository, consists of 165 patients diagnosed with HCC, and contains 49 features including demographic, risk factors, laboratory, and overall survival. Missing data represents 10.22% of the entire dataset, as all but eight patients have incomplete information (Dua & Graff, 2017). Therefore, data imputation was necessary. For each variable X_s, a random forest was first fitted on the observed values of X_s as well as the observed values of other variables, then applied to the missing values of other variables, which were finally used to predict the missing value of X_s. The imputation procedure was repeated until a stopping criterion was met, as described in the previous paragraph (Stekhoven & Bühlmann, 2011). Afterwards, two approaches were taken to predict survivability past one year: Random Forest and Principal Component Analysis combined with Support Vector Machine (PCA-SVM).

3.2 Method One: Random Forest

The constructed model used the random forest algorithm to divide the HCC patients into two classes: Positive (the patient will not survive past one year) and Negative (the patient will survive past one year). The basics of this algorithm are described in Section 3.1.1. The number of trees was set to 100 and the constant number of variables m randomly sampled as candidates at each node was set by default to $\sqrt{M}$, where $M = 49$ input variables (Breiman & Cutler, 2011).

3.3 Results of Method One

After 100 runs, the results (Table 1) were determined by comparing the model's predictions with the original dataset through three calculations: accuracy, true positive rate (TPR), and true negative rate (TNR), which are defined by the following

$$Accuracy = \frac{TP + TN}{P + N} \tag{3}$$

$$TPR = \frac{TP}{P} \tag{4}$$

$$TNR = \frac{TN}{N} \tag{5}$$

where TP represents the number of true positives, the instances when the model correctly predicts positive, TN represents the number of true negatives, when the model correctly predicts negative, P represents the total number of positives in the dataset, and N represents the total number of negatives. Figures 2, 3, and 4 show the distribution of the Random Forest model's accuracy, TPR, and TNR after 100 runs.

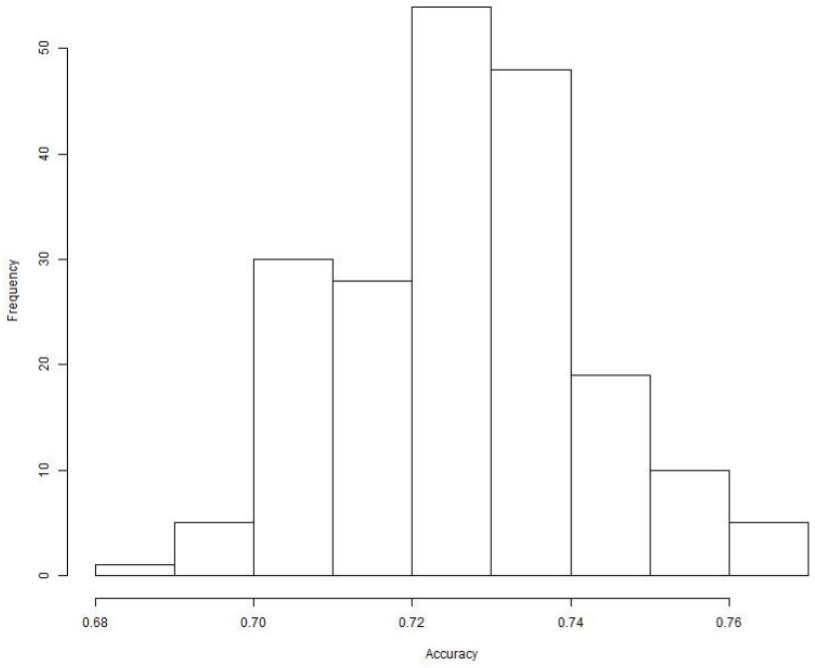

Figure 2 The distribution of the Random Forest model's overall accuracy after 100 runs, calculated by $\frac{TP+TN}{P+N}$, where TP represents the number of true positives, the instances when the model correctly predicts positive, TN represents the number of true negatives, when the model correctly predicts negative, P represents the total number of positives in the dataset, and N represents the total number of negatives (Created by Student Researcher).

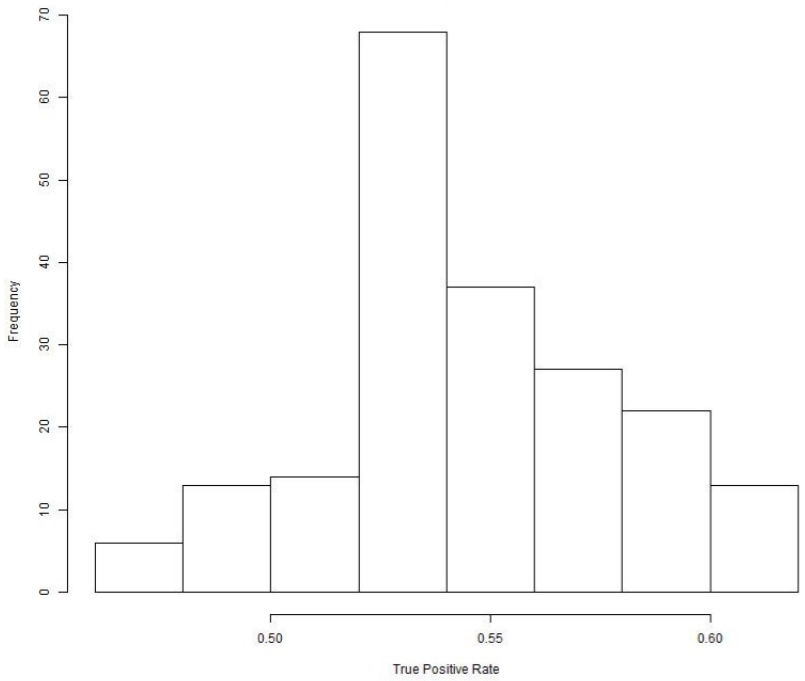

Figure 3 The distribution of true positive rates (TPR), after 100 runs of the Random Forest model, calculated by $\frac{TP}{P}$ (Created by Student Researcher).

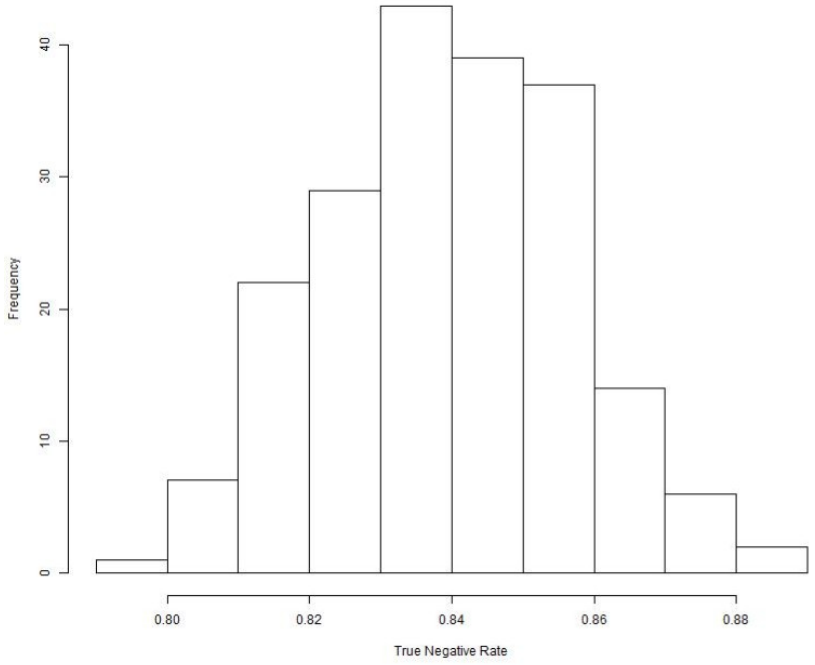

Figure 4 The distribution of true negative rates (TNR) after 100 runs of the Random Forest model, calculated by $\frac{TN}{N}$ (Created by Student Researcher).

Table 1 Comparison of the mean, maximum, minimum, and standard deviation for the overall accuracy, TPR, and TNR, taken across the 100 runs of the Random Forest model (Created by Student Researcher).

	Accuracy	**TPR**	**TNR**
Mean	0.727	0.547	0.838
Maximum	0.764	0.619	0.882
Minimum	0.685	0.476	0.794
Standard Deviation	0.0159	0.0326	0.0172

3.4 Method Two: PCA-SVM

Principal component analysis (PCA) is a popular statistical tool for analyzing the inherent structure of the data. It is typically used with high dimensional data, when there is a large number of features, or variables, which increases the difficulty of data mining tasks such as classification, as it reduces the dimensionality of the data through coordinate axis rotation (Lin, 2018). PCA is a method of feature extraction in which highly correlated input variables are grouped into new variables using a mapping function so that less influential variables are dropped while the most valuable parts of all variables are retained. After this process, each of the new variables are independent of each other (Astuti & Adiwijaya, 2018). In summary, PCA can transform a set of high-dimensional and correlated data to a set of lower-dimensional, uncorrelated data (Lin, 2018).

For PCA, the variance measures the amount of information in the data, and principal components can be viewed as the direction that captures the maximum amount of variance, mapping the data onto a subspace of lower dimensions. Accordingly, the most information is contained within the first principal component, or in other words, it accounts for the most variance, while the second principal component thus accounts for the second largest variance under the condition that it is not correlated to the first component, etc.

A support vector machine (SVM) is a classifier that is particularly suited for modeling complex data due to its high accuracy and a smaller tendency to overfit. SVM essentially searches for the optimal line or hyperplane (decision boundary) to separate data into classes, with the aim of maximizing the margin between datapoints and the boundary. When the data are not linearly separable, as in most cases, SVM maps the original data into a higher dimension nonlinearly in order to obtain the separating hyperplane (Astuti & Adiwijaya, 2018).

3.5 Results of Method Two

The results (Table 2) of the PCA-SVM model were determined with the same method used to find the results of the Random Forest model (Section 3.3). In summary, the model was run 100 times and the overall accuracy, TPR, and TNR were calculated (Equations 3, 4, and 5). Figures 5, 6, and 7 show the distribution of the PCA-SVM model's accuracy, TPR, and TNR.

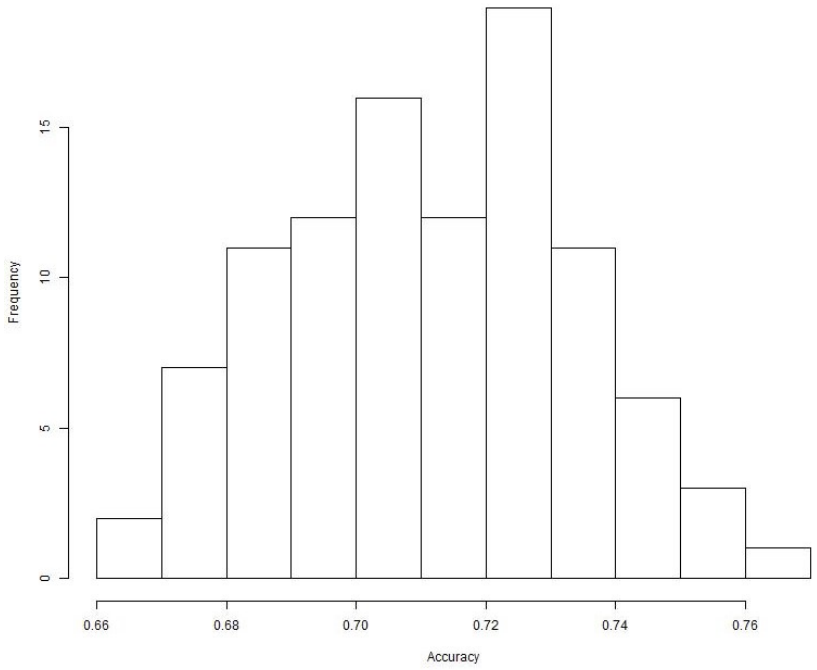

Figure 5 The distribution of the PCA-SVM model's overall accuracy after 100 runs, calculated by $\frac{TP+TN}{P+N}$ (Created by Student Researcher).

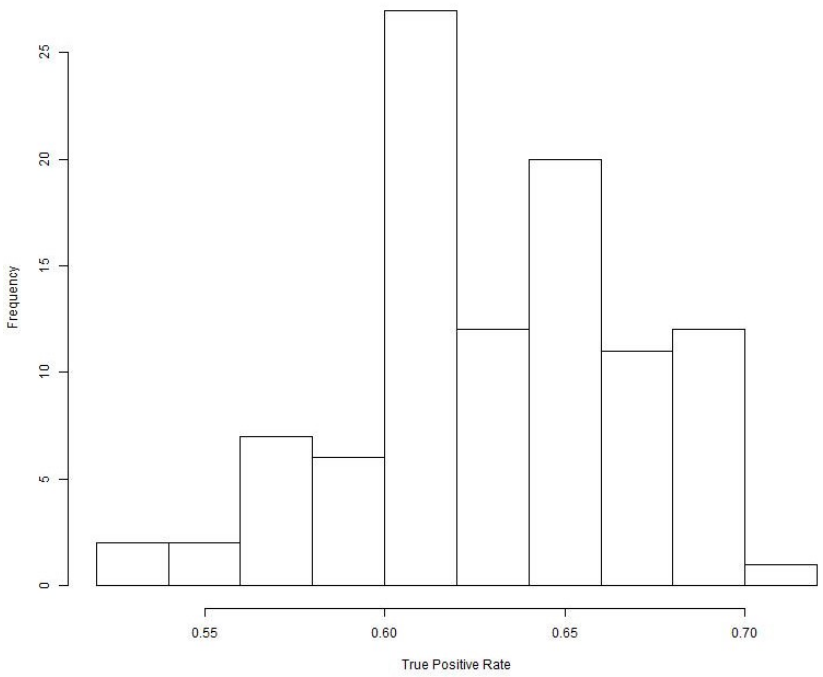

Figure 6 The distribution of true positive rates (TPR) after 100 runs of the PCA-SVM model, calculated by $\frac{TP}{P}$ (Created by Student Researcher).

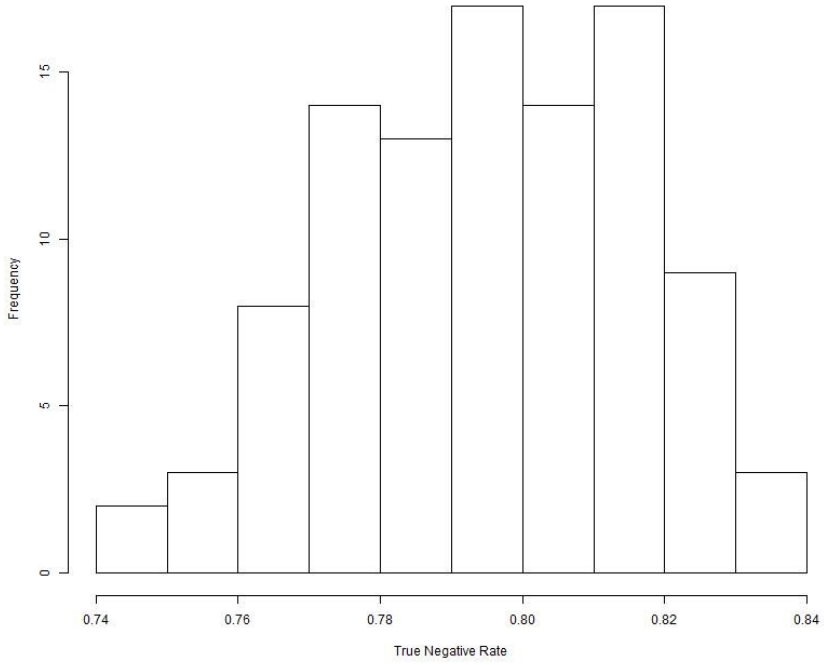

Figure 7 The distribution of true negative rates (TNR) after 100 runs of the PCA-SVM model, calculated by $\frac{TN}{N}$ (Created by Student Researcher).

Table 2 Comparison of the mean, maximum, minimum, and standard deviation for the overall accuracy, TPR, and TNR, taken across the 100 runs of the PCA-SVM model (Created by Student Researcher).

	Accuracy	**TPR**	**TNR**
Mean	0.712	0.630	0.794
Maximum	0.764	0.714	0.833
Minimum	0.667	0.540	0.745
Standard Deviation	0.0219	0.0382	0.0209

4 Analysis and Conclusion

Overall, the Random Forest model produced significantly higher accuracies and TNRs ($p < 0.05$) over the 100 runs when compared with the PCA-SVM model, while the latter had significantly higher TPRs ($p < 0.05$) (Tables 1, 2, and 3). Thus, the PCA-SVM model was superior in correctly predicting Positive prognoses; that is, it was more successful in determining when a patient was in critical condition and would not survive past one year. Although its overall TNR was significantly lower than that of the Random Forest model, indicating that it was more likely to falsely classify a patient as Positive, regarding matters of health, it is much more appropriate to err on the side of caution. Furthermore, although the overall accuracy of the PCA-SVM model was lower, the difference was not too large – on average, only a difference of about 0.015, and the p-value is much larger in comparison with those of the other two categories, indicating a lesser significance (Table 3). The difference between the accuracies of the two models, which measured overall performance, was much smaller than the difference between the TPRs and TNRs and thus can be disregarded. Accordingly, the PCA-SVM model was proposed as a reliable method of HCC prognosis evaluation, and can be extended to applications in other diseases as well.

Table 3 P-values of the two-sample t-test for the difference between the mean Accuracy, TPR, and TNR of the Random Forest model and the PCA-SVM model. All p-values were less than 0.05, and therefore the difference was statistically significant. The negative critical value t^* for TPR indicated that the second model (PCA-SVM) had a higher performance when predicting Positive prognoses; that is, it was more successful in determining that a patient was in critical condition and would not survive past one year (Created by Student Researcher).

	Accuracy	**TPR**	**TNR**
p-value	3.031E-7	3.51E-39	1.54E-37
critical value t^*	5.32	-16.6	16.1

This research in hepatocellular carcinoma (HCC) prognosis evaluation improves upon existing models, providing more accurate predictions of whether patients will survive past one year. With knowledge of the stage and severity of the cancer, more effective treatments that better target the patients specific needs can be provided, which would likely increase HCC survival rate. It is important now that hospitals take steps to implement the model in order to reap the full benefits, as the proposed HCC survival prediction model can additionally be extended to other areas.

References

Astuti, W., et al. (2018). Support vector machine and principal component analysis for microarray data classification. *Journal of Physics: Conference Series*, *971*(1), 012003.

Breiman, L., & Cutler, A. (2011). Manual–setting up, using, and understanding random forests v4. 0. 2003. Retrieved from `https://www.stat.berkeley.edu/\~{}breiman/Using\_random\_forests\_v4.0.pdf`

Breiman, L., & Cutler, A. (2015). Random forests. *Random Forests-Classification Description*.

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Ke, K., Chen, G., Cai, Z., Huang, Y., Zhao, B., Wang, Y., ... Liu, J. (2018). Evaluation and prediction of hepatocellular carcinoma prognosis based on molecular classification. *Cancer management and research*, *10*, 5291.

Key statistics about liver cancer. (2019). *American Cancer Society*. Retrieved from `https://www.cancer.org/cancer/liver-cancer/about/what-is-key-statistics.html`

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, *26*(9), 1011.

Piech, C. (2013). K means. *CS221*. Retrieved from `http://stanford.edu/~cpiech/cs221/handouts/kmeans.html`

Rawla, P., Sunkara, T., Muralidharan, P., & Raj, J. P. (2018). Update in global trends and aetiology of hepatocellular carcinoma. *Contemporary Oncology*, *22*(3), 141.

Reinstein, I. (2017, October). *Random forests(r), explained.* Retrieved from `https://www.kdnuggets.com/2017/10/random-forests-explained.html`

Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, *58*, 49–59.

Stekhoven, D. J., & Bühlmann, P. (2011). Missforest–non-parametric missing value imputation for

mixed-type data. *Bioinformatics*, *28*(1), 112–118.