Lingfei Zhao
ISEF Research Plan

Improving hepatocellular carcinoma survival prediction with artificial intelligence strategies

A. RATIONALE:

Currently, Hepatocellular Carcinoma (HCC) is the sixth most frequently diagnosed cancer and the third-leading cause of cancer-related deaths in the world (Blue Faery). 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 would include 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 (Santos et al., 2015). 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. Further research into methods of dealing with these problems in current HCC prognosis evaluation is particularly important, as accurate analysis of the tumor's behavior at a certain stage will allow for the correct treatment to be given, thus increasing survival rate.

dropped while the most valuable parts of all variables are retained. Thus, the high dimensionality of the data will be reduced, allowing the model to better predict Positive or Negative.

Risks and Safety

There are no potential risks or safety precautions necessary, as the data is de-identified from a public data source.

Data Analysis:

The results will be determined by comparing the model's predictions with the original dataset. That is, each predicted prognosis will be compared to the true prognosis of the patients from whom the data was obtained. For instance, if a patient is predicted to be Positive and the dataset reveals that he/she was indeed Positive, this will be considered a True Positive. Three calculations will be made: accuracy, true positive rate (TPR), and true negative rate (TNR), which are defined by the following

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

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

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

where *TP* represents the number of positives (the instances when the model correctly predicts positive), *TN* represents the number of true negatives (the model correctly predicts negative), *P* represents the total number of positives in the dataset, and *N* represents the total number of negatives.

The results of both models will be compared to determine the model with the highest overall success.

D. BIBLIOGRAPHY:

Breiman, L., & Cutler, A. (2011). Manual–setting up, using, and understanding random forests V4. 0. 2003. URL https://www. stat. berkeley. edu/~ breiman/Using_random_forests_v4. 0. pdf.

4. Hazardous Chemicals, Activities, and Devices:

No hazardous chemicals, activities, or devices will be used.