PanCan Diagnosed: Developing an Algorithm for the Accurate and Affordable Early Diagnosis of Pancreatic Cancer via Machine Learning and Bioinformatics

Goel, Siya (School: West Lafayette Junior/Senior High School)

Pancreatic cancer (PC) is the fourth leading cause of cancer death in the United States because it's five-year survival rate is 8.9%. This occurs because of late diagnosis which is affiliated with the hidden location of the pancreas, making current screening methods unavailable. Previous research also achieves low (70-75%) diagnostic accuracy, because 80.1% of PC cases are affiliated with diabetes, leading to misdiagnosis. To address the problems of frequent late diagnosis and misdiagnosis, an accessible, accurate, and affordable diagnostic tool for PC was developed by analyzing the expression of 19 genes in PC and diabetes. In the first stage, machine learning algorithms (Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Neural Network) were made to divide samples into four groups depending on the occurrence of PC and Diabetes. The models were analyzed with 400 PC samples of varying stages to ensure validity. Naive Bayes, Neural Network, and KNN models achieved the highest testing accuracy of around 92.6%. In the second stage, a novel user interface, PanCan Diagnosis, was designed to test an individual's likelihood of PC occurrence. In the third stage, the biological implication of the 19 genes was investigated using bioinformatics tools like String-DB, GeneCards, PathCards, and BioGPS. It was found that these genes were significantly involved in regulating the cytoplasm, cytoskeleton, and nuclear receptor activity in the pancreas. PanCan diagnosis is the first affordable clinical test in literature that achieves a PC diagnostic accuracy above 90%, potentially increasing five-year survival rate to 39.5%.