Creation of a Machine Learning App to Facilitate Pancreatic Cancer Prediction

Reid Fleishman

Abstract

While pancreatic cancer is a deadly disease that is difficult to diagnose at an early stage, early diagnosis yields a much larger survival rate, making it a key step for treatment. Machine learning has shown promise in the early diagnosis of many diseases by providing more accurate predictions than doctor's intuition. The goal of this study was to develop Decision Tree (DT), Random Forest (RF), Boosted Trees (BT), Logistic Regression (LR), and Support Vector Machine (SVM) models to predict one's risk of developing pancreatic cancer, improving upon previous Artificial Neural Network (ANN) and LR models. In addition, this study investigated the effects of including depression data on model performance, tested the models using an additional dataset, and created a mobile app to facilitate prediction by an end-user. The training dataset consisted of 17 pancreatic cancer risk factors, including demographic (i.e., age, race), lifestyle (i.e., smoking, alcohol), and disease (i.e., diabetes, hypertension) factors. Data were derived from two de-identified and publicly-available datasets: the Integrated Public Use Microdata Series (IPUMS) healthy surveys and the National Cancer Institute's Prostate, Lung, Colorectal and Ovarian (PLCO) study, encompassing a total of 752,527 patients, 983 of whom had pancreatic cancer. Data were re-coded, normalized, and split into training, validation, and testing sets. Measures were also taken to balance the classes and to handle missing values. The hyperparameters of the DT, RF, BT, LR, and SVM models were chosen to maximize area under the receiver operating characteristic curve (AUC), an important metric to evaluate the predictive power of the models. The BT model achieved a sensitivity (true positive rate) of 0.788, a specificity (true negative rate) of 0.791, and the highest AUC (0.870) out of the five algorithms and thus was chosen to be embedded within the mobile app. Models were also tested on an independent, de-identified dataset consisting of pancreatic cancer cases made available by the Pancreatic Cancer Action Network (PanCAN), but missing values and the absence of key variables limited usefulness and yielded inconsistent results. In addition, age, physical activity, smoking, and race were important predictors of pancreatic cancer, while depression, among others, were not. In this study, machine learning models were created and incorporated into an app to help identify one's risk of developing pancreatic cancer at an early stage.

Introduction

Pancreatic cancer is the third leading cause of death among the most commonly diagnosed cancers (Siegel et al., 2020). The average five-year relative survival rate of patients diagnosed with pancreatic cancer is 9%, and in 2020 an estimated 47,050 people will die from an estimated 57,600 new cases of pancreatic cancer (Siegel et al., 2020). However, while still low, the survival rate for patients diagnosed with pancreatic cancer at an early stage is 37%, indicating that early diagnosis is key for survival (Siegel et al., 2020). Effective screening for pancreatic cancer at an early stage is lacking, and with current techniques, it is impractical to screen the entire population on a regular basis since pancreatic cancer has an extremely low lifetime risk of 1% and screening can expose patients to harmful radiation (McGuigan et al., 2018; Poruk et al., 2013). As a result, it is vital that an alternative method is developed to identify those with a high risk of developing pancreatic cancer in order to increase their chance of survival.

As supervised machine learning has gained popularity, its use in the early detection of major diseases such as pancreatic cancer has been explored. With the increase in electronic patient records, more training data is available for use in machine learning models, increasing their effectiveness and popularity (Uddin et al., 2019). Moreover, machine learning models provide predictions of diseases with greater accuracy than traditional statistics or doctor's intuition because there is often human error and bias (Uddin et al., 2019; Palaniappan & Awang, 2008). In practice, using machine learning for disease prediction has been successful for at least 49 different diseases, including heart disease, diabetes, breast cancer, and Parkinson's disease (Uddin et al., 2019). As a result, developing an effective machine learning model to predict one's risk of developing pancreatic cancer could help with its early diagnosis.

Machine learning models for disease prediction are evaluated in three main ways — sensitivity, specificity, and AUC (area under the receiver operating characteristic (ROC) curve) — each of which provides different information about a model's performance. Sensitivity, or true positive rate, is defined as the rate at which the model correctly identifies patients who have the disease, while specificity, or true negative rate, is defined as the rate at which the model correctly identifies patients who do not have the disease (Uddin et al., 2019; Trevethan, 2017). AUC is the probability that a model ranks a random positive example it has never seen before closer to positive than a random negative example that it has never seen before ("Classification: ROC Curve and AUC," n.d.). In other words, the models best at discriminating between pancreatic cancer cases and healthy cases will have the highest sensitivity, specificity, and AUC values.

There is discussion over which type of machine learning algorithm yields the highest sensitivity, specificity, and AUC values. Uddin et al. (2019) examined multiple algorithms and their respective results encompassing 48 different studies for 49 different disease prediction problems. Each study used

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

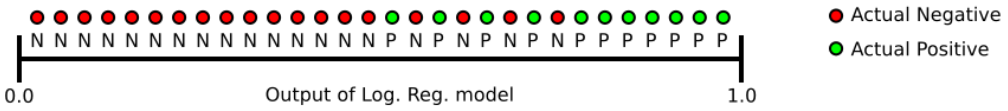

Figure 1: Top: Formulas for the calculation of sensitivity and specificity values (Trevethan, 2017). *TP* = True Positives; *TN* = True Negatives; *FP* = False Positives; *FN* = False Negatives. Bottom: In this scenario, AUC equals the probability that a random positive (green) example is ranked closer to 1.0 than a random negative (red) example ("Classification: ROC Curve and AUC," n.d.).

more than one algorithm, covering Artificial Neural Network (ANN), Logistic Regression (LR), Decision Tree (DT), K-nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). In their study, Uddin et al. (2019) found that the tree-based models, RF and DT, achieved superior performance the most and third-most times, respectively, compared to the other models tested. As Uddin et al. (2019) notes, the performance of the models can vary from problem to problem, as each algorithm has its own strengths and weaknesses. For instance, ANNs are strong at determining very complex non-linear relationships between variables, whereas LRs are effective with simpler problems (Uddin et al., 2019). The performance of various algorithms could also be a function of the dataset used, as one lab group, involved in the prediction of pancreatic cancer, skin cancer, lung cancer, prostate cancer, and colorectal cancer, consistently found that when using the same dataset, the ANN performed better than the Linear Discriminant Analysis (LDA), SVM, NB, DT, RF, and LR algorithms (Muhammad et al., 2019; Nartowt et al., 2020; Roffman et al., 2018a, Roffman et al., 2018b). Given the wide variability in results when training models, it is vital to evaluate multiple algorithms for each project.

There are many risk factors of pancreatic cancer, including age, sex, race, BMI, asthma, coronary heart disease/attack, diabetes, emphysema, hepatitis/cirrhosis, hypertension, stroke, alcohol consumption, smoking, and amount of exercise (Lowenfels & Maisonneuve, 2006; Arnold et al., 2009; Larsson et al., 2007; Gomez-Rubio et al., 2015; Muhammad et al., 2019; Everhart & Wright, 1995; Hassan et al., 2008; Ye et al., 2002; Lindgren et al., 2005; Zheng et al., 1993; Coughlin et al., 2000; Michaud et al., 2001; "Pancreatic Cancer Risk Factors," n.d.). There have been several machine learning models already created to predict one's risk of developing pancreatic cancer using commonly available patient data, such as Muhammad et al. (2019)'s ANN and Hsieh et al. (2018)'s ANN and LR models. While these models performed well in terms of sensitivity, specificity, and AUC (see **Table 1**), neither of them included another important risk factor for predicting pancreatic cancer: depression. There are a few studies that have shown this link. In one study, psychiatric illnesses (i.e., depression/anxiety) before medical symptoms were reported between 33% and 45% of pancreatic cancer patients (Kenner, 2018). Another

study found that twice as many patients reported depression one year before diagnosis of pancreatic cancer than those who were not diagnosed with pancreatic cancer (Olson et al., 2016). Given that the inclusion of relevant variables can improve model performance, it would be crucial to incorporate this risk factor to potentially achieve better predictions (Hall & Smith, 1998). In addition, neither of those studies developed a way for patients to use their models.

Table 1: Sensitivity, specificity, and AUC values of previous pancreatic cancer models

Model	Sensitivity	Specificity	AUC
Muhammad et al., 2019 (ANN)	0.807	0.807	0.850
Hsieh et al., 2018 (ANN)	0.873	(not calculated)	0.642
Hsieh et al., 2018 (LR)	0.998	(not calculated)	0.707

Muhammad et al. (2019); Hsieh et al. (2018).

There were three major goals of this study. The first goal was to build a model with higher sensitivity, specificity, and AUC values compared to the previous pancreatic cancer models shown in **Table 1** to predict one's risk of developing pancreatic cancer. This was proposed by including the depression variable in the models and using additional types of algorithms (DT, RF, Boosted Trees (BT), LR, and SVM). The second goal was to test the models on an additional dataset consisting of pancreatic cancer patients to further evaluate their performance. Finally, the third goal was to develop an iPhone app for doctors and/or patients to input data into the model to generate a prediction in an easy-to-use fashion.

Methods

Datasets

The models were trained on publicly-available, de-identified datasets from the Integrated Public Use Microdata Series (IPUMS) health surveys, a collection of data from the National Health Interview Survey (NHIS) dataset; as well as the "Pancreas" data from the National Cancer Institute's Prostate, Lung, Colorectal and Ovarian (PLCO) study (Blewett et al., 2019; "PLCO," n.d.). The models were also tested on an independent, de-identified dataset made available by the Pancreatic Cancer Action Network (PanCAN).

The IPUMS dataset consisted of 602,558 participants, 215 of whom had pancreatic cancer, from 1999-2018. Each participant was asked questions about their health and lifestyle and each answer was recorded as a numerical code. For example, one question was, "have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?" and the answers were represented as 1 for "No" and 2 for "Yes." Questions pertaining to continuous variables such as age were asked in a similar manner and answers were represented as the exact value given.

The PLCO dataset consisted of 149,969 participants, 768 of whom had pancreatic cancer, from 1993-2009. Data from the Baseline Questionnaire, Supplemental Questionnaire, and Diet History Questionnaire were used. Each participant was asked questions about their health and lifestyle and each answer was recorded as a numerical code. For example, one question was, "has a doctor ever told you that you have any of the following conditions?" with checkboxes identifying a wide variety of diseases/conditions, and the answers were represented as 0 for "No" and 1 for "Yes." Questions pertaining to continuous variables were asked in a similar manner and answers were represented as the exact value given.

The PanCAN dataset consisted of 1101 participants, all of whom had pancreatic cancer, from 2015-2020. Each participant was asked questions about their health and lifestyle and each answer was recorded as a numerical code or a string of text. For example, one question was, "I have been diagnosed with diabetes?" and the answers were represented as "No" and "Yes." Questions pertaining to continuous variables were asked in a similar manner and answers were represented as the exact value given.

The combined IPUMS + PLCO dataset consisted of 752,527 participants, 983 of whom had pancreatic cancer, from 1993-2018.

There were 17 variables and 1 label (pancreatic cancer) included in the combined IPUMS + PLCO dataset that were chosen based upon availability in both datasets along with their previously studied correlation with pancreatic cancer. However, only 6 of those variables were available in the PanCAN dataset. Further information about the variables used in this study are detailed in the **Appendix**. Please note that the numbers of participants in this section were calculated after the following re-coding processes were performed.

Re-Coding of the Datasets

Many processes were performed to handle missing data and to re-code values in the datasets. First, cases in all three datasets that were missing pancreatic cancer data were deleted. In the PanCAN dataset, cases which either only had the age variable or were missing the age variable entirely were deleted. In all datasets, cells with codes representing missing or unknown answers were cleared. Moreover, some variables were either combined with others or computed from others to match their definitions/criteria across all three datasets. Finally, in order for the models to make proper predictions, all variables were re-coded such that each specific value represented the same thing across all three datasets. Dataset re-coding was performed in IBM SPSS and Microsoft Excel.

Data Preprocessing and Split

After re-coding, the combined IPUMS + PLCO dataset was exported from SPSS to a *.csv* file. This file was uploaded to Google Colab for data pre-processing, where all code was custom written in the Python programming language version 3.6.9 using the Numpy, Pandas, Sklearn, and Matplotlib libraries.

Since the models cannot handle missing values, they were replaced with the mean of the values present in that particular column for continuous data and replaced with the mode of the values present in that particular column for categorical data (Aljuaid & Sasi, 2016). Data were then randomly split with 70% data for model training, 20% data for model validation, and 10% data for model testing. This split was chosen in order to maximize the amount of data used for model training while still leaving enough for testing. Continuous variables were then normalized as per standard machine learning practice (Muhammad et al., 2019; Roffman et al., 2018a, Roffman et al., 2018b).

The dataset was extremely unbalanced (out of the total 752,527 participants in the dataset, only 983 of them had pancreatic cancer, or 0.13%). If left untouched, a model would simply learn to predict each case as negative, therefore rendering it useless. As a result, random up-sampling of the minority class (pancreatic cancer cases) was performed on the training data until both classes were balanced, as per standard machine learning practice (Boyle, 2019).

Model Training, Validation, and Testing Procedures

Five algorithms — DT, RF, BT, LR, and SVM — were custom written in the Swift programming language version 5.3 with Apple's CreateML framework ("Create ML," n.d.). CreateML was the chosen framework for this study because it enabled the development of the above algorithms as well as the ability to embed them natively within an iPhone app. Unfortunately, an ANN was unable to be created due to limitations within CreateML.

Training: All models in this study were trained on the training data using a 2019 iMac with a 3.2 GHz 6-core Intel Core i7 processor. All code was compiled using Xcode 12.1 running on macOS Big Sur 11.0 beta 10.

Validation: During training, each model was evaluated on the validation data after each iteration and automatically tuned based on those results. To prevent overfitting, the models stopped training automatically upon an increase in the validation loss value.

Testing: After training, each model was saved as a *.mlmodel* file and imported into a Xcode project containing a simulated app environment within the CoreML framework ("Core ML," n.d.). The model was evaluated on each row of the testing data and its corresponding prediction value (0-1) was recorded to a *.csv* file. This file was then uploaded to Google Colab where sensitivity, specificity, and

AUC values were calculated using custom Python code. Testing of the models on the PanCAN dataset was done directly within CreateML.

Model Hyperparameter Tuning

Each algorithm has hyperparameters, or values that must be set manually, that when tuned can lead to better results. A baseline of each algorithm's AUC value was established using its default hyperparameters. Each hyperparameter was adjusted, one at a time, and compared to the baseline. If the new model performed better than the baseline in terms of the AUC value, then that new hyperparameter was kept. If not, it was reset back to its previous value. AUC was the chosen metric for directly comparing models here and throughout this study because it represents the overall superiority of one model compared to another (Uddin et al., 2019).

This process was repeated for each hyperparameter of all five algorithms. The same random seed (42) was used for splitting the dataset between train/validation/test sets, up-sampling, and model initialization in order to ensure that everything else was kept constant during hyperparameter tuning. **Table 2** displays the final hyperparameters selected for each algorithm.

Table 2: Final hyperparameters of each algorithm.

Hyperparameter	Decision Tree (DT)	Random Forest (RF)	Boosted Trees (BT)	Logistic Regression (LR)	Support Vector Machine (SVM)
maxDepth	5	5	5		
maxIterations		10	10	10	11
minLossReduction	0	0	0		
minChildWeight	1000	1000	0.1		
stepSize			0.3	1	
earlyStoppingRounds			None		
rowSubsample		0.7	1		
columnSubsample		0.7	0.8		
l1Penalty				0	
l2Penalty				0.01	
penalty					1
convergenceThreshold				0.01	0.01
featureRescaling				True	True

Empty cells indicate that the hyperparameter does not apply to the algorithm.

Addition of the PLCO Dataset

Each of the five algorithms were trained using their default hyperparameters on just the IPUMS dataset as well as the IPUMS + PLCO dataset in order to determine whether including the PLCO data improved performance. The dataset with the highest mean AUC across all five models was used to train the models in the remainder of this study.

Final Model Training/Testing

Each algorithm was initialized with its respective final hyperparameters as shown in **Table 2**. Each algorithm was trained 10 times on the same set of 10 randomly-generated random seeds. Each model was initialized with the same set of random seeds such that, for example, run 5 of one algorithm could be compared to run 5 of another algorithm. The models were trained, validated, and tested in the same fashion as described in the "Model Training, Validation, and Testing Procedures" section. Each model was also tested on the PanCAN dataset. Given that the PanCAN dataset only contained pancreatic cancer patients, only a sensitivity value was calculated.

After the five algorithms were run 10 times each, the model from each algorithm that had the highest AUC value was designated as the final model of that respective algorithm. The model among those five best models with the highest AUC value was designated as the final model used in the iPhone app.

Variable Importance

When using the model in the real world, it would be important for doctors and/or patients to understand which variables are the most important predictors of pancreatic cancer. This can help determine whether the patient can input enough information to generate as accurate of a prediction as possible. To determine variable importance, the standard variable permutation technique was used. Each variable in the testing set was randomly shuffled and then evaluated on the final model, and the difference in AUC between the final model tested on the normal dataset and the final model tested on the dataset with the permuted variable was calculated (Casalicchio et al., 2019). Since permuting the values of a variable changes the relationship between that variable, other variables, and the label, a large positive difference in AUC before and after the permutation would indicate high importance of that variable in predicting pancreatic cancer (Casalicchio et al., 2019).

iPhone App

The final model's *.mlmodel* file was imported into an Xcode simulated app environment within the CoreML framework, exactly as was done during model testing. An iOS app user interface and

backend code were custom developed using the SwiftUI framework within the Swift programming language in order to provide a seamless way for patients to enter data into the model, generate a prediction on-device, and view that prediction ("SwiftUI," n.d.). The app follows all Apple *App Store Review* and *Human Interface* guidelines ("App Store Review Guidelines," n.d.; "Human Interface Guidelines," n.d.).

Results

This study developed machine learning models that aimed to improve upon previous models for predicting one's risk of developing pancreatic cancer, test the models using an additional dataset, and develop an effective medium for generating predictions. Training data were obtained from IPUMS and PLCO, preprocessed, and used to train five model algorithms 10 times each. The models included 17 unique variables, each linked to pancreatic cancer, and were trained on 752,527 participants with 983 pancreatic cancer cases. Sensitivity, specificity, and AUC of the models were of interest and therefore were analyzed. The models were also tested on the PanCAN dataset, which only included pancreatic cancer cases, where a sensitivity value was calculated.

Effect of Combining the IPUMS and PLCO Datasets

Before further training of the models, a dataset which yielded the highest AUC needed to be chosen. The means of the sensitivity, specificity, and AUC values of the five algorithms trained solely on the IPUMS dataset vs. trained on the combined IPUMS + PLCO dataset are shown in **Table 3**. The mean sensitivity and AUC of the five models trained on the IPUMS + PLCO dataset was greater than that of the IPUMS-only dataset by 0.099 and 0.021, respectively, and lower in mean specificity by 0.032. Since the mean AUC of the models trained on the IPUMS + PLCO dataset was higher than that of the IPUMS-only dataset, the combined dataset was used for the remainder of this study.

Table 3: Means of the sensitivity, specificity, and AUC values of the five algorithms trained solely on the IPUMS dataset vs. trained on the combined IPUMS + PLCO dataset.

Dataset	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
IPUMS	0.658 (0.438-0.879)	0.788 (0.730-0.846)	0.813 (0.758-0.868)
IPUMS + PLCO	0.757 (0.662-0.852)	0.756 (0.651-0.861)	0.834 (0.809-0.860)

Model Performance

The results of the mean sensitivity, specificity, and AUC values of each algorithm are shown in **Figure 2** and were computed for all five algorithms over 10 trials each. The mean sensitivity of the

models ranged from 0.703 (RF; 95% CI: 0.666-0.740) to 0.826 (SVM; 95% CI: 0.802-0.850); mean specificity ranged from 0.644 (SVM; 95% CI: 0.640-0.648) to 0.806 (RF; 95% CI: 0.785-0.827); and mean AUC ranged from 0.795 (SVM; 95% CI: 0.785-0.805) to 0.855 (BT; 95% CI: 0.845-0.865). The highest sensitivity a model achieved out of the 50 total trials was 0.859 by the SVM, meaning that "pancreatic cancer" was correctly predicted among those who had it 85.9% of the time in that trial. The highest specificity observed was 0.861 by the RF, meaning that "healthy" was correctly predicted among those who were healthy 86.1% of the time in that trial. The highest AUC observed was 0.870 by the BT, meaning that the probability of the BT ranking a random positive pancreatic cancer case closer to positive than negative is 87.0%. The mean sensitivities of the DT, RF, and BT were lower than their respective specificities, whereas the mean sensitivities of the LR and SVM were higher than their respective specificities. This resulted in higher mean AUCs for the DT, RF, and BT compared to the LR and SVM.

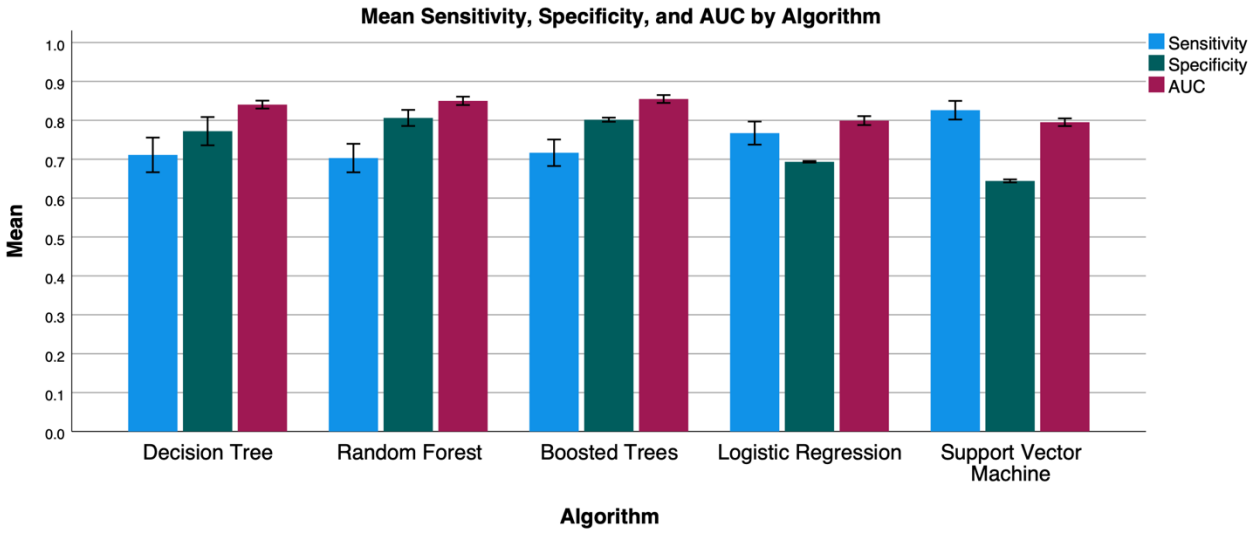

Figure 2: Data were obtained by evaluating the sensitivity, specificity, and AUC each model after training. Error bars represent the 95% confidence interval. n = 10 for all models.

Additionally, the metrics of the highest-AUC model from each algorithm are shown in **Table 4**. Of these models, the highest sensitivity (0.856) was achieved by the SVM; the highest specificity (0.791) was achieved by the BT; and the highest AUC (0.870) was achieved by the BT. The final model used in the app was the BT model since it achieved the highest AUC of these models.

Model Performance on the PanCAN Dataset

The results of the mean sensitivity of each algorithm tested on the PanCAN dataset are shown in **Figure 3** and were computed for all five algorithms over 10 trials each. The mean sensitivity of the models tested on the PanCAN dataset ranged from 0.411 (BT) to 1.000 (LR & SVM). The highest

Table 4: Sensitivity, specificity, and AUC values of the highest-AUC model from each algorithm.

Algorithm	Random Seed	Sensitivity	Specificity	AUC
Decision Tree (DT)	8242	0.798	0.711	0.861
Random Forest (RF)	8242	0.778	0.754	0.866
Boosted Trees (BT)	8242	0.788	0.791	0.870
Logistic Regression (LR)	604	0.814	0.693	0.823
Support Vector Machine (SVM)	604	0.856	0.640	0.819

sensitivity, 1.000, was observed in every trial of the LR and SVM algorithms, meaning that pancreatic cancer was correctly predicted among those who had it 100.0% of the time in those trials. The sensitivities of the DT, RF, and BT had very high variability (95% CI: 0.307-0.867, 0.112-0.724, 0.244-0.576, respectively), whereas the sensitivities of the LR and SVM had zero variability (95% CI: 1.000-1.000).

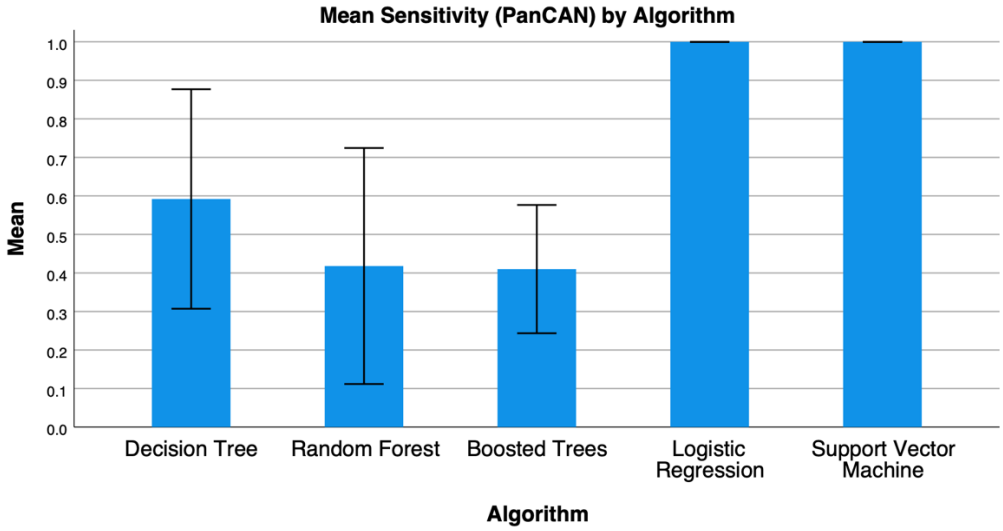

Figure 3: Data were obtained by evaluating the sensitivity of the models tested on the PanCAN dataset after training. Error bars represent the 95% confidence interval. n = 10 for all models.

Variable Importance

Figure 4 displays the difference in AUC between the final BT model and that from each variable permutation (see the **Appendix** for variable descriptions). The permutation of *Age* yielded the largest difference in AUC of 0.242, followed by *Moderate Activity Times Per Week* (0.071). *Strenuous Activity Times Per Week*, *Smoke 100*, *Race*, *Diabetes*, *Alcohol Days Past Year*, *Smoke Years*, *Asthma*, and *BMI* were also of diminishing but relative importance in predicting pancreatic cancer. The remaining variables indicated little importance (difference in AUC < 0.0005), with the permutations of *Functionally Limiting*

Depression, *Emphysema*, *Coronary HD or Attack*, and *Hypertension* actually improving AUC, though only by a maximum of 0.0008 (*Hypertension*).

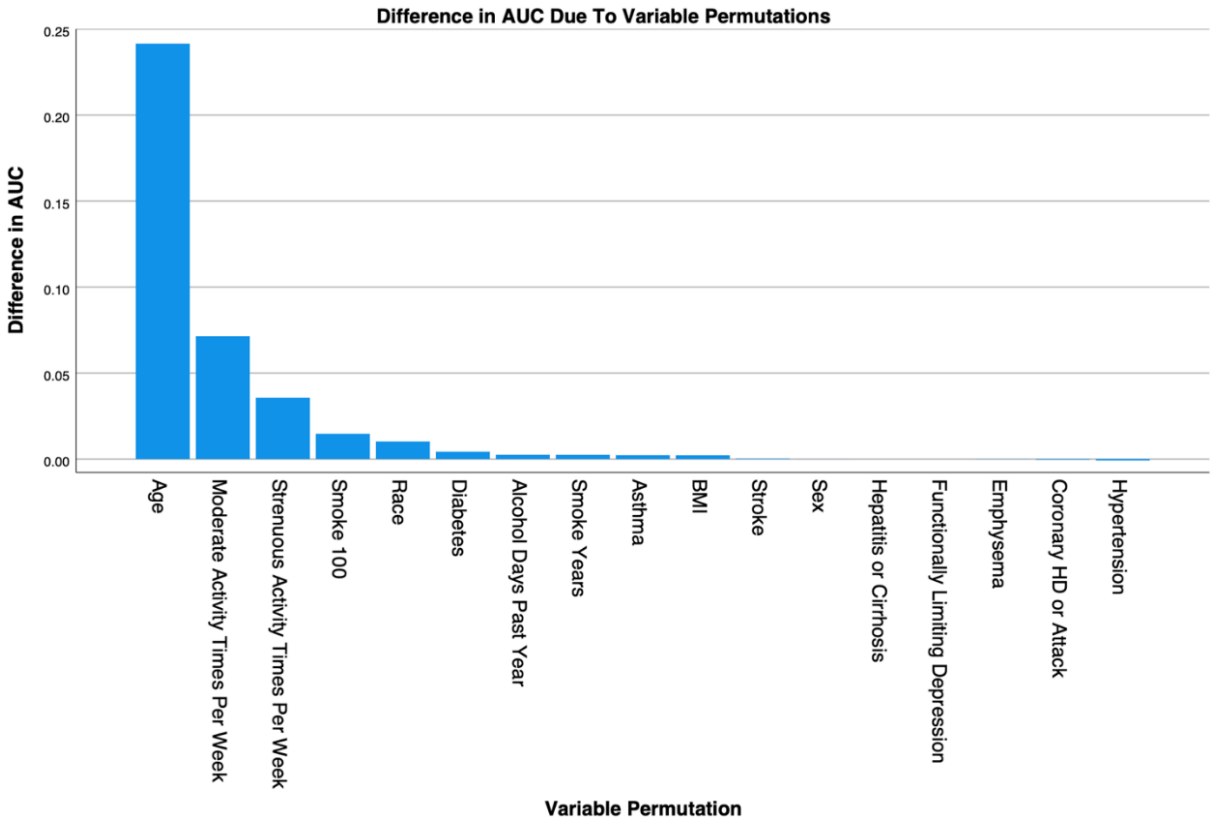

Figure 4: Data were obtained by evaluating the difference in AUC between the final BT model and that from each variable permutation. The greater the difference in AUC, the greater importance that given variable has on predicting pancreatic cancer.

iPhone App

If being used in a clinical setting, it would be advantageous to provide a simple, efficient, and easy-to-use way for doctors and/or patients to input data into the model to generate a prediction. Since smartphones are very prevalent in our everyday lives, creating an app to generate fast predictions would be ideal. As shown in **Figure 5**, an iPhone app was developed with data input and prediction output functionality. With privacy in mind, all data input by the user are fed into the model on-device and can be deleted after a prediction is generated. In addition, all predictions are stored on-device and can be discarded as well.

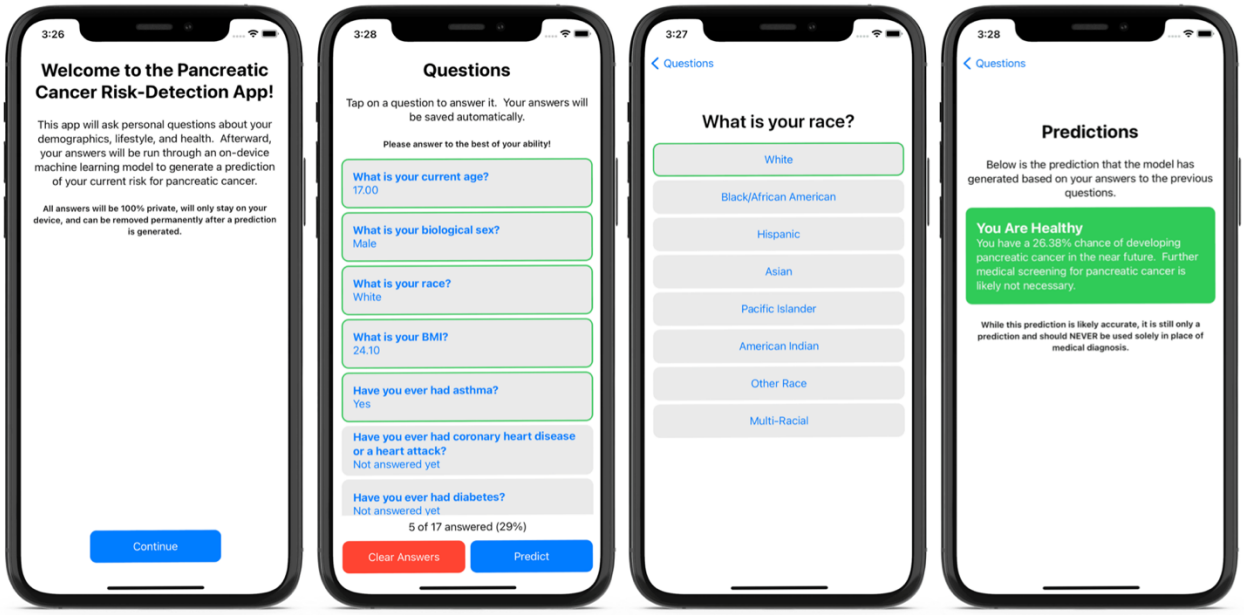

Figure 5: Screenshots of the iPhone app created to facilitate user input of data and prediction output. The final BT model was embedded within the app. From left to right: welcome screen; question list; categorical answer view; final predictions.

Discussion

The results of this study indicate that the inclusion of the PLCO dataset compared to the IPUMS dataset alone resulted in greater model performance in terms of sensitivity and AUC. In addition, the DT, RF, BT, LR, and SVM models showed promising results, each receiving AUCs of ~0.82 or greater. While the LR and SVM models achieved perfect predictions on the PanCAN dataset, the DT, RF, and BT models achieved lower sensitivities with a large variability. For the BT model, which received the highest AUC of all models, the age, physical activity, smoking, and race variables were important predictors of pancreatic cancer, while depression, among others, were not.

Comparison to Previous Pancreatic Cancer Models

Only the top-performing LR and SVM models achieved higher sensitivities than the ANN model from Muhammad et al. (2019), and no model achieved a higher sensitivity than the ANN or LR models from Hsieh et al. (2018). In addition, no model achieved a higher specificity than Muhammad et al. (2019)'s ANN model (specificity was not calculated for Hsieh et al. (2018)'s models). While only the top-performing DT, RF, and BT models achieved higher AUCs than Muhammad et al. (2019)'s ANN model, all models achieved higher AUCs than Hsieh et al. (2018)'s models.

The final model used in the app, the BT, achieved a higher AUC than all three of the previously published models, outperforming the highest-AUC model, Muhammad et al. (2019)'s ANN, by 0.02. As

such, the BT model is superior in discriminating between pancreatic cancer and healthy cases compared to the other models in this study and all models of previous studies. To put this into perspective, a comparison of receiver operating characteristic (ROC) curves between the BT model and Muhammad et al. (2019)'s ANN model is shown in **Figure 6**.

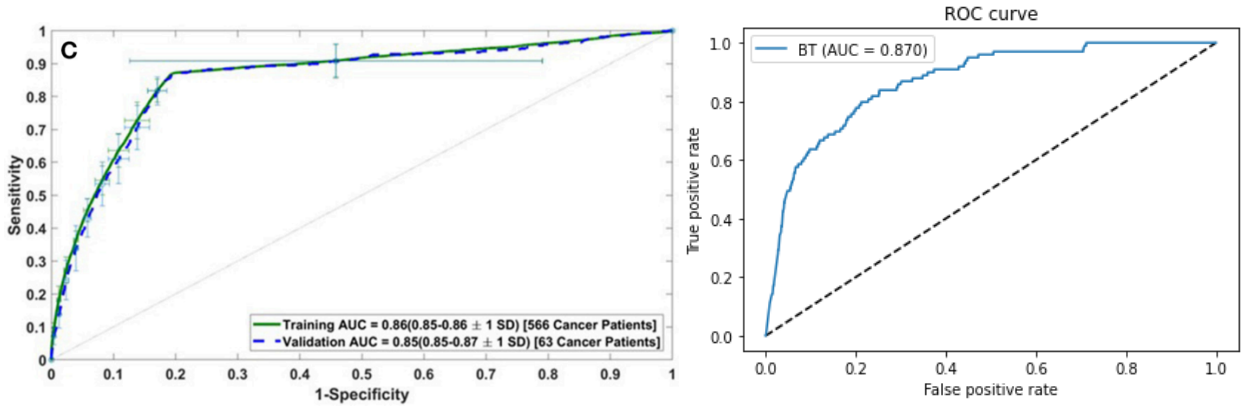

Figure 6: Comparison of ROC curves (area under the ROC curve = AUC). Left: Muhammad et al. (2019)'s ANN model. Right: BT model from this study.

This improvement suggests that the dataset, algorithms, and model hyperparameters used in this study were more effective for this task than those used in previous studies. Muhammad et al. (2019) created an ANN model using similar versions of the IPUMS and PLCO datasets consisting of 800,114 participants with 898 pancreatic cancer cases. Although the present study only included 752,527 participants, 983 of them had pancreatic cancer, making the ratio of pancreatic cancer to healthy cases higher than that of Muhammad et al. (2019)'s study (0.13% > 0.11%).

Although machine learning models generally yield better results with larger datasets (Barbedo, 2018), perhaps the greater number of pancreatic cancer cases that the model could "learn" from more than compensated for the ~50,000 fewer total participants compared to Muhammad et al. (2019)'s ANN. The increase in pancreatic cancer cases may also explain why the addition of the PLCO dataset improved AUC in both the present study and in Muhammad et al. (2019)'s study by 0.021 and 0.140, respectively.

The number and relevance of variables used to train a model may also play a role. Hall & Smith (1998) stresses the importance of not just including variables which are relevant but also eliminating ones which are extraneous in order to achieve the best performance. Hsieh et al. (2018) created ANN and LR models using the Longitudinal Cohort of Diabetes Patients dataset from the National Health Insurance, consisting of 1,358,634 participants with 3,092 pancreatic cancer cases. Though this is the largest dataset with the highest percentage of pancreatic cancer cases, most of the variables only applied to diabetes patients and many of the most common pancreatic cancer risk factors were not included. This may explain the weak performance of their ANN and LR models in terms of AUC.

Moreover, in the present study, the tree-based algorithms (DT, RF, and BT) had higher AUCs than that of the LR and SVM algorithms. As mentioned by Uddin et al. (2019), tree-based algorithms and ANNs tend to perform well for large datasets consisting of multiple data types (i.e., continuous & categorical) with high-complexity relationships between variables, whereas LRs and SVMs are better for lower-complexity problems. Since the DT, RF, BT, and Muhammad et al. (2019)'s ANN had higher AUCs than all of the current and previous LRs/SVMs, perhaps predicting pancreatic cancer is a high-complexity problem for which tree-based models and ANNs are better suited for.

Impact of Depression on Model Performance

One of the reasons for determining variable importance of the final BT model was to investigate the effects of including depression on model performance. Interesting to note is that the permutation (random shuffling) of the depression variable (*Functionally Limiting Depression*) actually improved AUC, indicating that while depression is a risk factor of pancreatic cancer, it had a negative effect on model performance.

There are a few reasons that may explain this. Adding additional variables generally results in better model performance; however, it is more important that those variables are relevant, clean, and reliable (Hall & Smith, 1998). The depression variable is relevant since it is known to be a predictor of pancreatic cancer (Kenner, 2018; Olson et al., 2016); however, this variable was unreliable in the form of missing values (*Functionally Limiting Depression* was not included in the PLCO dataset, accounting for a total of 20.0% missing values). This resulted in only 21.9% of the total pancreatic cancer cases (215/983) available for the model to learn the relationship of depression from, thereby decreasing the model's ability to discriminate between pancreatic cancer and healthy cases, which may explain the decrease in AUC.

Before the present study, I conducted a study aimed at improving upon Muhammad et al. (2019) and Hsieh et al. (2018)'s models in terms of sensitivity, specificity, and AUC values. Thirty variables were used from the IPUMS Health Surveys dataset encompassing 602,558 anonymized participants to create an ANN and LR model. Though these models failed to improve upon the previously published models (hence why the present study aimed to do so with additional data, algorithms, and refined variables), I found that the inclusion of three depression variables (two of which had > 75% missing values) decreased AUC as well. As such, it would be beneficial to include depression variables with significantly fewer missing values in order to potentially improve performance (Hall & Smith, 1998).

Yet a closer look at the dataset in the present study reveals that only 8 of the 215 patients with pancreatic cancer reported some form of depression, indicating that it may only occur in a subset of cases. In addition, this variable does not distinguish between those with a sudden onset of depression and those with long-term depression. Moreover, *Moderate Activity Times Per Week* and *Strenuous Activity Times*

Per Week each had the 2nd and 3rd highest percentage of missing values out of all variables (8.5% and 7.8%, respectively) and still were the 2nd and 3rd most important variables, respectively. Given this information, depression may simply not be a useful predictor of pancreatic cancer.

Variability in Model Sensitivities Tested on the PanCAN Dataset

As indicated in the results, the sensitivities of the DT, RF, and BT models tested on the PanCAN dataset contained a notably high variability while every trial of the LR and SVM models yielded perfect sensitivity (1.000). As such, the LR and SVM models may be of interest even though the BT had the highest AUC. However, as can be seen in the **Appendix**, the PanCAN dataset contained only 6 of 17 variables that were used to train the models, and all of which except *Age*, *Sex*, and *Race* contained over 85% missing values. As a result, the abundance of missing data from the PanCAN dataset could have resulted in inaccuracies when testing the models (Hall & Smith, 1998). Coupled with that, three of the six variables available in the PanCAN dataset (*Smoke 100*, *Race*, and *Diabetes*) were identified as important variables for predicting pancreatic cancer, so their limited presence due to missing values in the PanCAN dataset could have impacted performance as well (Hall & Smith, 1998). The fact that the LR and SVM models achieved a perfect sensitivity score for every trial is unusual and further indicates that the missing values and small number of variables may lead to misleading results. Although it was of great interest to test model performance on an additional dataset, the poor/inconsistent performance may just reflect the need for more complete data.

A Delicate Balance Between True Positives and True Negatives

In this problem, it can be more important to identify true positives (sensitivity) than to identify true negatives (specificity) since that could make the difference between life and death. If the model were to predict that a patient would not develop pancreatic cancer, but they do end up getting it, then the patient could die. However, it is also important that the model does not yield too many false positives in order to prevent unnecessary panicking and screening.

Figure 7 displays confusion matrices (diagrams representing true/false positives/negatives) of the final BT model compared to the highest-sensitivity model, the SVM. As shown in the confusion matrices, the BT missed 21 pancreatic cancer cases while the SVM only missed 14. However, the SVM incorrectly classified 11,286 more patients as having pancreatic cancer than the BT. Though the BT missed 7 more pancreatic cancer cases than the SVM, one could argue that the SVM's 11,286 additional false positives could create unnecessary anxiety and expose patients to the potentially detrimental effects of screening, outweighing the benefit of identifying more true positives. Yet at the same time, identifying

as many of these true pancreatic cancer patients as possible is crucial. As such, maximizing both sensitivity and specificity is key if a model was to be deployed in the real world.

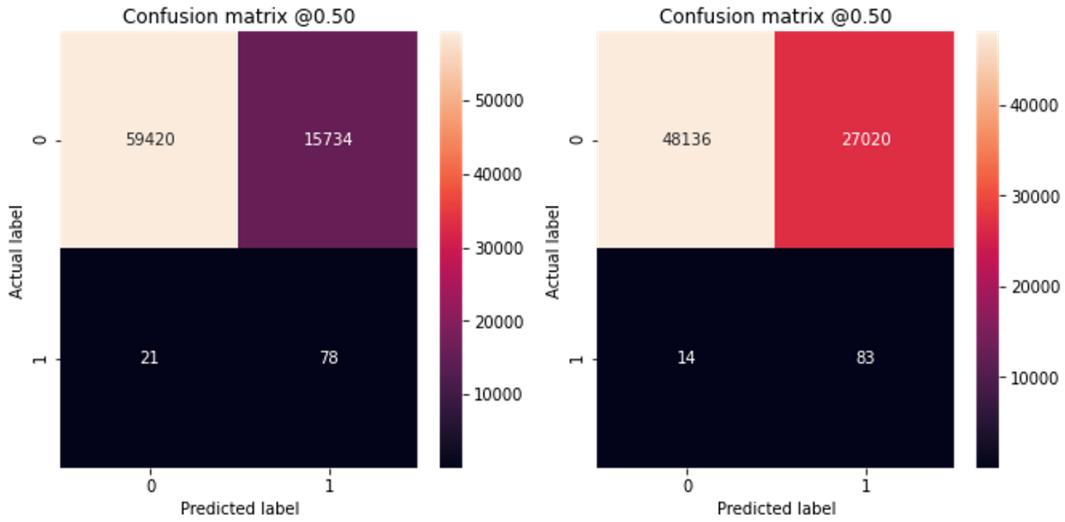

Figure 7: Confusion matrices of the final BT (left) and SVM (right) models. The confusion matrices are applied with the standard cutoff of 0.5 for discriminating between pancreatic cancer and healthy cases. '0' represents a negative case and '1' represents a positive case.

Variable Importance

Another reason for determining variable importance in this study was to inform users of the confidence of the model's predictions. While the models were trained on variables that most patients would know the answer to, there may still be instances in which a certain variable does not apply to a patient and/or a patient does not know the answer. For example, if the user fails to input *Age*, *Moderate Activity Times Per Week*, *Strenuous Activity Times Per Week,* or *Smoke 100* data, they should be made aware that predictions could be slightly less accurate. However, if the user fails to input *Hypertension, Coronary HD or Attack*, or *Emphysema* data, they should be made aware that it will likely not impact the accuracy of predictions. The app alerts the user if they are missing data that may cause an impact in this regard.

Interesting to note is that among the top five most influential variables on the BT model, age, smoking, and race remain consistent with high levels of importance as reported by the American Cancer Society's collection of pancreatic cancer risk factors ("Pancreatic Cancer Risk Factors," n.d.). In addition, although physical activity is shown as important in this study, the American Cancer Society notes that it has an "unclear effect" on one's risk ("Pancreatic Cancer Risk Factors," n.d.).

Limitations and Future Research

One significant limitation in this study was the lack of pancreatic cancer cases in the training dataset. While this study did employ methods to up-sample the pancreatic cancer cases and included additional data to help further balance the cases, it would be best to have a dataset which naturally contains more balanced cases to increase generalization and therefore model performance (Boyle, 2019). In addition, the depression variable did not exist in the PLCO dataset, resulting in a significant percentage (20.0%) of missing values in the combined dataset. As such, finding more depression data may help determine the impact of depression on model performance and perhaps increase the overall performance of the model as well. Furthermore, having a dataset like PanCAN to test the models on additional cases, except with a more complete set of data, would help provide further insight into model performance.

Though this study presents the five models as effective tools in identifying those at risk for pancreatic cancer, the inclusion of an ANN, which has also been shown to be effective for disease prediction, should be considered in order to potentially achieve better performance (Muhammad et al., 2019; Nartowt et al., 2020; Roffman et al., 2018a, Roffman et al., 2018b). Improvements could also be made to the resampling of the data, such as evaluating a popular up-sampling method, SMOTE, that was not explored in this study (Brownlee, 2020). Lastly, before deploying the app into the real world, testing should be done to ensure that it is a viable method to include in modern medical workflows and should be improved/modified as necessary.

Conclusions

This study investigated the use of machine learning algorithms to identify one's risk of developing pancreatic cancer. Five algorithms — Decision Tree, Random Forest, Boosted Trees, Logistic Regression, and Support Vector Machine — are shown to be effective in using machine learning to discriminate between patients with and without pancreatic cancer, achieving the highest AUC of 0.870 (Boosted Trees model). The addition of the PLCO data to the IPUMS data, which increased sample size by 24.9% and more than quadrupled the number of pancreatic cancer cases, improved mean AUC by 0.021. The testing of the models on the PanCAN dataset consisting of pancreatic cancer cases supported the effectiveness of the final models; however, a large variability in sensitivity between trials limits evidence. In addition, the inclusion of depression did not improve performance, whereas risk factors such as age, physical activity, smoking, and race were the most influential. Finally, an iPhone app was created to facilitate data input and prediction output by an end-user. Once the models and app are fully optimized, they could become valuable tools assisting in the early diagnosis of pancreatic cancer by identifying high-risk candidates for further medical screening.

Acknowledgements

I would like to sincerely thank Dr. Carol Hersh for supporting me throughout my research virtually during the COVID-19 pandemic. I would also like to thank Benjamin Newman, graduate of Great Neck South High School and current computer science student at Stanford University for his occasional support at the beginning of this project.

I also thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The statements contained herein are solely mine and do not represent or imply concurrence or endorsement by NCI. I would also like to thank the Pancreatic Cancer Action Network (PanCAN) for the dozens of emails back-and-forth and support from the team in order to deliver the data used in this study.

Appendix

Table 1: Variables from the IPUMS, PLCO, and PanCAN datasets used to train and/or test the models.

Variable	Description	Values	IPUMS	PLCO	IPUMS + PLCO	PanCAN
Age	Age of person	18 - 85	Mean: 48.13 SD: 18.15	Mean: 62.64 SD: 5.37	**Mean: 51.02 SD: 17.41**	Mean: 60.41 SD: 10.77
Sex	Biological sex of person	Male	44.2%	49.2%	**45.2%**	52.0%
		Female	55.8%	50.8%	**54.8%**	48.0%
Race	Race of person	White	74.1%	88.4%	**77.0%**	88.5%
		Black/African American	14.4%	5.1%	**12.5%**	2.0%
		Hispanic	N/A	1.9%	**0.4%**	2.5%
		Asian	2.9%	3.7%	**3.1%**	2.9%
		Pacific Islander	N/A	0.6%	**0.1%**	0.1%
		American Indian	0.9%	0.3%	**0.8%**	0.3%
		Multi-racial	0.3%	N/A	**0.3%**	2.5%
		Other/Missing	7.3%	< 0.05%	**5.9%**	1.3%
BMI	Body Mass Index of person	6.6 - 99.8	Mean: 27.48 SD: 6.09	Mean: 27.28 SD: 4.92	**Mean: 27.44 SD: 5.87**	N/A
		Missing	4.0%	1.5%	**3.5%**	
Asthma	Person ever told by a doctor they had asthma	No	88.2%	58.9%	**82.4%**	N/A
		Yes	11.7%	6.2%	**10.6%**	
		Missing/Unknown	0.1%	34.9%	**7.0%**	

Coronary HD or Attack	Person ever told by a doctor they had coronary heart disease or a heart attack	No	93.7%	90.3%	**93.0%**	N/A
		Yes	6.1%	9.0%	**6.7%**	
		Missing/Unknown	0.2%	0.7%	**0.3%**	
Diabetes	Person ever told by a doctor they had diabetes	No	89.5%	91.7%	**89.9%**	3.5%
		Yes	10.5%	7.7%	**9.9%**	9.2%
		Missing/Unknown	< 0.05%	0.6%	**0.2%**	87.3%
Emphysema	Person ever told by a doctor they had emphysema	No	98.1%	96.8%	**97.9%**	N/A
		Yes	1.8%	2.5%	**2.0%**	
		Missing	0.1%	0.6%	**0.2%**	
Hepatitis or Cirrhosis	Person ever told by a doctor they had hepatitis or cirrhosis	No	90.2%	95.5%	**91.2%**	N/A
		Yes	3.0%	3.7%	**3.1%**	
		Missing/Unknown	6.9%	0.8%	**5.7%**	
Hypertension	Person ever told by a doctor they had hypertension	No	69.5%	65.4%	**68.8%**	N/A
		Yes	30.4%	34.0%	**31.2%**	
		Missing/Unknown	0.1%	0.6%	**0.2%**	
Stroke	Person ever told by a doctor they had a stroke	No	96.9%	96.9%	**97.1%**	N/A
		Yes	3.0%	2.4%	**2.9%**	
		Missing/Unknown	0.1%	0.6%	**0.2%**	
Alcohol Days Past Year	The number of days in which a person had at least one alcoholic drink in the past year	0 - 365	Mean: 47.56 SD: 88.85	Mean: 119.41 SD: 143.35	**Mean: 59.31 SD: 103.30**	N/A
		Missing/Unknown	1.0%	22.2%	**5.2%**	
Smoke 100	Person smoked 100+ cigarettes in the past year	No	57.7%	31.1%	**56.5%**	5.9%
		Yes	41.6%	35.8%	**43.5%**	7.4%
		Missing/Unknown	0.7%	33.1%	**7.2%**	86.7%
Smoke Years	The number of years a person has smoked in his/her lifetime	0 - 85	Mean: 9.77 SD: 15.18	Mean: 14.74 SD: 17.10	**Mean: 10.77 SD: 15.71**	N/A
		Missing/Unknown	1.6%	1.1%	**1.5%**	
Moderate Activity Times Per Week	The amount of times a person has engaged in moderate activity (i.e., a walk) per week over the past year	0 - 28	Mean: 2.61 SD: 3.76	Mean: 2.54 SD: 2.14	**Mean: 2.60 SD: 3.57**	N/A
		Missing/Unknown	2.1%	34.3%	**8.5%**	
		Missing/Unknown	1.3%	33.9%	**7.8%**	

Strenuous Activity Times Per Week	The amount of times a person has engaged in strenuous activity (i.e., a run) per week over the past year	0 - 28	Mean: 1.59 SD: 3.02	Mean: 1.82 SD: 2.09	**Mean: 1.63 SD: 2.91**	N/A
		Missing/Unknown	1.3%	33.9%	**7.8%**	
Functionally Limiting Depression	How long the person has had a depression, anxiety, or emotional problem that limits everyday activities	None	97.5%	N/A	**78.0%**	5.6%
		< 3 months	< 0.05%		**< 0.05%**	3.3%
		3-5 months	< 0.05%		**< 0.05%**	3.0%
		6-12 months	0.2%		**0.1%**	0.5%
		> 12 months	2.3%		**1.8%**	0.4%
		Missing/Unknown	< 0.05%		**20.0%**	87.2%
PC (Label)	Person ever told by a doctor they had pancreatic cancer	No	100.0%	99.5%	**99.9%**	0.0%
		Yes	< 0.05%	0.5%	**0.1%**	100.0%

Value counts of categorical variables are represented as percentages while continuous variables are represented with the mean and standard deviation (SD). "Missing/Unknown" values include both values missing in the original dataset as well as unknown answers re-coded as missing. "N/A" values indicate that a variable or one or more of its codes do not exist in its respective dataset. Fully re-coded variable names, descriptions, and values are displayed in this table and thus should not be directly compared to those of the raw IPUMS, PLCO, and PanCAN datasets.

Bibliography

Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*. https://doi.org/10.1109/icdse.2016.7823957

App Store Review Guidelines. Apple Developer. https://developer.apple.com/app-store/review/guidelines/.

Arnold, L. D., Patel, A. V., Yan, Y., Jacobs, E. J., Thun, M. J., Calle, E. E., & Colditz, G. A. (2009). Are Racial Disparities in Pancreatic Cancer Explained by Smoking and Overweight/Obesity? *Cancer Epidemiology Biomarkers & Prevention*, *18*(9), 2397–2405. https://doi.org/10.1158/1055-9965.epi-09-0080

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, *153*, 46–53. https://doi.org/10.1016/j.compag.2018.08.013

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King and Kari C.W. Williams. *IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]*. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D070.V6.4*

Boyle, T. (2019, February 3). *Dealing with Imbalanced Data*. Medium. https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18.

Brownlee, J. (2020, August 21). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. *Machine Learning and Knowledge Discovery in Databases*, 655–670. https://doi.org/10.1007/978-3-030-10925-7_40

Classification: ROC Curve and AUC. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Core ML. Apple Developer Documentation. https://developer.apple.com/documentation/coreml.

Coughlin, S. S., Calle, E. E., Patel, A. V., & Thun, M. J. (2000). Predictors of pancreatic cancer mortality among a large cohort of United States adults. *Cancer Causes & Control*, *11*, 915–923. https://doi.org/10.1023/a:1026580131793

Create ML. Apple Developer Documentation. https://developer.apple.com/documentation/createml.

Everhart, J., & Wright, D. (1995). Diabetes Mellitus as a Risk Factor for Pancreatic Cancer. *Jama*, *273*(20), 1605–1609. https://doi.org/10.1001/jama.1995.03520440059037

Gomez-Rubio, P., Zock, J.-P., Sharp, L., Hidalgo, M., Carrato, A., Ilzarbe, L., … Malats, N. (2015). Reduced risk of pancreatic cancer associated with asthma and nasal allergies. *Gut*. https://doi.org/10.1016/j.pan.2015.05.439

Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning. In *Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98* (pp. 181–191). Berlin: Springer.

Hassan, M. M., Li, D., El-Deeb, A. S., Wolff, R. A., Bondy, M. L., Davila, M., & Abbruzzese, J. L. (2008). Association Between Hepatitis B Virus and Pancreatic Cancer. *Journal of Clinical Oncology*, *26*(28), 4557–4562. https://doi.org/10.1200/jco.2008.17.3526

Hsieh, M. H., Sun, L.-M., Lin, C.-L., Hsieh, M.-J., Hsu, C.-Y., & Kao, C.-H. (2018). Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer Management and Research*, *10*, 6317–6324. https://doi.org/10.2147/cmar.s180791

Human Interface Guidelines. Apple Developer. https://developer.apple.com/design/human-interface-guidelines/.

Kenner, B. J. (2018). Early Detection of Pancreatic Cancer. *Pancreas*, *47*(4), 363–367. https://doi.org/10.1097/mpa.0000000000001024

Larsson, S. C., Orsini, N., & Wolk, A. (2007). Body mass index and pancreatic cancer risk: A meta-analysis of prospective studies. *International Journal of Cancer*, *120*(9), 1993–1998. https://doi.org/10.1002/ijc.22535

Lindgren, A. M., Nissinen, A. M., Tuomilehto, J. O., & Pukkala, E. (2005). Cancer pattern among hypertensive patients in North Karelia, Finland. *Journal of Human Hypertension*, *19*(5), 373–379. https://doi.org/10.1038/sj.jhh.1001834

Lowenfels, A. B., & Maisonneuve, P. (2006). Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology*, *20*(2), 197–209. https://doi.org/10.1016/j.bpg.2005.10.001

McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., & Mccain, R. S. (2018). Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World Journal of Gastroenterology*, *24*(43), 4846–4861. https://doi.org/10.3748/wjg.v24.i43.4846

Michaud, D. S., Giovannucci, E., Willett, W. C., Colditz, G. A., Stampfer, M. J., & Fuchs, C. S. (2001). Physical Activity, Obesity, Height, and the Risk of Pancreatic Cancer. *Jama*, *286*(8), 921–929. https://doi.org/10.1001/jama.286.8.921

Muhammad, W., Hart, G. R., Nartowt, B., Farrell, J. J., Johung, K., Liang, Y., & Deng, J. (2019). Pancreatic Cancer Prediction Through an Artificial Neural Network. *Frontiers in Artificial Intelligence*, *2*. https://doi.org/10.3389/frai.2019.00002

Nartowt, B. J., Hart, G. R., Muhammad, W., Liang, Y., Stark, G. F., & Deng, J. (2020). Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. *Frontiers in Big Data*, *3*. https://doi.org/10.3389/fdata.2020.00006

Olson, S. H., Xu, Y., Herzog, K., Saldia, A., DeFilippis, E. M., Li, P., … Kurtz, R. C. (2016). Weight Loss, Diabetes, Fatigue, and Depression Preceding Pancreatic Cancer. *Pancreas*, *45*(7), 986–991. https://doi.org/10.1097/MPA.0000000000000590

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. *International Journal of Computer Science and Network Security*, *8*(8). https://doi.org/10.1109/aiccsa.2008.4493524

Pancreatic Cancer Risk Factors. American Cancer Society. https://www.cancer.org/cancer/pancreatic-cancer/causes-risks-prevention/risk-factors.html.

PLCO. The Cancer Data Access System. https://cdas.cancer.gov/plco/.

Poruk, K. E., Firpo, M. A., Adler, D. G., & Mulvihill, S. J. (2013). Screening for Pancreatic Cancer. *Annals of Surgery*, *257*(1), 17–26. https://doi.org/10.1097/sla.0b013e31825ffbfb

Roffman, D. A., Hart, G. R., Leapman, M. S., Yu, J. B., Guo, F. L., Ali, I., & Deng, J. (2018). Development and Validation of a Multiparameterized Artificial Neural Network for Prostate

Cancer Risk Prediction and Stratification. *JCO Clinical Cancer Informatics*, *2*, 1–10. https://doi.org/10.1200/cci.17.00119

Roffman, D., Hart, G., Girardi, M., Ko, C. J., & Deng, J. (2018). Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-19907-9

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

SwiftUI. Apple Developer Documentation. https://developer.apple.com/documentation/swiftui/.

Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00307

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*. https://doi.org/10.1186/s12911-019-1004-8

Ye, W., Lagergren, J., Weiderpass, E., Nyren, O., Adami, H.-O., & Ekbom, A. (2002). Alcohol abuse and the risk of pancreatic cancer. *Gut*, *51*(2), 236–239. https://doi.org/10.1136/gut.51.2.236

Zheng, W., Mclaughlin, J. K., Gridley, G., Bjelke, E., Schuman, L. M., Silverman, D. T., … Fraumeni, J. F. (1993). A cohort study of smoking, alcohol consumption, and dietary factors for pancreatic cancer (United States). *Cancer Causes & Control*, *4*(5), 477–482. https://doi.org/10.1007/bf00050867

*Blewett et. al requires their citation to appear as such. See: https://nhis.ipums.org/nhis/citation.shtml