Quality in Primary Care
ISSN: 1479-1064

Open access Research Article

Using Machine Learning to Identify the Risk Factors of Pancreatic Cancer from the NCI PLCO Dataset

Ananya Dutta*

Department of Electrical and Communications Engineering, Gauhati University, India

ABSTRACT

Background: Pancreatic cancer (PC) is a disease with poor prognosis and survival rate. There is a pertinent need to identify the risk factors of this disease. The purpose of this study is to identify a subset of factors (a.k.a. features) as predictors of PC from the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer dataset consisting of responses to 65 questions about demographics, cancer and health history, medication usage, and smoking habits from 154,897 participants.

Method: There are two challenges to selecting the subset of features that predict PC with highest probability: The problem is computationally intractable, and the PLCO dataset is highly imbalanced. We use an innovative method to use the dataset in a balanced way, without involving up or down-sampling. We use nine feature selection methods to select the optimal subset of features from the preprocessed and balanced dataset.

Results: Our preprocessed dataset consists of 32 risk factors (8 demographics, 5 cancer history, 13 health histories, 2 medication usage, 4 smoking habits). Risk factors belonging to cancer and health history, followed by smoking habits, were consistently chosen by the feature selection methods. We also discuss findings in the medical sciences literature that corroborate our findings.

Conclusions: The study found that risk factors belonging to cancer and health history are the most prominent ones for PC. In particular, previously diagnosed with PC is chosen as the most prominent risk factor by majority of methods. While most of our findings are consistent with the literature, some of our findings shed light on novel factors that may not have received their due attention by the research community.

Keywords: Pancreatic cancer; NCI PLCO dataset; Feature selection; Classification

INTRODUCTION

Pancreatic Cancer (PC) is a disease with poor prognosis and survival rate. About 95% of people who contract PC would not make it to the 5-year survival period [1]. Pancreas is an inner organ of the human body, surrounded by the duodenum and the small intestine; hence early symptoms are hard to detect [2]. Malicious cells in the pancreas are typically detected at a very advanced stage when it is impossible to save the patient. There is a pertinent need for a prediction model that can lead to early detection of this disease.

Biomarkers for early diagnosis of PC have been investigated (see for example, [3-8]). However, evidence for identified biomarkers has not been very conclusive. Image analysis and machine learning algorithms have been used for distinguishing between benign and malignant tissues in endoscopic ultrasound and computed tomography images (see for example, [9-12]). However, these models can detect PC only at an advanced stage and hence are not very useful.

The purpose of this study is to identify a set of factors as predictors of PC. We use a cancer dataset collected from 154,897 participants, each responding to 65 questions (or factors) about demographics, cancer and health history, medication usage, and smoking habits. There are two challenges to selecting the

Received:	29-August-2023	**Manuscript No:**	IPQPC-23-17722
Editor assigned:	31-August-2023	**PreQC No:**	IPQPC-23-17722 (PQ)
Reviewed:	14-September-2023	**QC No:**	IPQPC-23-17722
Revised:	19-September-2023	**Manuscript No:**	IPQPC-23-17722 (R)
Published:	26-September-2023	**DOI:**	10.36648/1479-1064.23.31.34

Corresponding author Ananya Dutta, Department of Electrical and Communications Engineering, Gauhati University, India Email: adutta@memphis.edu

Citation Dutta A (2023) Using Machine Learning to Identify the Risk Factors of Pancreatic Cancer from the NCI PLCO Dataset. Qual Prim Care. 31:34.

subset of 65 factors that predict PC with highest probability: The problem is computationally intractable, and the dataset is highly imbalanced. Our approach consists of balancing and pre-processing the dataset, and rank the risk factors based on their ability to predict PC.

Our study found that risk factors belonging to cancer and health history are the most prominent ones for PC. In particular, previously diagnosed with PC is chosen as the most prominent risk factor by majority of methods.

We also discuss findings in the medical sciences literature that corroborate our findings. Some of our findings shed light on novel factors that may not have received their due attention by the research community.

MATERIALS AND METHODS

Problem Statement

Our problem is to predict whether a subject is diagnosed with PC or not, given information about his demographic characteristics, health history, medication usage, smoking habits, and his and his family's cancer diagnosis history. This information is encoded as a vector of predictor variables where each predictor represents a risk factor (a.k.a. feature). The predictors are discrete and finite random variables.

Formally, given a set of data points $X=[x_1,..., x_N] \in N^{d\times N}$, N={0, 1, 2,...}, and a set of labels {True, False}, the task is to map each data point $xi \in N^d$ into one of the labels, where d is the dimension of each data point, and N is the number of data points in the dataset. This is a binary classification problem. Our goal is to select a subset of predictors such that classification using the subset is at least as accurate as that using the entire set. It has been shown that the accuracy does not always improve with increase in number of variables [13], hence choosing the optimal subset of predictors is imperative for accurate prediction of PC.

MATERIALS

The Prostate, Lung, Colorectal and Ovarian (PLCO) cancer dataset [14] is collected by the National Cancer Institute from 154,897 participants. Among them, 76,678 or 49.5% were males, and 132,572 or 85.6% were non-Hispanic White. The participants, randomly selected based on a set of criteria from different parts of United States, were between 55-74 years of age with no history of prostate, lung, colorectal or ovarian cancer. Each participant filled out three questionnaires, thereby responding to 65 questions about demographics, cancer and health history, medication usage, and smoking habits. Therefore, N=154, 897 and d=65 for our problem. The dataset is highly imbalanced; only 749 or 0.48% of the participants were diagnosed with pancreatic cancer (Table 1). Visualizations of the PLCO dataset in two-dimensional (2D) space are shown in Figure 1 lists the 65 risk factors.

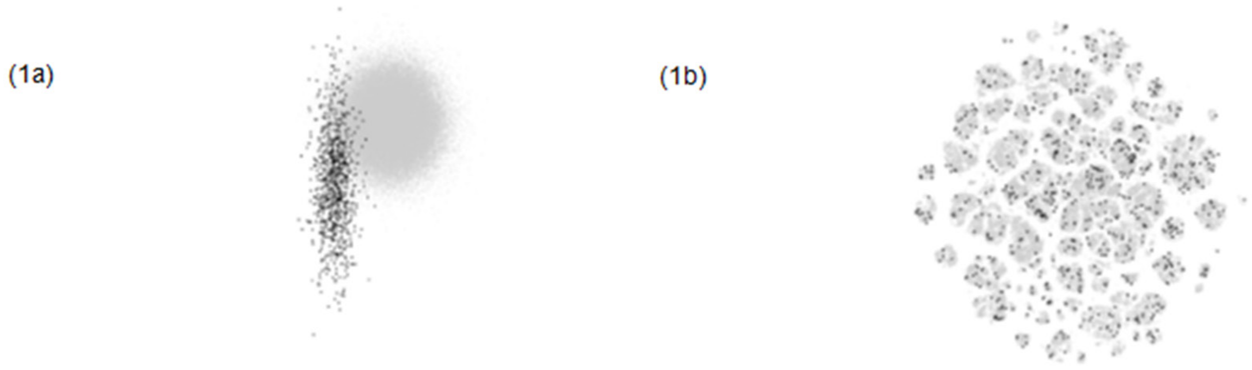


Figure 1: PLCO dataset visualized in 2D using (a) ADASYN algorithm [15] and (b) t-SNE algorithm [16]. Data points corresponding to PC=True and PC=False are shown in black and gray respectively.

Table 1: The risk factors considered in the PLCO dataset. The ones marked "removed" are not considered in our analysis as there are not enough responses from the participants on these questions

Risk factor categories	Risk factors (values)	Male risk factors (total 47, removed 15)	Female risk factors (total 52, removed 20)
Cancer history	59 participants Has relative with cancer (yes, no)	✓	✓
	60 participants Has relative with PC (yes, no)	✓	✓
	61 participants No. of relatives with PC (0,1,2,3,...)	✓	✓
	62 participants Diagnosed with any cancer (yes, no)	✓	✓
	63 participants Diagnosed with PC (yes, no)	✓	✓

Demographics	64 participants Gender (male, female)	✓	✓
	38 participants Race (White, Black, Asian, Pacific Islander, American Indian/Alaskan Native)	✓	✓
	39 participants Hispanic origin (yes, no)	✓	✓
	1 participant Education level completed (<8 yrs, 8-11 yrs, 12 yrs, 12 yrs+ some college, college grad, post grad)	✓	✓
	2 participants Marital status (married, widowed, divorced, separated, never married)	✓	✓
	3 participants Occupation (homemaker, working, unemployed, retired, extended sick leave, disabled, other)	✓	✓
	6 participants No. of sisters (0,1,2,3,4,5,6, ≥ 7)	✓	✓
	7 participants No. of brothers (0, 1, 2, 3, 4, 5,6, ≥ 7)	✓	✓
Medication usage	8 participants Used aspirin regularly (yes, no)	✓	✓
	9 participants Used ibupr ofen regularly (yes, no)	✓	✓
	52 participants Taken birth control pills (yes, no)		✓ (removed)
	20 participant Age started taking birth control pills (<30 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, ≥ 60 yrs)		✓ (removed)
	21 participants Currently taking female hormones (yes, no)		✓ (removed)
	22 participants No. of years taking female hormones (≤ 1, 2-3, 4-5, 6-9, ≥ 10)		✓ (removed)
	53 participants Taken female hormones (yes, no, don't know)		✓ (removed)
Health history	27 participants Had high blood pressure (yes, no)	✓	✓
	28 participants Had heart attack (yes, no)	✓	✓
	29 participants Had stroke (yes, no)	✓	✓
	30 participants Had emphysema (yes, no)	✓	✓
	31 participants Had bronchitis (yes, no)	✓	✓
	32 participants Had diabetes (yes, no)	✓	✓
	33 participants Had colorectal polyps (yes, no)	✓	✓
	34 participants Had arthritis (yes, no)	✓	✓
	35 participants Had osteoporosis (yes, no)	✓	✓
	36 participants Had diverculitis (yes, no)	✓	✓
	37 participants Had gall bladder inflammation (yes, no)	✓	✓
	57 participants Had colon comorbidity (yes, no)	✓	✓
	58 participants Had liver comorbidity (yes, no)	✓	✓
	40 participants Had biopsy of prostrate (yes, no)	✓ (removed)	
	41 participants Had transurethral resection of prostate (yes, no)	✓ (removed)	
	42 participants Had prostatetomy of benign disease (yes, no)	✓ (removed)	
	43 participants Had prostate surgery (yes, no)	✓ (removed)	
	47 participants Had enlarged prostate (yes, no)	✓ (removed)	

	48 participants Had inflamed prostate (yes, no)	✓ (removed)	
	49 participants Had prostate problem (yes, no)	✓ (removed)	
	50 participants No. of times wakes up to urinate at night (0,1,2,3, >3)	✓ (removed)	
	23 participants Age started to urinate more than once at night (<30 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, ≥ 70 yrs)	✓ (removed)	
	24 participants Age when told had enlarged prostate (<30 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, ≥ 70 yrs)	✓ (removed)	
	25 participants Age when told had inflammed prostate (<30 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, ≥ 70 yrs)	✓ (removed)	
	26 participants Age at vasectomy (<25 yrs, 25-34 yrs, 35-44 yrs, ≥ 45 yrs)	✓ (removed)	
	51 participants Had vasectomy (yes, no)	✓ (removed)	
	44 participants Been pregnant (yes, no, don’t know)		✓ (removed)
	45 participants Had hysterectomy (yes, no)		✓ (removed)
	46 participants Had ovaries removed (yes, no)		✓ (removed)
	10 participants No. of tubal pregnancies (0, 1, ≥ 2)		✓ (removed)
	11 participants Had tubal ligation (yes, no, don’t know)		✓ (removed)
	12 participants Had benign ovarian tumor (yes, no)		✓ (removed)
	13 participants Had benign breast disease (yes, no)		✓ (removed)
	14 participants Had endometriosis (yes, no)		✓ (removed)
	15 participants Had uterine fibroid tumors (yes, no)		✓ (removed)
	16 participants Tried to become pregnant without success (yes, no)		✓ (removed)
	17 participants No. of pregnancies (0,1,2,3, 4-9, ≥ 10)		✓ (removed)
	18 participants No. of stillbirth pregnancies (0,1, ≥ 2)		✓ (removed)
	19 participants Age at hysterectomy (<40 yrs, 40-44 yrs, 45-49 yrs, 50-54 yrs, ≥ 55 yrs)		✓ (removed)
Smoking habits	4 participants Smoked pipe (never, currently, formerly)	✓	✓
	5 participants Smoked cigar (never, currently, formerly)	✓	✓
	54 participants Smoked cigarettes regularly (yes, no)	✓	✓
	55 participants Smoke regularly now (yes, no)	✓ (removed)	✓ (removed)
	56 participants Usually filtered or not filtered (filter more often, non-filter more often, both about equally)	✓ (removed)	✓ (removed)
	65 participants No. of cigarettes smoked daily (0, 1-10, 11-20, 21-30, 31-40, 41-60, 61-80, >80)	✓	✓

Dataset Balancing

A balanced dataset contains equal number of data points in all classes. Usually, an imbalanced dataset is balanced using methods such as fixed-rate downsampling or clustering that downsample the majority subset, or using methods such as the SMOTE algorithm [15-17] that upsample the minority subset. Both approaches inherit drawbacks unless the true distribution generating the data is known. The true distribution is unknown for the current problem.

We use a balancing method, similar to that proposed in [18], whereby the majority subset is iteratively and randomly subsampled such that in each iteration, the sampled subset is balanced. This method refrains from eliminating any data point from or introducing any new data point into the given dataset. A feature selection method is applied independently on each subset. The final result is obtained by computing the mean over all the subsets.

Data Preprocessing

The PLCO dataset has a number of missing values. We employ two steps iteratively to obtain a less incomplete dataset. First, we eliminate factors that are either missing responses from more than 10% of the participants, or responses from all participants are same. Next, we eliminate participants who did not respond to more than 10% of the remaining factors. The two steps are again applied to the resulting dataset. Application of the two steps continues until there is no change in the dataset between two consecutive iterations.

Each feature is standardized by subtracting its mean and dividing by its standard deviation. The missing values in the resulting dataset are filled in. The j^{th} element of the i^{th} data point, if missing, is filled by:

Equation 1:

$$\hat{x}_{ij} = \sum_{\substack{k=1 \\ k\neq i}}^{N} x_{kj}\, dist\left(x_i, x_k\right) / \sum_{\substack{k=1 \\ k\neq i}}^{N} dist\left(x_i, x_k\right)$$

Where $dist\left(x_i, x_k\right) = \dfrac{\left|x_i \ . \ x_k\right|}{\left\|x_i\right\| \ . \ \|x_k\|}$

$$x_i, x_k = \sum_{\substack{m=1 \\ m \text{ not missing}}}^{d} x_{im} x_{km},$$

$$\left\|x_i\right\| = \sqrt{\sum_{\substack{m= \\ m \text{ not missing}}}^{d} x_{im}}$$

|.| denotes the absolute value, “m not missing” refers to the mth element of a data point that is not missing, and dist is the absolute of the cosine similarity (or normalized dot product) of two data points. Therefore, 0 ≤ dist ≤ 1; as two data points get closer, their dist increases. In Eq. 1, a missing element of a given data point is computed as the weighted mean of that element from all data points in which values of all elements are present, and the weights are proportional to the absolute cosine similarity. After filling in all missing values, each feature is standardized again.

Variable or Feature Selection

Our problem of selecting the optimal subset of features is intractable as a total of $\sum_{n=1}^{d}(\frac{d}{n})$O ($2^d$) subsets are possible. Computing $O(2^d)$ subsets to determine the optimal one is impractical for the PLCO dataset with d=65. Hence we resort to variable or feature selection methods [19,20]. We used several feature selection algorithms suitable for categorical and continuous features and classification task [21-31], implemented in MATLAB, to rank the features, such as rank features using chi-square tests (‘fscchi2’ in MATLAB), rank features for classification using minimum redundancy maximum relevance (MRMR) algorithm (‘fscmrmr’ in MATLAB), estimate predictor importance for classification using ensemble of decision trees (‘fit-censemble’ in MATLAB), estimate predictor importance for classification using a binary decision tree (‘fitctree’ in MATLAB), estimate predictor importance for classification with an ensemble of bagged decision trees (e.g., random forest) which assigns positive and negative scores to the predictors (‘fitcensemble’ with method ‘bag’ and ‘oobPermutedPredictorImportance’ in MATLAB), rank key features by class separability criteria (‘rankfeatures’ with criteria ‘ttest,’ ‘entropy,’ ‘bhattacharyya,’ ‘roc,’ and ‘wilcoxon’ in MATLAB), and Pearson correlation between each feature/predictor variable and response variable (‘corrcoef’ in MATLAB) with correlation set to zero if not significant (i.e. p > 0.01). Figure 2 shows the ranking of the features by each of these algorithms for males and females respectively. A brief description of each of these algorithms is presented in Appendix.

RESULTS

Our analysis is done on the entire dataset as well as separately on the male and female participants. After Pre-processing (Ref: Data Processing), the PLCO dataset containing 65 features and 154897 points (76682 male, 749 True, 430 male True) reduces to 32 features and 148315 points (73162 male, 706 True, 405 male True). For balancing (Ref: Data Balancing), we randomly sample [148315/706]=210 non-overlapping subsets for PC=False, each containing 706 or 707 data points. Thus after balancing, each subset contains a total of 1412 or 1413 points. Similarly, for male only analysis, we obtain [73162/405]=180 balanced subsets, each containing 810 or 811 points. For female only analysis, we obtain [75153/301]=249 balanced subsets, each containing 602 or 603 points.

Classification

Some machine learning algorithms, briefly described in Section 6.4 were used and their statistical parameters are reported.

Using classification ensemble: In this ensemble algorithm, the weights or costs can be modified to correctly train the algorithm to predict PC. The weights are normalized to add unity, depicting the prior probabilities. Suppose $\in_{ij}$ (i, j ∈ {1...c},$\in_{ii}$=0) is the cost of misclassification of the example of the i^{th} class to the j^{th} class, where c is the number of classes. Then, the weight assigned to the i^{th} class after rescaling is given as [32]:

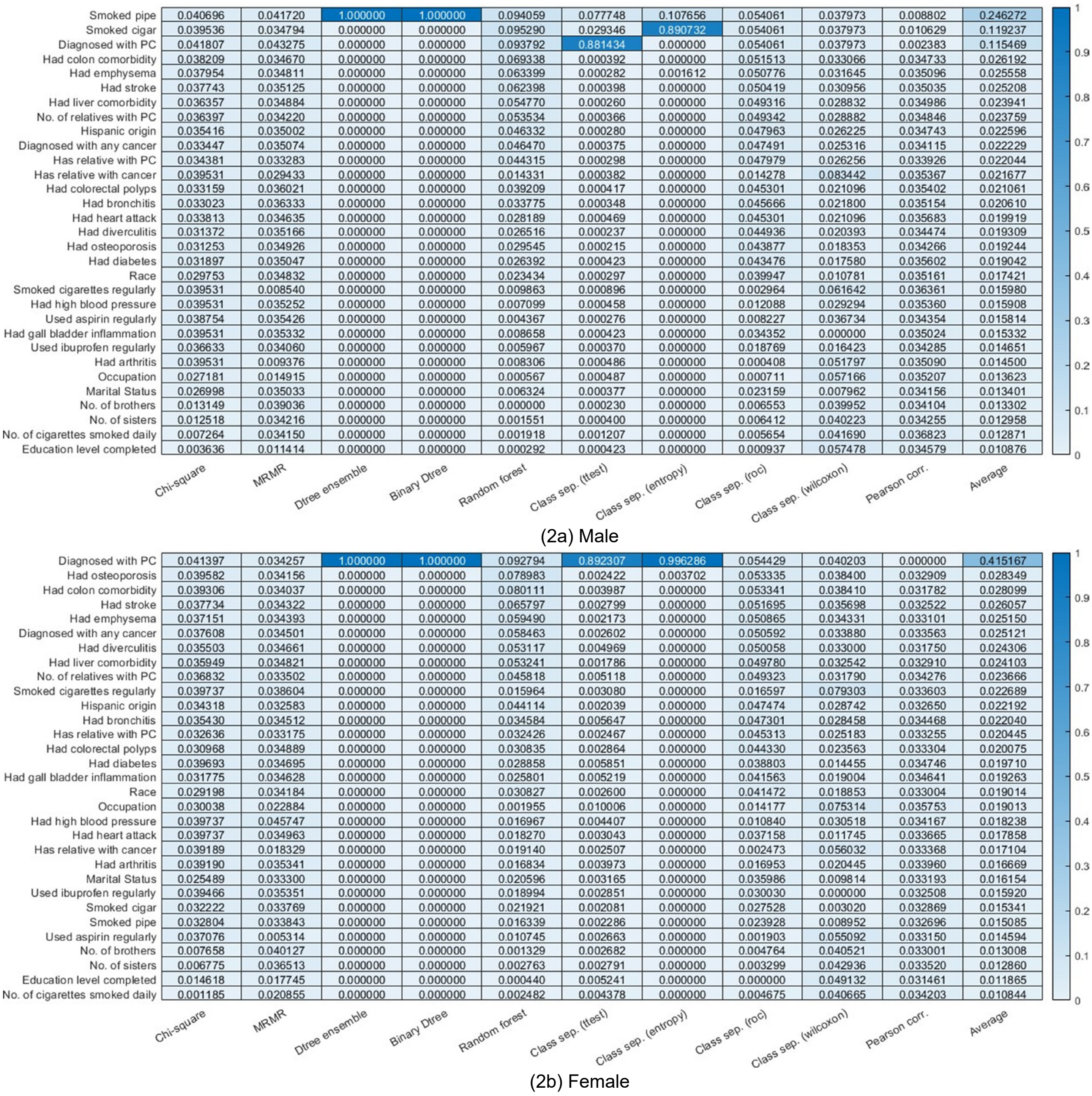
Equation 2:

$$wi = \frac{n X \in i}{\sum_{k=1}^{c} n_k X \in_k}$$

where n is the number of training samples, $\in_i = \sum_{j=1}^{c} \in_{ij}$

It uses the algorithms as described in [32-34]. For example, we can say the weight of predicting no PC for subjects with PC (False positive) is 1000 times more serious than predicting PC for subjects with no PC (False negative). Accordingly, we can change the weights to get a confusion matrix as per our need.

Feature selection: We used several feature selection algorithms [21-31], implemented in MATLAB, to rank the features. Figure 2 show the ranking of the features by each of these algorithms for males and females respectively.



	Chi-square	MRMR	Dtree ensemble	Binary Dtree	Random forest	Class sep. (ttest)	Class sep. (entropy)	Class sep. (roc)	Class sep. (wilcoxon)	Pearson corr.	Average
Smoked pipe	0.040696	0.041720	1.000000	1.000000	0.094059	0.077748	0.107656	0.054061	0.037973	0.008802	0.246272
Smoked cigar	0.039536	0.034794	0.000000	0.000000	0.095290	0.029346	0.890732	0.054061	0.037973	0.010629	0.119237
Diagnosed with PC	0.041807	0.043275	0.000000	0.000000	0.093792	0.881434	0.000000	0.054061	0.037973	0.002383	0.115469
Had colon comorbidity	0.038209	0.034670	0.000000	0.000000	0.069338	0.000392	0.000000	0.051513	0.033066	0.034733	0.026192
Had emphysema	0.037954	0.034811	0.000000	0.000000	0.063399	0.000282	0.001612	0.050776	0.031645	0.035096	0.025558
Had stroke	0.037743	0.035125	0.000000	0.000000	0.062398	0.000398	0.000000	0.050419	0.030956	0.035035	0.025208
Had liver comorbidity	0.036357	0.034884	0.000000	0.000000	0.054770	0.000260	0.000000	0.049316	0.028832	0.034986	0.023941
No. of relatives with PC	0.036397	0.034220	0.000000	0.000000	0.053534	0.000366	0.000000	0.049342	0.028882	0.034846	0.023759
Hispanic origin	0.035416	0.035002	0.000000	0.000000	0.046332	0.000280	0.000000	0.047963	0.026225	0.034743	0.022596
Diagnosed with any cancer	0.033447	0.035074	0.000000	0.000000	0.046470	0.000375	0.000000	0.047491	0.025316	0.034115	0.022229
Has relative with PC	0.034381	0.033283	0.000000	0.000000	0.044315	0.000298	0.000000	0.047979	0.026256	0.033926	0.022044
Has relative with cancer	0.039531	0.029433	0.000000	0.000000	0.014331	0.000382	0.000000	0.014278	0.083442	0.035367	0.021677
Had colorectal polyps	0.033159	0.036021	0.000000	0.000000	0.039209	0.000417	0.000000	0.045301	0.021096	0.035402	0.021061
Had bronchitis	0.033023	0.036333	0.000000	0.000000	0.033775	0.000348	0.000000	0.045666	0.021800	0.035154	0.020610
Had heart attack	0.033813	0.034635	0.000000	0.000000	0.028189	0.000469	0.000000	0.045301	0.021096	0.035683	0.019919
Had diverculitis	0.031372	0.035166	0.000000	0.000000	0.026516	0.000237	0.000000	0.044936	0.020393	0.034474	0.019309
Had osteoporosis	0.031253	0.034926	0.000000	0.000000	0.029545	0.000215	0.000000	0.043877	0.018353	0.034266	0.019244
Had diabetes	0.031897	0.035047	0.000000	0.000000	0.026392	0.000423	0.000000	0.043476	0.017580	0.035602	0.019042
Race	0.029753	0.034832	0.000000	0.000000	0.023434	0.000297	0.000000	0.039947	0.010781	0.035161	0.017421
Smoked cigarettes regularly	0.039531	0.008540	0.000000	0.000000	0.009863	0.000896	0.000000	0.002964	0.061642	0.036361	0.015980
Had high blood pressure	0.039531	0.035252	0.000000	0.000000	0.007099	0.000458	0.000000	0.012088	0.029294	0.035360	0.015908
Used aspirin regularly	0.038754	0.035426	0.000000	0.000000	0.004367	0.000276	0.000000	0.008227	0.036734	0.034354	0.015814
Had gall bladder inflammation	0.039531	0.035332	0.000000	0.000000	0.008658	0.000423	0.000000	0.034352	0.000000	0.035024	0.015332
Used ibuprofen regularly	0.036633	0.034060	0.000000	0.000000	0.005967	0.000370	0.000000	0.018769	0.016423	0.034285	0.014651
Had arthritis	0.039531	0.009376	0.000000	0.000000	0.008306	0.000486	0.000000	0.000408	0.051797	0.035090	0.014500
Occupation	0.027181	0.014915	0.000000	0.000000	0.000567	0.000487	0.000000	0.000711	0.057166	0.035207	0.013623
Marital Status	0.026998	0.035033	0.000000	0.000000	0.006324	0.000377	0.000000	0.023159	0.007962	0.034156	0.013401
No. of brothers	0.013149	0.039036	0.000000	0.000000	0.000000	0.000230	0.000000	0.006553	0.039952	0.034104	0.013302
No. of sisters	0.012518	0.034216	0.000000	0.000000	0.001551	0.000400	0.000000	0.006412	0.040223	0.034255	0.012958
No. of cigarettes smoked daily	0.007264	0.034150	0.000000	0.000000	0.001918	0.001207	0.000000	0.005654	0.041690	0.036823	0.012871
Education level completed	0.003636	0.011414	0.000000	0.000000	0.000292	0.000423	0.000000	0.000937	0.057478	0.034579	0.010876

(2a) Male

	Chi-square	MRMR	Dtree ensemble	Binary Dtree	Random forest	Class sep. (ttest)	Class sep. (entropy)	Class sep. (roc)	Class sep. (wilcoxon)	Pearson corr.	Average
Diagnosed with PC	0.041397	0.034257	1.000000	1.000000	0.092794	0.892307	0.996286	0.054429	0.040203	0.000000	0.415167
Had osteoporosis	0.039582	0.034156	0.000000	0.000000	0.078983	0.002422	0.003702	0.053335	0.038400	0.032909	0.028349
Had colon comorbidity	0.039306	0.034037	0.000000	0.000000	0.080111	0.003987	0.000000	0.053341	0.038410	0.031782	0.028099
Had stroke	0.037734	0.034322	0.000000	0.000000	0.065797	0.002799	0.000000	0.051695	0.035698	0.032522	0.026057
Had emphysema	0.037151	0.034393	0.000000	0.000000	0.059490	0.002173	0.000000	0.050865	0.034331	0.033101	0.025150
Diagnosed with any cancer	0.037608	0.034501	0.000000	0.000000	0.058463	0.002602	0.000000	0.050592	0.033880	0.033563	0.025121
Had diverculitis	0.035503	0.034661	0.000000	0.000000	0.053117	0.004969	0.000000	0.050058	0.033000	0.031750	0.024306
Had liver comorbidity	0.035949	0.034821	0.000000	0.000000	0.053241	0.001786	0.000000	0.049780	0.032542	0.032910	0.024103
No. of relatives with PC	0.036832	0.033502	0.000000	0.000000	0.045818	0.005118	0.000000	0.049323	0.031790	0.034276	0.023666
Smoked cigarettes regularly	0.039737	0.038604	0.000000	0.000000	0.015964	0.003080	0.000000	0.016597	0.079303	0.033603	0.022689
Hispanic origin	0.034318	0.032583	0.000000	0.000000	0.044114	0.002039	0.000000	0.047474	0.028742	0.032650	0.022192
Had bronchitis	0.035430	0.034512	0.000000	0.000000	0.034584	0.005647	0.000000	0.047301	0.028458	0.034468	0.022040
Has relative with PC	0.032636	0.033175	0.000000	0.000000	0.032426	0.002467	0.000000	0.045313	0.025183	0.033255	0.020445
Had colorectal polyps	0.030968	0.034889	0.000000	0.000000	0.030835	0.002864	0.000000	0.044330	0.023563	0.033304	0.020075
Had diabetes	0.039693	0.034695	0.000000	0.000000	0.028858	0.005851	0.000000	0.038803	0.014455	0.034746	0.019710
Had gall bladder inflammation	0.031775	0.034628	0.000000	0.000000	0.025801	0.005219	0.000000	0.041563	0.019004	0.034641	0.019263
Race	0.029198	0.034184	0.000000	0.000000	0.030827	0.002600	0.000000	0.041472	0.018853	0.033004	0.019014
Occupation	0.030038	0.022884	0.000000	0.000000	0.001955	0.010006	0.000000	0.014177	0.075314	0.035753	0.019013
Had high blood pressure	0.039737	0.045747	0.000000	0.000000	0.016967	0.004407	0.000000	0.010840	0.030518	0.034167	0.018238
Had heart attack	0.039737	0.034963	0.000000	0.000000	0.018270	0.003043	0.000000	0.037158	0.011745	0.033665	0.017858
Has relative with cancer	0.039189	0.018329	0.000000	0.000000	0.019140	0.002507	0.000000	0.002473	0.056032	0.033368	0.017104
Had arthritis	0.039190	0.035341	0.000000	0.000000	0.016834	0.003973	0.000000	0.016953	0.020445	0.033960	0.016669
Marital Status	0.025489	0.033300	0.000000	0.000000	0.020596	0.003165	0.000000	0.035986	0.009814	0.033193	0.016154
Used ibuprofen regularly	0.039466	0.035351	0.000000	0.000000	0.018994	0.002851	0.000000	0.030030	0.000000	0.032508	0.015920
Smoked cigar	0.032222	0.033769	0.000000	0.000000	0.021921	0.002081	0.000000	0.027528	0.003020	0.032869	0.015341
Smoked pipe	0.032804	0.033843	0.000000	0.000000	0.016339	0.002286	0.000000	0.023928	0.008952	0.032696	0.015085
Used aspirin regularly	0.037076	0.005314	0.000000	0.000000	0.010745	0.002663	0.000000	0.001903	0.055092	0.033150	0.014594
No. of brothers	0.007658	0.040127	0.000000	0.000000	0.001329	0.002682	0.000000	0.004764	0.040521	0.033001	0.013008
No. of sisters	0.006775	0.036513	0.000000	0.000000	0.002763	0.002791	0.000000	0.003299	0.042936	0.033520	0.012860
Education level completed	0.014618	0.017745	0.000000	0.000000	0.000440	0.005241	0.000000	0.000000	0.049132	0.031461	0.011865
No. of cigarettes smoked daily	0.001185	0.020855	0.000000	0.000000	0.002482	0.004378	0.000000	0.004675	0.040665	0.034203	0.010844

(2b) Female

Figure 2: Weights assigned by 9 feature selection algorithms (columns 1-9) to risk factors in the PLCO dataset.

Finding probability feature combination using a bayesian network: Russell and Norvig in their book, Artificial Intelligence: A Modern Approach [35] have illustrated about Bayes Theorem and joint probability. Consider that the symptoms E1, E2 are conditionally independent. Then their co-occurrence is as follows:

Using the above equation,

Equation 3:

$$P(E_1, E_2 \mid C) = P(E_1 \mid C) P(E_2 \mid C)$$

Equation 4:

$$P(C \mid E_1, E_2) = \frac{(P(C) P(E_1, E_2 \mid C))}{\mathrm{P}(E_1 \quad E_2)}$$

Equation 5:

As any individual will either have PC or not have PC with the given symptoms, considering a universal set, P(E1ΔE2) can be resolved using normalization as follows:

$$P(C\mid E_1,E_2)+P(\bar{C}\mid E_1,E_2)=1$$

Equation 6:

P(¯) is the probability of non-occurrence of PC. Hence,

$$P\left(E_1\Delta E_2\right)=P\left(C\right)P(E_1,E_2\mid C)+P\left(\bar{C}\right)P(E_1,E_2\mid \bar{C})$$

Equation 7:

Substituting equation 3 in equation 6

$$P\left(E_1\Delta E_2\right)=P\left(C\right)P(E_1\mid C)P(E_2\mid C)+P\left(\bar{C}\right)P(E_1\mid \bar{C})P(E_2\mid \bar{C})$$

Equation 8:

Substituting equation 4 in equation 7,

$$P(C\mid E_1,E_2)=\frac{P\left(C\right)P(E_1,E_2\mid C)}{P\left(C\right)P(E_1,E_2\mid C)+P\left(\bar{C}\right)P\left(E_{1,}E_2\right)\mid \bar{C})}$$

Equation 9:

Substituting equation 3 in equation 8,

$$P(C\mid E_1,E_2)=\frac{P\left(C\right)P(E_1\mid C)P(E_2\mid C)}{P\left(C\right)P(E_1\mid C)P(E_2\mid C)+P\left(\bar{C}\right)P(E_1\mid \bar{C})P(E_2\mid \bar{C})}$$

DISCUSSION

A number of risk factors of PC have been identified [36-38], such as, smoking, obesity, exposure to certain chemicals (e.g., pesticides, benzene, certain dyes, petrochemicals), age (older than 55 years), gender (male), race/ethnicity (Blacks, Ashkenazi Jewish heritage), family history (two or more first-degree relatives with PC), inherited genetic syndromes, diabetes, pancreatic cysts and chronic pancreatitis. Several different genes are associated with increased risk of PC. However, genetic risk factors are beyond the scope of this work as our dataset does not contain genetic information. Table 2 shows the ratio in which each symptom was distributed in the HPT (High Probability Table) chosen by selecting top percentage values of probability for that feature combination and in the LPT (Low Probability Table) chosen by selecting bottom percentage values of probability for that feature combination.

Table 2: Table containing features from PLCO dataset that is plausible to being indicators of risk of PC

Symptoms	Results	Conclusion
Occupation	All subjects in HPT were retired (category 4) and in LPT, they were extended sick leave (category 5)	Older people have a greater risk of PC
Smoked pipe	Subjects in HPT were in ratio 0.22 (never smoked): 0.5 (currentsmoker): 0.28 (pastsmoker) whereas subjects in LPT were in ratio 0.37 (never smoked): 0.27 (current smoker): 0.36 (past smoker)	Subjects who never smoked have a lesser risk than past smokers and risk for current smokers was doubled
Heart Attack	Subjects in HPT were in ratio 0.23 (never had heart at-tack): 0.77 (had heart attack) whereas subjects in LPT were in ratio 0.8 (never had heart attack): 0.2 (had heart attack)	Subjects who had heart attack at least once have a greater risk for PC
Hypertension	Subjects in HPT were in ratio 0.36 (not diagnosed with hypertension): 0.63 (diagnosed with hypertension) whereas subjects in LPT were in ratio 0.68 (not diagnosed with hypertension): 0.32 (diagnosed with hypertension)	Stress (or hypertension) is directly proportional to risk for PC
Taken female hormones	Subjects in HPT were in ratio 0.7 (never taken): 0.3 (taken) whereas subjects in LPT were in ratio 0.23 (never taken): 0.77 (taken)	Somehow female hormones reduces risk of PC
Race	Subjects in HPT were mostly Asian (0.38) and only 0.3 were Pacific Islander whereas subjects in LPT were mostly American Indian (0.85)	Clearly shows that Asians are at a higher risk of PC while Pacific Islander and American Indian were at lower risk.
Diabetes	Subjects in HPT were in ratio 0.17 (never had diabetes): 0.83(had diabetes) whereas subjects in LPT were in ratio 0.75 (never had diabetes): 0.25 (had diabetes)	Diabetes is a clear risk factor for PC
Bronchitis	Subjects in HPT were in ratio 0.27 (never had): 0.73 (had) whereas subjects in LPT were in ratio 0.68 (never had): 0.32 (had)	Bronchitis is a risk factor for PC
Liver comorbidities	Subjects in HPT were in ratio 0.39 (never had): 0.61 (had) whereas subjects in LPT were in ratio 0.62 (never had): 0.38(had)	Liver comorbidities is a risk factor for PC
Colorectal Polyps	Subjects in HPT were in ratio 0.36 (never had): 0.64 (had) whereas subjects in LPT were in ratio 0.62 (never had):0.38 (had)	Colorectal Polyps is a risk factor for PC
Gender	Subjects in HPT were in ratio 0.53 (male): 0.47 (female) whereas subjects in LPT were in ratio 0.35 (male): 0.65 (female)	Male were at higher risk of PC than female
No of Relatives with pancreatic cancer	Subjects in HPT were in ratio 0.02 (no relative):0.1 (1 relative): 0.88 (2 relatives) whereas subjects in LPT were in ratio 0.71(no relative: 0.29 (1 relative)	Risk of PC increases as incidence of PC on family members increases.

Ever take birth control pills?	Subjects in HPT were in ratio 0.76 (no history): 0.24 (has history) whereas subjects in LPT were in ratio 0.17 (no history): 0.83 (has history)	birth control pills may lower risk of PC
Smoke regularly now?	Subjects in HPT were in ratio 0.12 (no history): 0.88 (has history) whereas subjects in LPT were in ratio 0.95 (no history): 0.05 (has history)	Current smokers have higher risk of PC
Ever smoke regularly more than 6 months?	Subjects in HPT were in ratio 0.22 (no history): 0.78 (has history) whereas subjects in LPT were in ratio 0.85 (no history): 0.15 (has history)	Smoking in excess of 6 months also poses higher risk of PC

Factors with unclear effect on risk include nature of diet, lack of physical activity, coffee and alcohol consumption, and certain infections (see for example, [38]).

Smoking: Several studies have shown that smoking has a significant relationship with PC (see for example [39-44]). Yadav et al. found that smoking cessation can significantly reduce risk of PC [43]. Raimondi et al. argue that smoking is the most common risk factor and accounts for 20-25% of all pancreatic tumors [44].

Diabetes: Diabetes also has a positive correlation with PC [45]. Huxley et al. shows that individuals who have had type-II diabetes for less than four years were at a 50% higher risk of contracting PC than individuals who have had type-II diabetes for more than four years [3]. Everhart et al. have concluded that subjects with long standing diabetes have a higher relative risk of PC [4]. Ben et al. have also found similar relationship between diabetes and PC [46]. Liao et al shows that subjects in Taiwan who have had diabetes for less than 2 years are at elevated risk of PC [5]. Long standing diabetes did not pose a strong risk. Also concurrent occurrence of diabetes and chronic pancreatitis puts subjects at a higher risk.

Reproductive history in women: Lo, et al. has shown that women with 7 or more live births had a lower risk of PC. Lactation period also had a significant effect on the possibility of PC [47]. This study shows that women who lactated for 144 months or more had a one-fifth the risk of PC than women who lactated only for 89 months or less. Kreiger et al. have shown that PC is an estrogen-dependent disease and aspects of reproductive history and hormone replacement are associated with a greater risk of this disease. Reduced risks were observed with 3 or more pregnancies and with the use of oral contraceptives [48].

Marriage: Baine et al. shows marriages improves the survival rate and longevity of patients with PC [31]. This paper also shows using Kaplan-Meier analysis, that patients who were married had a median survival rate of 4 months in comparison to unmarried patients who had a survival rate of 3 months. Aizer et al. have shown that marriage has a beneficial effect on any cancer with regards to detection, treatment and survival [49]. This improvement was observed more in males than females, highlighting the socio-economic elevation that a married person could have. The paper concluded that ”married people were less likely to present metastatic disease, more likely to receive definitive therapy, and less likely to die as a result of their cancer after adjusting for demographics, stage, and treatment than unmarried patients.” Multivariate logistic and Cox regression were used to analyze the patients.

Occupation: Logan et al. shows how specific types of occupation pose higher risk to exposure to carcinogenic substances [50]. In 1961 and 1971, for men, occupation categories of clothing, food, drink and tobacco and armed forces had higher Standardised Mortality Rates (SMR) and relative standardised mortality rates (RSMR) whereas people in the clerical and leather industry saw low SMR and RSMR. For men in the occupation categories of mining, labourers and service, sport and recreation saw elevated but reduced RSMR. For men in administrative and managerial, and professional and technical disciplines, the trend was reduced SMR and elevated RSMR. In case of married women, if husbands worked in engineering, leather, wood, sales, clothing, construction work, both SMR and RSMR were high. For wives of husbands working in farm, gas, coke and chemicals industry, glass and ceramics and warehouse, both SMR and RSMR were low. In 1961, wives of husband in food, drink and tobacco had high SMR and RSMR and values were low for husbands in painting and decoration industry, and the trend was reversed in 1971.

Family composition: Gharidian et al. have found an interesting relationship wherein there is the occurrence of this disease in two brothers and one sister in all the seventh decade of their life [51]. This study was based in Montreal and there was no pancreatitis history between the patients or their relatives.

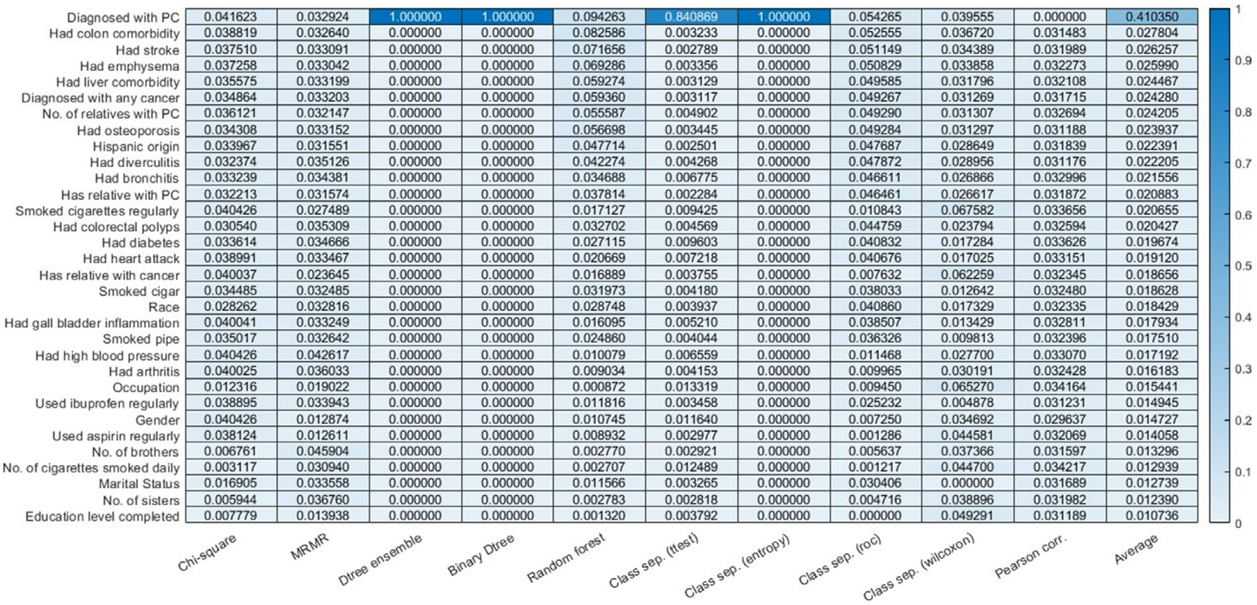
Use of certain medications: Tan et al. have shown that aspirin use decreases risk of procuring PC [52]. Aspirin use for 1day/ month or greater was associated with a lower risk of PC than subjects who had aspirin for less than 1day/month. According to this study, there are no relationships between non-aspirin non-steroidal anti-inflammatory drugs (NSAID) and PC. Larsson et al. have provided doubtful evidence that regular use of aspirin over longer duration increases risk of PC [53]. No relationship was found between use of frequent aspirin (7 tablets or more/week) or prolonged use of aspirin (more than 20 years) and the increase/decrease in PC. Harris et al. have found a relationship between aspirin, ibuprofen, and other Non-Steroidal Anti-Inflammatory Drugs (NSAID) and cancer prevention [54]. However, results varied for different types of cancer.

Surgical history: Rosenberg et al. have shown a positive correlation in increase in risk of PC by 1.8% because of vasectomy [55].

Inherited genetic syndrome: Certain rare genetic conditions cause almost 10% of all PCs. In our investigation, it can be found under family history of PC that have been chosen by two of the feature-selection algorithms, viz, Relieff and Lasso in Table 3. Also, no of relatives with PC has been chosen by 4 of the feature selection algorithms, viz, ECFS, UDFS, LLCFS and CFS. From the graphs in (Figure 3), it can be seen that if subject has family history of PC or any form of cancer, there is increase in probability of PC. Further the trend of increase is almost exponential as no. of relatives with PC increases, which strongly suggests that genetics play an important role in determination of possibility of PC. Such rare genetic conditions include [36]:

Table 3: Table containing 2 features combinations from PLCO dataset that produces highest risk of PC for male

Symptom 1	Symptom 2	Probability
Age when told had inflamed prostate=70+	No of cigarettes smoked daily=80+	0.032
Age when told had inflamed prostate=70+	Prior history of any cancer?=Yes	0.03
Prior history of any cancer?=Yes	No of cigarettes smoked daily=80+	0.026
Age when told had inflamed prostate=70+	Age when told had enlarged prostate=70+	0.026
Age when told had inflamed prostate=70+	Family history of PC?=Yes	0.026
Age when told had inflamed prostate=70+	No of relatives with PC=1	0.026
Age when told had enlarged prostate=70+	No of cigarettes smoked daily=80+	0.024
Family history of PC=Yes	No of cigarettes smoked daily=80+	0.024
No of relatives with PC=1	No of cigarettes smoked daily=80+	0.024
Age when told had inflamed prostate=70+	Bronchitis history?=Yes	0.022
Age when told had enlarged prostate=70+	Prior history of any cancer?=Yes	0.022
Prior history of any cancer?=Yes	Family history of PC=Yes	0.022
Prior history of any cancer?=Yes	No of relatives with PC=1	0.021
Age when told had inflamed prostate=70+	Gall bladder stone or inflammation=Yes	0.021
Age when told had inflamed prostate=70+	Smoke regularly now?=Yes	0.021
No of cigarettes smoked daily=80+	Bronchitis history?=Yes	0.021
Age when told had inflamed prostate=70+	During past year, how many times wake up in the night to urinate?=Thrice	0.021
Age when told had inflamed prostate=70+	Smoked pipe=current smoker	0.021
Age when told had inflamed prostate=70+	Diabetes history=yes	0.02
Age when told had inflamed prostate=70+	No. of brother=7+	0.02



	Chi-square	MRMR	Dtree ensemble	Binary Dtree	Random forest	Class sep. (ttest)	Class sep. (entropy)	Class sep. (roc)	Class sep. (wilcoxon)	Pearson corr.	Average
Diagnosed with PC	0.041623	0.032924	1.000000	1.000000	0.094263	0.840869	1.000000	0.054265	0.039555	0.000000	0.410350
Had colon comorbidity	0.038819	0.032640	0.000000	0.000000	0.082586	0.003233	0.000000	0.052555	0.036720	0.031483	0.027804
Had stroke	0.037510	0.033091	0.000000	0.000000	0.071656	0.002789	0.000000	0.051149	0.034389	0.031989	0.026257
Had emphysema	0.037258	0.033042	0.000000	0.000000	0.069286	0.003356	0.000000	0.050829	0.033858	0.032273	0.025990
Had liver comorbidity	0.035575	0.033199	0.000000	0.000000	0.059274	0.003129	0.000000	0.049585	0.031796	0.032108	0.024467
Diagnosed with any cancer	0.034864	0.033203	0.000000	0.000000	0.059360	0.003117	0.000000	0.049267	0.031269	0.031715	0.024280
No. of relatives with PC	0.036121	0.032147	0.000000	0.000000	0.055587	0.004902	0.000000	0.049290	0.031307	0.032694	0.024205
Had osteoporosis	0.034308	0.033152	0.000000	0.000000	0.056698	0.003445	0.000000	0.049284	0.031297	0.031188	0.023937
Hispanic origin	0.033967	0.031551	0.000000	0.000000	0.047714	0.002501	0.000000	0.047687	0.028649	0.031839	0.022391
Had diverculitis	0.032374	0.035126	0.000000	0.000000	0.042274	0.004268	0.000000	0.047872	0.028956	0.031176	0.022205
Had bronchitis	0.033239	0.034381	0.000000	0.000000	0.034688	0.006775	0.000000	0.046611	0.026866	0.032996	0.021556
Has relative with PC	0.032213	0.031574	0.000000	0.000000	0.037814	0.002284	0.000000	0.046461	0.026617	0.031872	0.020883
Smoked cigarettes regularly	0.040426	0.027489	0.000000	0.000000	0.017127	0.009425	0.000000	0.010843	0.067582	0.033656	0.020655
Had colorectal polyps	0.030540	0.035309	0.000000	0.000000	0.032702	0.004569	0.000000	0.044759	0.023794	0.032594	0.020427
Had diabetes	0.033614	0.034666	0.000000	0.000000	0.027115	0.009603	0.000000	0.040832	0.017284	0.033626	0.019674
Had heart attack	0.038991	0.033467	0.000000	0.000000	0.020669	0.007218	0.000000	0.040676	0.017025	0.033151	0.019120
Has relative with cancer	0.040037	0.023645	0.000000	0.000000	0.016889	0.003755	0.000000	0.007632	0.062259	0.032345	0.018656
Smoked cigar	0.034485	0.032485	0.000000	0.000000	0.031973	0.004180	0.000000	0.038033	0.012642	0.032480	0.018628
Race	0.028262	0.032816	0.000000	0.000000	0.028748	0.003937	0.000000	0.040860	0.017329	0.032335	0.018429
Had gall bladder inflammation	0.040041	0.033249	0.000000	0.000000	0.016095	0.005210	0.000000	0.038507	0.013429	0.032811	0.017934
Smoked pipe	0.035017	0.032642	0.000000	0.000000	0.024860	0.004044	0.000000	0.036326	0.009813	0.032396	0.017510
Had high blood pressure	0.040426	0.042617	0.000000	0.000000	0.010079	0.006559	0.000000	0.011468	0.027700	0.033070	0.017192
Had arthritis	0.040025	0.036033	0.000000	0.000000	0.009034	0.004153	0.000000	0.009965	0.030191	0.032428	0.016183
Occupation	0.012316	0.019022	0.000000	0.000000	0.000872	0.013319	0.000000	0.009450	0.065270	0.034164	0.015441
Used ibuprofen regularly	0.038895	0.033943	0.000000	0.000000	0.011816	0.003458	0.000000	0.025232	0.004878	0.031231	0.014945
Gender	0.040426	0.012874	0.000000	0.000000	0.010745	0.011640	0.000000	0.007250	0.034692	0.029637	0.014727
Used aspirin regularly	0.038124	0.012611	0.000000	0.000000	0.008932	0.002977	0.000000	0.001286	0.044581	0.032069	0.014058
No. of brothers	0.006761	0.045904	0.000000	0.000000	0.002770	0.002921	0.000000	0.005637	0.037366	0.031597	0.013296
No. of cigarettes smoked daily	0.003117	0.030940	0.000000	0.000000	0.002707	0.012489	0.000000	0.001217	0.044700	0.034217	0.012939
Marital Status	0.016905	0.033558	0.000000	0.000000	0.011566	0.003265	0.000000	0.030406	0.000000	0.031689	0.012739
No. of sisters	0.005944	0.036760	0.000000	0.000000	0.002783	0.002818	0.000000	0.004716	0.038896	0.031982	0.012390
Education level completed	0.007779	0.013938	0.000000	0.000000	0.001320	0.003792	0.000000	0.000000	0.049291	0.031189	0.010736

Figure 3: Weights assigned by 9 feature selection algorithms (columns 1-9) to risk factors in the PLCO dataset.

- Hereditary breast and ovarian cancer syndrome, caused by mutations in the BRCA1 or BRCA2 genes,
- Hereditary breast cancer, caused by mutations in the PALB2 gene,
- Familial atypical multiple mole melanoma (FAMMM) syndrome, caused by mutations in the p16/CDKN2A gene and
- associated with skin and eye melanomas,
- Familial pancreatitis, usually caused by mutations in the PRSS1 gene,
- Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC), most often caused by a defect in the
- MLH1 or MSH2 genes,
- Peutz-Jeghers syndrome, caused by defects in the STK11 gene. This syndrome is also linked with polyps in the

digestive

- Tract and several other cancers.

Race: Race has been a predominant factor in the determination of the risk of PC [36-38]. According to literature, blacks or African American people have a higher risk of contracting PC. This could be attributed to their dietary habits or smoking history. Race has been chosen as one of the features by 3 of our feature-selection algorithms, viz, Laplacian, FSASL and LLCFS in 2 and also Asian race has been chosen as one of the highest probability of PC causing feature in Table 3.

Gender: Literature has shown that men are more likely to contract PC than women [36-38]. This could be because men are more likely to smoke than women and smoking has a significant effect on PC. Gender has been chosen as one of the features by 3 of our feature-selection algorithms, viz, Laplacian, CFS and ECFS in Table 3 and also gender is male in the highest probability of PC in Table 3.

Female hormones: Experimental findings from this article by on use of affect of female hormones suggest that female hormones have a protective role towards incidence of PC [56].

Bronchitis: Although there is no direct evidence between bronchitis and PC risk, is a study conducted on male smokers in Finland that suggests that bronchial asthma predict the subsequent risk of developing PC in male smokers and that greater physical activity may decrease the risk [57]. Also bronchial asthma can increase chances of developing bronchitis.

Heart attack: Many references suggest the increased association between heart attack and stroke with any type of cancer (not necessarily PC). It shows the increased risk of heart attack and stroke in the months leading up to cancer diagnosis. In another article [58,59], it shows that recent epidemiological analyses suggest that cancer incidence is more common among subjects with a history of heart failure versus subjects with no history of heart failure.

Hypertension: Some references, for example suggestion that hypertension at baseline was associated with an increased risk of PC incidence [60]. Although the above factors-inherited genetic syndrome, race, diabetes history and gender have a strong relationship with PC, yet they were not one of the highly selected features by our algorithms, probably because other features have a stronger dependence when considered in unison.

Most of the remaining features as seen in Table 2 do not have a strong evidence yet to their dependency with PC, however they can act as a guide to biologists and researchers to delve into possible correlation between these symptoms Tables 4 and 5.

Table 4: Table containing 2 features combinations from PLCO dataset that produces highest risk of PC for female

Symptom 1	Symptom 2	Probability
No of cigarettes smoked daily=61-80	No of relatives with PC=2+	0.156
No of tubal/ectopic pregnancies=2+	No of relatives with PC=2+	0.137
Usually filtered or not filtered?=Both	No of relatives with PC=2+	0.115
No of cigarettes smoked daily=61-80	No of relatives with PC=2+	0.095
No of tubal/ectopic pregnancies=1	No of relatives with PC=2+	0.084
Heart attack history?=yes	No of relatives with PC=2+	0.08
No of cigarettes smoked daily=21-30	No of relatives with PC=2+	0.077
No of relatives with PC=2+	Race=Asian	0.076
No of relatives with PC=2+	No of still births=1	0.074
No of relatives with PC=2+	Diabetes history?=Yes	0.0737
No of relatives with PC=2+	Race=American Indian	0.0737
No of relatives with PC=2+	Emphysema history?=Yes	0.0737
No of relatives with PC=2+	No of cigarettes smoked daily=31-40	0.0708
No of relatives with PC=2+	Colorectal Polyps history?=Yes	0.0708
No of relatives with PC=2+	Stroke history?=Yes	0.0704
No of relatives with PC=2+	Age at hysterectomy=40-44	0.0686
No of relatives with PC=2+	No. of brothers=7+	0.0645
No of relatives with PC=2+	Bronchitis history?=2+	0.064
No of relatives with PC=2+	Liver comorbidities history?=Yes	0.063
No of relatives with PC=2+	No of cigarettes smoked daily=11-20	0.063

Table 5: Table containing 3 features combinations from PLCO dataset that produces highest risk of PC for male and female

Male			
Symptom 1 conditional probability	Symptom 2 conditional probability	Symptom 3 conditional probability	Total probability
No of cigarettes smoked Daily is 61-80=0.005	Age when told had enlarged Prostate is 70+=0.0175	Prior history of cancer is Yes=0.05	0.00521
No of cigarettes smoked Daily is 61-80=0.005	Age when told had enlarged Prostate is 70+=0.0175	Family history of Pc=yes=0.533	0.002368
No of cigarettes smoked Daily is 61-80=0.005	Age when told had enlarged Prostate is 70+=0.0175	No. of relatives with pc is 1=0.04	0.004593
Prior history of cancer is Yes=0.05	Age when told had enlarged Prostate is 70+=0.0175	Family history of pc is Yes=0.533	0.002153

Prior history of cancer is Yes=0.05	Age when told had enlarged Prostate is 70+=0.0175	No. of relatives with pc is 1=0.04	0.004177
Age when told had enlarged Prostate is 70+=0.0175	Family history of pc is Yes=0.533	No. of relatives with pc is 1=0.04	0.00189
No of cigarettes smoked Daily is 61-80=0.005	Prior history of cancer is Yes=0.05	Family history of Pc=yes=0.533	0.02742
No of cigarettes smoked Daily is 61-80=0.005	Prior history of cancer is Yes=0.05	No. of relatives with pc is 1=0.04	0.05197055
No of cigarettes smoked Daily is 61-80=0.005	Family history of pc is Yes=0.533	No. of relatives with pc is 1=0.04	0.024218
Prior history of cancer is Yes=0.05	Family history of pc is Yes=0.533	No. of relatives with pc is 1=0.04	0.02206
Female			
No of tubal/ectopic pregnancies is 1=0.003	No. of relatives with pc is 2+=0.011	No. of cigarettes smoked is 61-80=0.007	0.3578

CONCLUSION

We have used widely used algorithms for our prediction for PC. Since the exact relationship between features and the cause of PC cannot be ascertained for sure, for example, some factors like education, marital status and several others could have an indirect causal relationship with this disease, hence these factors were not excluded from our prediction study. After running all the above algorithms, it is observed that k-means clustering and SMOTE method of oversampling are some of the superior algorithms for PC prediction. The artificial intelligence based Bayesian network prediction model can signify which individuals are at an elevated risk for PC.

Until now, very limited work has been done in PC prediction, so the accuracy obtained by our research is significant. Lack of online available datasets for PC has limited the work that can be done in this field. Still the PLCO dataset by NIH has been a very valuable resource. Future improvements can be made based on taking into account other features that would have been found as a possible precursor to PC, based on further research and availability of more datasets.

ACKNOWLEDGEMENT

The author of this paper would like to thank Dr Bonny Banerjee and Dr Chrysanthe Preza from the University of Memphis, Tennessee and Dr Subhash Chauhan and Dr Sheema Khan from University of Texas, Rio Grand Valley, Texas for their support and guidance in writing the paper.

CONFLICT OF INTEREST

No conflict of interest associated with this work.

REFERENCES

1. Wikipedia contributors (2019) Wikipedia, the free encyclopedia.
2. Ik-Gyu J (2019) Method of providing information for the diagnosis of pancreatic cancer using bayesian network based on artificial intelligence, computer program, and computer-readable recording media using the same. Cancers Basel 14(21): 5382.
3. Huxley R, Moghaddam AA, Berrington De Gonz´alez A, Barzi F (2005) Type-2 Diabetes and Pancreatic Cancer: A meta-analysis of 36 studies. Br J Cancer 92(11): 2076.
4. James E, David W (1995) Diabetes mellitus as a risk factor for pancreatic cancer: A meta-analysis. Jama 273(20): 1605–1609.
5. Ben Q, Xu M, Ning X, Liu J, Hong S, et al. (2011) Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. Eur J Cancer 47(13): 1928–1937.
6. Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, et al. (2009) Exomic sequencing identifies palb2 as a pancreatic cancer susceptibility gene. Sci 324(5924): 217–217.
7. Barton CM, Staddon SL, Hughes CM, Hall PA, Sullivan CO, et al. (1991) Abnormalities of the p53 tumour suppressor gene in human pancreatic cancer. Br J Cancer 64(6): 1076.
8. Donahue CAL, Fu B, Yachida S, Luo M, Abe H, et al. (2009) Dpc4 gene status of the primary carcinoma correlates with patterns of failure in patients with pancreatic cancer. J Clin Oncol 27(11): 1806.
9. Das A, Nguyen CC, Li F, Li B (2008) Digital image analysis of eus images accurately differentiates pancreatic cancer from chronic pancreatitis and normal tissue. Gastrointest Endosc 67(6): 861–867.
10. Guangtao G, Wong GW (2008) Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinform 9(1): 275
11. Saftoiu A, Vilmann P, Gorunescu F, Gheonea DL, Gorunescu M, et al. (2008) Neural network analysis of dynamic sequences of eus elastography used for the differential diagnosis of chronic pancreatitis and pancreatic cancer. Gastrointest Endosc 68(6): 1086–1094.
12. Zhang MM, Yang H, Jin ZD, Yu JG, Cai ZY, et al. (2010) Differential diagnosis of pancreatic cancer from normal tissue with digital imaging processing and pattern recognition based on a support vector machine of EUS

images. Gastrointest Endosc 72(5): 978–985.

13. Baruah M, Banerjee B (2020) Modality selection for classification on time-series data. MileTS 20: 6.

14. National Cancer Institute (2019) Pancreas-Datasets-PLCO-The Cancer Data Access System.

15. He H, Bai Y, Garcia EA, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. IJCNN 1322–1328.

16. Maaten LV, Hinton G. Visualizing data using T-SNE (2008) J Mach Learn Res 9: 2579–2605.

17. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. J Artif Intell Res 16: 321–357.

18. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11(1): 1–13.

19. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3: 1157–1182.

20. Tang J, Alelyani S, Liu H (2014) Feature selection for classification: A review. Data Cls Anal Algo Appl 37-39.

21. Roffo G, Melzi S (2016) Ranking to learn. In International workshop on new frontiers in mining complex patterns Springer 19: 35.

22. Mangasarian OL, Bradley PS (1998) Feature selection via concave minimization and support vector machines. Technical Report.

23. He X, Cai D, Niyogi P (2006) Laplacian score for feature selection. Adv Neural Inf Process 507: 514.

24. Yang Y, Shen HT, Ma Z, Huang Z, Zhou X (2011) L2, 1-norm regularized discriminative feature selection for unsupervised. IJCAI 1589-1594.

25. Du l, Shen YD (2015) Unsupervised feature selection with adaptive structure learning. ACM SIGKDD 209: 218.

26. Hall MA (1999) Correlation-based feature selection for machine learning. J Algorithms Comput Technol 6:3

27. Fonti V, Belitser E (2017) Feature selection using lasso. VU Amsterdam Research Paper.

28. Guo J, Zhu W (2018) Dependence guided unsupervised feature selection. AAAI Conf AI 32:1

29. Roffo G (2017) Ranking to learn and learning to rank: On the role of ranking in pattern recognition applications. Arxiv 1706:05933.

30. Roffo G, Melzi S, Castellani U, Vinciarelli A (2017) Infinite latent feature selection: A probabilistic latent graph-based ranking approach. ICC 1398–1406.

31. Baine M, Sahak F, Lin C, Chakraborty S, Lyden E, et al. (2011) Marital status and survival in pancreatic cancer patients: A seer based analysis. PloS one 6(6):21052.

32. Zhou ZH, Liu XY (2010) On multi-class cost-sensitive learning. Comput Intell 26(3):232–257.

33. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees, belmont, ca: Wadsworth. Int GP 432:151–166.

34. Zadrozny B, Langford J, Abe N (2003) Cost-sensitive learning by cost-proportionate example weighting. ICDM 3: 435.

35. Russell SJ, Norvig P (2016) Artificial intelligence: A modern approach. PEL

36. Dutta A “What are risk factors for Pancreatic Cancer?”

37. Pancreatic Cancer: Risk Factors Approved by the Cancer J Cell Biol

38. Lowenfels AB, Maisonneuve P (2005) Pancreatic cancer risk factors. J Cell Biol 95(4): 649-656

39. Li D, Xie K, Wolff R, Abbruzzese JL (2004) Pancreatic cancer. The Lancet 363(9414): 1049–1057.

40. Lynch SM, Vrieling A, Lubin JH, Kraft P, Mendelsohn JB, et al. (2009) Cigarette smoking and Pancreatic Cancer: A pooled analysis from the Pancreatic Cancer cohort consortium. Am J Epidemiol 170(4): 403–413.

41. Christos L, Max W, Diederik K (2017) Learning sparse neural networks through l 0 regularization. J Mach Learn Res 144(6): 1252–1261

42. Muscat JE, Stellman SD, Hoffmann D, Wynder EL (1997) Smoking and pancreatic cancer in men and women. Cancer Epidemiol Biomarkers Prev 6(1): 15–19.

43. Yadav D, Lowenfels AB (2013) The epidemiology of pancreatitis and pancreatic cancer. J Gastroenterol 144(6):1252–1261.

44. Raimondi S, Maisonneuve P, Lowenfels AB (2009) Epidemiology of pancreatic cancer: An overview. Nat Rev Gastroenterol Hepatol 6(12):699.

45. DT Silverman, M Schiffman, J Everhart, A Goldstein, KD Lillemoe et al. (1999) Diabetes mellitus, other medical conditions and familial history of cancer as risk factors for pancreatic cancer. Br J Cancer 80(11): 1830.

46. Liao KF, Lai SW, Li CI, Chen WC (2012) Diabetes mellitus correlates with increased risk of pancreatic cancer: A population-based cohort study in Taiwan. J Gastroenterol Hepatol 27(4):709–713.

47. Lo AC, Soliman AS, Ghawalby NE, Wahab MA, Fathy O, et al. (2007) Lifestyle, occupational, and reproductive factors in relation to pancreatic cancer risk. Pancreas 35(2): 120–129.

48. Nancy K, Jeanie L, Margaret S (2001) Hormonal factors and pancreatic cancer in women. Ann Epidemiol 11(8): 563–567.

49. Aizer AA, Chen MH, McCarthy EP, Mendu ML, Koo S, et al. (2013) Marital status and survival in patients with cancer. J Clin Oncol 31(31): 3869.

50. Logan WP (1982) Cancer mortality by occupation and social class 1851-1971.

51. Ghadirian P, Simard A, Baillargeon J (1987) Cancer of the pancreas in two brothers and one sister. Int J Pancreatol 2(5-6): 383–391.

52. Xiang-Lin Tn, Reid Lombardo KM, William R Bamlet, Ann LO, et al. (2011) Aspirin,nonsteroidal anti-inflammatory drugs, acetaminophen, and pancreatic cancer risk: A clinic-based case-control study. Cancer Prev 4(11): 1835–1841.

53. Susanna CL, Edward G, Leif B, Alicja W (2006) Aspirin and nonsteroidal anti-inflammatory drug use and risk of pancreatic cancer: A meta-analysis. Cancer Epidemiol Biomarkers Prev 15(12):2561–2564.

54. Harris RE, Donk JB, Doss H, Doss DB (2005) Aspirin, ibuprofen, and other non-steroidal anti-inflammatory drugs in cancer prevention: A critical review of non-selective cox-2 blockade. Oncol Rep 13(4): 559–583.

55. Rosenberg L, Palmer JR, Zauber AG, Warshauer ME, Strom BL et al. (1994) The relation of vasectomy to the risk of cancer. Am J Epidemiol 140(5): 431– 438

56. Andersson G, Borgquist S, Jirstrom K (2018) Hormonal factors and pancreatic cancer risk in women: The malmo diet and cancer study. Int J Cancer 143(1): 52–62.

57. Solomon RZS, Pietinen P, Taylor PR, Virtamo J, Albanes D (2013) A prospective study of medical conditions, anthropometry, physical activity, and pancreatic cancer in male smokers (finland). Cancer causes cntrl 31(2): 417–426

58. Navi BB, Reiner AS, Kamel H, Ladecola C, Okin PM et al. (2019) Arterial thromboembolic events preceding the diagnosis of cancer in older persons. Clinical trials and observations 133(8): 781–789

59. Bertero E, Canepa M, Maack C, Ameri P (2018) Linking heart failure to cancer Circ 138(7): 735–742.

60. Wang Z, White DL, Hoogeveen R, Chen L, Whitsel EA, et al. (2018) Anti-hypertensive medication use, soluble receptor for glycation end products and risk of pancreatic cancer in the womens health initiative study. J Clin Med 197(7).

61. Zhou H, Wang F, Tao P (2018) T-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations J Chem Theory Comput 14(11): 5499-5510.

62. Likas A, Vlassis N, Verbeek JJ (2003) The global k-means clustering algorithm. Pattern Recognit 36(2): 451–461

63. Train models to classify data using supervised machine learning - MATLAB (2019)

64. Manning C, Raghavan P, Schutze H (2008) Introduction to Information Retrieval, volume 39. Cambridge University Press

65. Radovic M, Ghalwash M, Filipovic N, Obradovic Z (2017) Minimum redundancy maximum relevance feature selection approach for temporal gene expression data BMC Bioinform 18(1): 1–14.

66. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data J Bioinform Comput Biol 3(02): 185–205. ,

67. Breiman L (1996) Bagging predictors Mach Learn 24(2): 123–140.

68. Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees 8(5):534-535.

69. Coppersmith D, Hong SJ, Hosking JRM (1999) Partitioning nominal attributes in decision trees Data Min Knowl Discov 3(2): 197–217.

70. Rugeles D (2012) Study of feature ranking using bhattacharyya distance. 1: 1

71. King AP, Eckersley R (2019) Statistics for biomedical engineers and scientists: How to visualize and analyze data. Academic Press

72. Biesiada J, Duch WLL, Kachel A, Maczka K, Palucha S (2005) Feature ranking methods based on information entropy with parzen windows 1: 1.

73. Liu H, Motoda H (2012) Feature selection for knowledge discovery and data mining.

74. https://en.wikipedia.org/wiki/Wilcoxon_signedrank_testhttps://en.wikipedia.org/Wilcoxon_%20signed-rank_test (2022).

75. Loh WY, Shih YH (1997) Split selection methods for classification trees. Stat Sin 7(4): 815–840.

76. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and inter- pretation. Anesth Analg 126(5): 1763–1768.

77. https://en.wikipedia.org/Pearson_correlation_coefficient) 2022

78. Roffo G, Melzi S, Cristani M (2015) Infinite feature selection Proc IEEE Int Conf Comput 4202–4210.

79. Kira K, Rendell LA (1992)The feature selection problem: Traditional methods and a new algorithm 2(1992a): 129–134.

80. Kira K, Rendell LA (1992) A practical approach to feature selection Mach Learn Proc 249–256.

81. https://en.wikipedia.org/wiki/Relief_(feature_selection) (2022).

82. Wikipedia contributors. Wikipedia, the free encyclopedia. 2022. [Online; accessed 5-Jun-2022].

83. Seiffert C, Khoshgoftaar TM, Hulse JV, Napolitano (2009) A Rusboost: A hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern 40(1): 185–197.

84. Wikipedia contributors. Wikipedia, the free encyclopedia:https://en.wikipedia.org/Precision_and_recall) 2022.