ISSN: 2476-2377

Research Article

International Journal of Cancer Research & Therapy

Mathematical Modeling of Cancers Using Machine Learning Algorithms

Ananya Dutta*

Gauhati University, Guwhati, Assam, India-781014

***Corresponding Author**
Ananya Dutta, Gauhati University, Guwhati, Assam, India-781014

Submitted:2023, Aug 08 ; **Accepted**: 2023, Sep 10: **Published**: 2023, Sep 22

Citation: Dutta, A. (2023). Mathematical Modeling of Cancers Using Machine Learning Algorithms. *Int J Cancer Res Ther, 8*(3), 116-126.

Abstract

This paper shows a mathematical modeling method using different machine learning algorithms for prediction of probability of procuring Pancreatic Cancer (PC). Each algorithm reports its own accuracy, precision, recall and F1-score. Also, a Bayesian network model is used to determine the probability each subject has in contracting PC on the basis of certain preconditions, like his dietary habits and other biological attributes. This paper makes use of the PC dataset as provided by the National Cancer Institute in collaboration with National Institute of Health (NIH). The features obtained from this dataset can have either a binary value or a scalar value. The dataset consists of three questionnaires distributed to 155000 subjects. In each of these questionnaires, the subject is asked about his dietary habits and illness history.

Keywords: NIH-PLCO Dataset, Feature Selection, Bayesian Network, Prediction Model, Feature Graph-Trends

1. 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 five-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 PC prediction model that can lead to early detection of this disease.

Many researchers are in search of biomarkers for early diagnosis of PC (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 (EUS) and computed tomography (CT) images see for example, [9–12]. However, these models can detect PC only at an advanced stage and hence are not very useful.

As a follow-up on our previous work on pancreatic cancer, this paper uses machine learning algo- rithms to identify a subset of features from the PLCO dataset as useful predictors of PC [13]. The Prostate, Lung, Colorectal and Ovarian (PLCO) cancer dataset is collected by the National Cancer Institute from approximately 155,000 participants. Each participant responded to three questionnaires consisting of 65 questions (or features) about demographics, dietary habits, illness history, and family background. The dataset is highly imbalanced. To solve the unbalancing problem, we can use some data balancing algorithms to oversample the minority datapoints or undersample the majority datapoints so that both these datapoints are equal in number.

2. Problem Statement

Our problem is to infer whether a subject has pancreatic cancer or not given information about his demographic characteristics, dietary habits, illness history, and family background. This information is encoded as a vector of features. Formally, given a set of data points $X = [\vec{x}_1, ..., \vec{x}_N]^{d \times N}$ and a set of labels T rue, F alse , the goal is to map each data point $\vec{x}_i^{\,d}$ 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.

In this paper, we will use the PLCO dataset where N = 155, 000. Each data point represents a subject as a d = 65dimensional feature vector. The biggest challenge with the PLCO dataset is that it is highly imbalanced. Only 0.48% of the data points belong to the T rue class; rest are F alse. To classify this imbalanced dataset, we use techniques from data visualization, data balancing using oversampling and undersampling, feature selection, and probabilistic inference. Along the way, we find interesting insights into the correlates of pancreatic cancer, some of which are consistent with what has been reported in the medical sciences/ healthcare literature while a few others are yet to be investigated thoroughly.

3. Models and Methods

3.1 Data Visualization Methods

t-distributed stochastic neighbor embedding (t-SNE) algorithm. t-SNE is a widely-used algo- rithm for dimensionality reduction that can be used to visualize high dimensional data by embedding the datapoints into a 2D or 3D-space. As Van der Maaten and Hinton explained [1, 14]. “The similarity of datapoint

x_j to datapoint x_i is the conditional probability p_j *i*, **that** x_i would pick as its neighbor x_j if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i."

$$p_{j|i} = \frac{exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})}{\sum_{k \neq i} exp(-\frac{\|x_i - x_k\|^2}{2\sigma^2})} \quad (1)$$

Adaptive Synthetic (ADASYN) algorithm. ADASYN is a method of generating synthetic examples for minority classes using a weighted distribution as shown in Figure 2. The algorithm flowchart is described in detail in [15].

3.2 Data Balancing Methods

k-means clustering. k-means clustering performed on the majority class of the dataset yielded 743 cluster centers. The value of k is based on idea of equalizing minority class with majority class and generating 743 clusters for majority class. These points were mixed with the datapoints of the minority class to remove bias, and generate a total of 1486 datapoints. The 24 prediction algorithms were run on this new dataset and the results were validated using 5-fold cross validation. Table .1 shows the validation results.

	ILFS	ECFS	Relieff	FSV	Laplacian	UDFS	LLCFS	CFS	FASAL	Lasso	DGUFS	No feature selection
Fine Decision Tree	0.998, 1, 0.996, 0.998	0.9832, 0.996, 0.9711, 0.9834	0.9879, 0.9812, 0.9945, 0.9878	0.9919, 0.9879, 0.9959, 0.9919	0.9919, 0.9879, 0.9959, 0.9919	0.998, 0.9973, 0.9987, 0.998	1,1,1,1	1,1,1,1	1,1,1,1	0.996, 0.996, 0.996, 0.996	0.9899, 0.9892, 0.9906, 0.9899	1, 1,1,1
Medium Decision Tree	0.998, 1, 0.996, 0.998	0.9832, 0.996, 0.9711, 0.9834	0.9879, 0.9812, 0.9945, 0.9878	0.9919, 0.9879, 0.9959, 0.9919	0.9919, 0.9879, 0.9959, 0.9919	0.998, 0.9973, 0.9987, 0.998	1,1,1,1	1,1,1,1	1,1,1,1	0.996, 0.996, 0.996, 0.996	0.9899, 0.9892, 0.9906, 0.9899	1, 1,1,1
Coarse Decision Tree	0.998, 1, 0.996, 0.998	0.9832, 0.996, 0.9727, 0.9834	0.9657, 0.9583, 0.9727, 0.9654	0.969, 0.969, 0.969, 0.969	0.969, 0.969, 0.969, 0.969	0.9825, 0.965, 1, 0.9822	1,1,1,1	1,1,1,1	1,1,1,1	0.998, 0.996, 1, 0.998	0.9859, 0.9717, 1, 0.9857	1, 1,1,1
Linear Discriminant	0.9643, 1, 0.9334, 0.9656	0.683, 0.961, 0.6716, 0.9834	0.7995, 0.9973, 0.7146, 0.8326	0.998, 1, 0.996, 0.998	0.998, 1, 0.996, 0.998	0.998, 1, 0.996, 0.998	0.9623, 1, 0.9299, 0.9637	0.9314, 0.9358, 1, 0.8793	0.9987, 1, 0.9987, 0.9973	0.8493, 0.8291, 0.8640, 0.8462	0.5599, 0.5882, 0.5567, 0.572	0.9987, 1, 0.9987, 0.9973
Quadratic Discriminant	0.9906, 0.9812, 1, 0.9905	0.6938, 0.9825, 0.6229, 0.7624	0.9246, 0.9650, 0.8929, 0.9276	0.9933, 0.9865, 1, 0.9932	0.9933, 0.9865, 1, 0.9932	0.9926, 0.9854, 1, 0.9927	0.9987, 1, 0.9987, 0.9973	0.9933, 0.9865, 1, 0.9932	0.9933, 0.9865, 1, 0.9932	0.8614, 0.9731, 0.7954, 0.8753	0.9092, 0.9502, 0.8781, 0.9127	0.998, 0.996, 1, 0.998
Logistic Regression	0.9684, 1, 0.9405, 0.9693	0.6992, 0.9987, 0.6246, 0.7685	0.7705, 0.9341, 0.7039, 0.8028	0.9933, 1, 0.9987, 0.9933	0.9993, 1, 0.9987, 0.9993		0.9939, 1, 0.9880, 0.994	0.998, 1, 0.996, 0.998	0.9987, 1, 0.9987, 0.9973	0.9307, 0.9906, 0.8846, 0.9346	0.5673, 0.5882, 0.5646, 0.5761	0.998, 0.996, 1, 0.998
Gaussian Naive Bayes	0.9919, 0.9865, 0.9973, 0.9919	0.9758, 0.9731, 0.9783, 0.9757	0.9098, 0.9664, 0.8682, 0.9146	0.9825, 0.9650, 1, 0.9822	0.9825, 0.965, 1, 0.9822	0.9933, 0.9867, 1, 0.9933	1,1,1,1	0.8163, 0.9919, 0.7341, 0.8437	1,1,1,1	0.7429, 0.9812, 0.6645, 0.7924	0.6703, 0.9664, 0.6069, 0.7456	0.9926, 0.9852, 1, 0.9925
Kernel Naive Bayes	0.998, 0.996, 1, 0.998	0.9791, 0.9785, 0.9798, 0.9791	0.9132, 0.9879, 0.8595, 0.9192	0.9966, 0.9933, 1, 0.9966	0.9966, 0.9933, 1, 0.9966	0.9987, 0.9973, 1, 0.9987	0.9953, 0.9906, 1, 0.9953	0.8183, 0.9973, 0.7344, 0.8459	0.9987, 0.9973, 1, 0.9987	0.7651, 0.9892, 0.6831, 0.8081	0.7544, 0.9892, 0.6731, 0.8011	0.987, 0.9744, 1, 0.987
Linear SVM	0.9933, 1, 0.9867, 0.9933	0. 6965, 0.9919, 0.6235, 0.7657	0.7974, 0.9933, 0.7137, 0.8306	1,1,1,1	1,1,1,1	1,1,1,1	0.9919, 1, 0.9841, 0.992	0.9926, 1, 0.9854, 0.9927	1,1,1,1	0.9374, 1, 0.8888, 0.9411	0.5727, 0.6420, 0.5638, 0.6004	0.9987, 1, 0.9973, 0.9987
Quadratic SVM	0.9926, 1, 0.9854, 0.9927	0. 6964, 0.992, 0.6239, 0.766	0.9805, 0.9919, 0.9697, 0.9807	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0.9906, 0.9946, 0.9866, 0.9906	0.9933, 0.996, 0.9906, 0.9933	1, 1,1,1
Cubic SVM	0.9913, 1, 0.9828, 0.9913	0.6985, 0.9973, 0.6243, 0.7679	0.9906, 0.9946, 0.9866, 0.9906	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0.9872, 0.9946, 0.9801, 0.9873	0.9778, 0.9892, 0.9671, 0.978	1, 1,1,1
Fine Gaussian SVM	0.9953, 0.9906, 1, 0.9953	0.6989, 1, 0.6238, 0.7684	0.9744, 0.9515, 0.9972, 0.9738	0.9852, 0.9704, 1, 0.985	0.9852, 0.9704, 1, 0.9852	0.9966, 0.9933, 1, 0.9966	0.9778, 0.9556, 1, 0.9773	0.9946, 0.9892, 1, 0.9946	1,1,1,1	0.9926, 0.9852, 1, 0.9925	0.9576, 0.9287, 0.9857, 0.9563	0.9845, 0.969, 1, 0.9843
Medium Gaussian SVM	0.9939, 1, 0.988, 0.994	0.70, 1, 0.6249, 0.7692	0.895, 0.9865, 0.8339, 0.9038	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0.9502, 0.9973, 0.9114, 0.9524	0.8801, 0.9098, 0.8439, 0.8756	1, 1,1,1
Coarse Gaussian SVM	0.9933, 1, 0.9867, 0.9933	0.6904, 0.9785, 0.6208, 0.7597	0.8001, 0.9919, 0.7169, 0.8323	1,1,1,1	1,1,1,1	1,1,1,1	0.9872, 1, 0.9751, 0.9874	0.9973, 1, 0.9945, 0.9973	1,1,1,1	0.9374, 1, 0.8888, 0.9411	0.6090, 0.8264, 0.5760, 0.6788	1, 1,1,1
Fine KNN	0.9933, 1, 0.988, 0.994	0. 6999, 1, 0.6249, 0.7692	0.9441, 0.9879, 0.9084, 0.9465	0.9993, 1, 0.9987, 0.9993	0.9919, 0.9879, 0.9959, 0.9919	1,1,1,1	0.998, 1, 0.996, 0.998	1,1,1,1	1,1,1,1	0.9939, 0.9973, 0.9906, 0.994	0.8977, 0.9825, 0.84, 0.9057	1, 1,1,1
Cosine KNN	0.9764, 0.9906, 0.9634, 0.9768	0. 6972, 1, 0.6228, 0.7676	0.8855, 0.9879, 0.7813, 0.8726	1,1,1,1	1,1,1,1	0.998, 0.9973, 0.9987, 0.998	0.9919, 1, 0.9841, 0.992	0.9993, 1, 0.9987, 0.9933	1,1,1,1	0.9724, 0.9892, 0.957, 0.9729	0.7443, 0.9906, 0.6637, 0.7948	0.998, 0.996, 1, 0.998
Medium KNN	0.976, 1, 0.9538, 0.9763	0. 6972, 1, 0.6228, 0.7676	0.8432, 0.9987, 0.7618, 0.8643	0.9973, 1, 0.9946, 0.9973	0.9919, 0.9879, 0.9959, 0.9919	0.998, 0.996, 1, 0.998	0.9865, 1, 0.9738, 0.9867	0.9993, 1, 0.9987, 0.9933	1,1,1,1	0.9711, 0.9973, 0.9476, 0.9718	0.5444, 0.9906, 0.5235, 0.685	1, 1,1,1
Coarse KNN	0.9179, 1, 0.9538, 0.9763	0. 6972, 1, 0.6228, 0.7676	0.7995, 1, 0.7137, 0.8323	0.9913, 1, 0.9828, 0.9913	0.9913, 1, 0.9828, 0.9913	1,1,1,1	0.967, 1, 0.9381, 0.9681	0.9953, 1, 0.9907, 0.9953	1,1,1,1	0.9374, 1, 0.8888, 0.9411	0.7773, 0.9704, 0.7, 0.8133	1, 1,1,1
Cubic KNN	0.9738, 0.996, 0.9536, 0.9743	0. 6972, 1, 0.6228, 0.7676	0.8149, 0.9987, 0.7303, 0.8437	0.9973, 1, 0.9946, 0.9973	0.9973, 1, 0.9946, 0.9973	0.9966, 0.9933, 1, 0.9966	0.961, 1, 0.9276, 0.9624	0.9973, 1, 0.9946, 0.9973	1,1,1,1	0.9563, 0.9973, 0.9216, 0.958	0.6676, 0.9865, 0.6023, 0.748	0.9946, 1, 0.9893, 0.9946
Weighted KNN	0.9872, 1, 0.9751, 0.9874	0. 6985, 1, 0.6238, 0.7684	0.8930, 0.9960, 0.8259, 0.9030	0.9993, 1, 0.9987, 0.9993	0.9993, 1, 0.9987, 0.9993	1,1,1,1	0.9953, 1, 0.9907, 0.9953	1,1,1,1	1,1,1,1	0.9933, 0.9987, 0.988, 0.9933	0.8466, 0.9906, 0.7691, 0.8659	1, 1,1,1
Ensemble Boosted Trees	0.4987, 0.5989, 0.4989, 0.5443	0. 784, 0.5976, 0.9528, 0.7345	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.9951, 1, 0.9952, 0.9976
Ensemble Bag Trees	0.9993, 0.9987, 1, 0.9993	0. 9872, 0.996, 0.9788, 0.9873	0.9939, 0.9892, 0.9986, 0.9939	0.996, 0.9919, 1, 0.9959	0.996, 0.9919, 1, 0.9959	0.9987, 1, 0.9973, 0.9987	1,1,1,1	1,1,1,1	1,1,1,1	0.9953, 0.9946, 0.996, 0.9953	0.9953, 0.9946, 0.996, 0.9953	0.4987, 0.5989, 0.4989, 0.5443
Ensemble Subspace Discriminant	0.969, 1, 0.9417, 0.97	0. 7759, 0.9515, 0.7402, 0.8094	0.7995, 0.9960, 0.7150, 0.8324	0.9993, 1, 0.9987, 0.9993	0.9993, 1, 0.9987, 0.9993	0.9933, 0.9987, 1, 0.9933	0.967, 1, 0.9381, 0.9681	0.9017, 1, 0.8358, 0.9105	1,1,1,1	0.9367, 0.9987, 0.8886, 0.9404	0.5585, 0.5989, 0.5542, 0.5757	0.999, 1, 0.9952, 0.9976
Ensemble subspace KNN	0.9987, 1, 0.9973, 0.9987	0.9132, 1, 0.8521, 0.9201	0.9926, 0.9946, 0.9906, 0.9926	0.9993, 1, 0.9987, 0.9993	0.9993, 1, 0.9987, 0.9993	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0.9926, 0.996, 0.9893, 0.9926	1, 1,1,1
Ensemble RUS Boosted Trees	0.4987, 0.5989, 0.4989, 0.5443	0.9697, 0.996, 0.9463, 0.9705	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.5989, 0.4989, 0.5443	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.3984, 0.4983, 0.4428	0.4987, 0.5989, 0.4989, 0.5443

Table 1: Table showing accuracy, precision, recall and F1-score using 24 prediction algorithms with selected features for the feature selection algorithms for k-means clustering

SMOTE method of oversampling. The minority class was oversampled and the new dataset was run through the algorithms. The highest accuracy was given by Fine Decision Tree of 95.4%.

Downsampling method. The majority class dataset can be downsampled by an integer sampling factor, n. It samples the dataset by keeping the first sample and then every nth sample after that. In case of several columns in the dataset, each column will be treated as a separate sequence. After feeding the downsampled dataset into 24 algorithms, the highest accuracy was reported by *Quadratic SVM of 56.4%* only.

3.3 Feature Selection Methods

Infinite Latent Feature Selection (ILFS) [16]. Consider a training set $X = \{\vec{x}_1, ..., \vec{xn}\}$, such that the distribution of the values assumed by the i^{th} features is given by m 1 vector $\vec{\mathbf{x}}_i$, taking into account m samples. An undirected graph G is formed so that the features are represented by the nodes and the inter- node relationships are represented by the edges. If aij is an element of the adjacency matrix, A associated with G, that represents the pairwise relationship between the features xi and xj ($1 \leq i, j \leq n$). G can be represented by the binary function [16].

$$a_{ij} = \phi\ (x_i, x_j) \qquad (2)$$

where ϕ is a real valued potential function. The probability of each co-occurrence in $\mathbf{x}_i$ and $\mathbf{x}_j$ is framed as a mixture of conditionally independent multinomial distribution, where parameters are learned using Expectation Maximization (EM) algorithm.

Feature selection via Eigenvector Centrality (ECFS). The adjacency matrix of the above graph G can be written as [17].

$$A = \alpha k + (1 - \alpha)\ \ \sum(\mathbf{i, j}) \qquad (3)$$

where $\alpha \in [0, 1]$ is a loading coefficient. In Eigenvector Centrality measure (EC), v_o is calculated as the eigen vector of A associated with the largest eigen value. If e is any vector,

$$\lim_{l \to L} [A^l e] = v_o \qquad (4)$$

Relieff. Relieff is an algorithm developed by Kira and Rendell in 1992 that uses filter-method approach for feature selection. If a dataset consists of n instances of p features, belonging to two classes. At each iteration, X is a feature vector belonging to one random instance and the feature vectors of the instances closest to X from each class using Manhattan L1 norm are chosen. 'Near hit' is the closest instance of the same class and 'Near miss' is the closest instance of different class. The weight vector is updated as follows [1].

$$W_i = W_i - (x_i - \text{nearHit}_i)^2 - (x_i - \text{nearM iss}_i)^2 \qquad (5)$$

Feature Selection Concave (FSV). If matrices $\in R^{mxn}$ and $B \in R^{kxn}$ are two point sets, then they can be discriminated by a separating plane, P as in [18].

$$P = \{\ x | x \in R^n, x^T w = \gamma\ \} \qquad (6)$$

where normal $w \in R^n$ and 1-norm distance to the origin is defined as $\frac{|\gamma|}{\|w\|_{inf}}$

Laplacian. A parameter used in this algorithm is the Laplacian Score (LS) which means that two points

are related to the same topic if they are close to each other. Laplacian score of the r^{th} feature is calculated as follows [19].

$$L_r = \frac{\tilde{f}_r^T L \tilde{f}_r}{\tilde{f}_r^T D \tilde{f}_r} \qquad (7)$$

Unsupervised Discriminative Feature Selection (UDFS). UDFS aims to select the most discrim- inative features for data representation, where manifold structure is considered. $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ is the training set, $x_i \in R^d$ ($1 < i < n$) is the i^{th} datum and n is the number of data points in the training set. The objective function of this algorithm is: For an arbitrary matrix, $A \in R^{r \times p}$, its $l_{2,1}$-norm [20]. Is

$$\|A\|_{2,1} = \sum_{i=1}^{R} \sqrt{\sum_{j=1}^{p} A_{ij}^2} \qquad (8)$$

Local Learning Clustering based Feature Selection (LLCFS). This algorithm constructs the *k*-nearest neighbor graph in the weighted feature space. It performs joint clustering and feature weight learning [21].

Correlation based Feature Selection (CFS). This algorithm performs feature selection on the basis of the hypothesis,"good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" [22]. A merit function is a function that measures the agreement between data and the fitting model, for a particular choice of parameters. By definition, the merit function is small when the agreement is good. The merit function of a feature subset S consisting of *k* features is given as [1].

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (9)$$

Where is the mean of all feature-classification correlations, and rff is the mean of all feature-feature correlations.

Feature Selection with Adaptive Structure Learning (FSASL). In this algorithm the features are ranked in descending order of their weights [21]. The optimal sparse combination weight matrix $S\ R^{n \times n}$ can be obtained by solving the following problem [23].

$$\min_S \sum_{i=1}^{n} \left(\|x_i - X s_i\|^2 + \alpha \|s_i\|_1 \right), such\ \ that\ \ S_{ii} = 0 \qquad (10)$$

Lasso. Consider a sample consisting of N cases, each of which consists of *p* covariates and a single outcome. If y_i be the outcome and $x_i = (x_1, x_2, ..., x_p)^T$ be the covariate vector for the i^{th} case. The parameters are estimated by solving the following optimization problem [24].

$$\widehat{\beta}(\lambda) = \arg\min_{\beta} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right) \quad (11)$$

where $\|Y - X\beta\|_2^2 = \sum_{i=0}^{n}(Y_i - (X\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$. $\lambda \geq 0$, is the parameter that controls the strength of the penalty.

Dependence Guided Unsupervised Feature Selection (DGUFS). The objective of this algorithm is to select *m* most discriminative features (m < d) whose learned pseudo-label indicators are much closer to the cluster groups. It can be stated by the following discrimination promotion function [25].

$$\min_{Y,V} J(X, V, Y), \quad (12)$$

where Y = diag(s)X, V ∈ Ω, S ∈ $\{0, 1\}^d$, $s^T 1_d$ = m, Ω=candidate set of data cluster label matrices.

3.4 Classification Methods

Bayesian network. A Bayesian network is a directed acyclic graph with some quantitative probability information assigned to each node that corresponds to a random variable. It has many other synonyms, viz, belief networks, probabilistic network, causal network and knowledge map [26]. A conditional probability distribution P (x_i parents (X_i)) defines the relationship between each node and its parents. It is defined by the following equation [26].

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i)) \quad (13)$$

where P (x_1,.....x_n) = probability of joint conjunction of events x_1, x_2,……x_n.

Decision tree. The goal attribute is true if and only if the input attributes follow the paths towards a leaf with value true. This assertion gives a decision tree and its propositional logic can be written as follows [26].

$$\text{Goal} \Leftrightarrow \text{P ath}_1 \text{V P ath}_2 \text{V...} \quad (14)$$

In MATLAB definition, *fine trees* have the highest model flexibility as they have many leaves to make many fine distinctions between classes [27]. They allow a maximum of 100 splits. In case of medium trees, the model flexibility is *medium*. They allow a maximum of 20 splits. In case of *coarse* trees, the model flexibility is low and they allow a maximum of 4 splits.

Logistic regression. The logistic function is given by the following equation [26].

$$Logistic(z) = \frac{1}{1 + e^{-z}} \quad (15)$$

It gives the *probability* of belonging to the class labeled 1. The process of fitting the weights of this model to minimize loss on a data set is called *logistic regression*.

RUS boosted trees. Random under-sampling (RUS) is used to balance an imbalanced class that
is a common problem for any datasets having rare occurrences of a particular event, the algorithm of which can be found I [28].

Bagged trees. "Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predicton"- Breiman [29].

Consider data (y_n, x_n), n = 1..., N in a learning set, where the *y's* are either class labels or a numerical response. If the input is *x* we predict y by ϕ (x, L), taking repeated bootstrap samples L^B from L, and forming ϕ (x, L^B) and if *y* is numerical

$$\phi_B(x) = av_b \phi\ (x, L^{(Bj)}) \quad (16)$$

If y is a class label, let the ϕ(x, L^B) vote to form ϕ_B(x). This is called "bootstrap aggregating" or bagging. k-**means clustering**. In k-means clustering [30, 1]. the data (x_1, x_2, ..., x_n) with n observations is segregated into k clusters (k ≤ n), S = S_1, S_2, ..., S_k so that the *within cluster sum of squares* (WCSS) or
variance is minimum.

$$\arg\min_{s} \sum_{j}^{k} \sum_{x \epsilon S_i} \|x - \mu_i\|^2 = \arg\min_{s} \sum_{i=1}^{k} |S_i| \text{ Var } S_i \quad (17)$$

where μ_i is the centroid of cluster S_i imbalanced dataset by increasing the number of samples of the minority class. The algorithm flowchart is described in detail in [31].

Support Vector Machine. SVM is a type of supervised learning, where data that is not linearly separable can be easily separated by mapping them into higher dimensional space. The optimal SVM separator is found by solving the following [26].

$$\arg \max_{a} \sum_{j} \alpha_j - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_j y_j y_k (x_j . x_k)) \quad (18)$$

where α_j 0 and $\alpha_j y_j = 0$. Solution of this equation is done using software called as quadratic programming.

K-nearest neighbor KNN algorithm classification is a type of clustering where the nearest *k* datapoints $NN(k, x_q)$ are considered. The distance metric is measured using Minkowski distance as follows [26].

$$L^p(x_j, x_q) = (\sum_{i} |x_{j,i} - x_{q,i}|^p)^{\frac{1}{p}} \quad (19)$$

When p = 2, it is called Euclidean distance and if p = 1, it is Manhattan distance.

3.5 Evaluation Matrices

The statistical parameters calculated are as follows [1].

$$Accuracy = \frac{t_p + t_n}{total} \quad (20)$$

$$Precision = \frac{t_p}{t_p + f_p} \quad (21)$$

$$Recall = \frac{t_p}{t_p + f_n} \quad (22)$$

$$F1 - Score = \frac{2 \text{ x } Precision \text{ x } Recall}{Precision + Recall} \quad (23)$$

where t_p, t_n, f_p, f_n are the number of true positives, true negatives, false positives, false negatives respectively.

4. Experimental Results

4.1 Dataset

The dataset used for this work is the Pancreas Cancer Dataset made accessible by the National Cancer Institute by NIH. Around 155,000 individuals have participated in the PLCO data collection [32]. Each of them have filled out three questionnaires-the *Baseline Questionnaire (male-BQM/female-BQF), Other Cancer Form (OCF), and the Annual Study Update (ASU)* form. Figure 2 shows the 2D representation using AdaSyn algorithm and t-SNE algorithm [14] for the down-sampled dataset using selective sampling of majority class (PC=0), down-sampled dataset using *k*-means clustering and oversampled minority class(PC=1) of dataset using SMOTE algorithm.

4.2 Classification

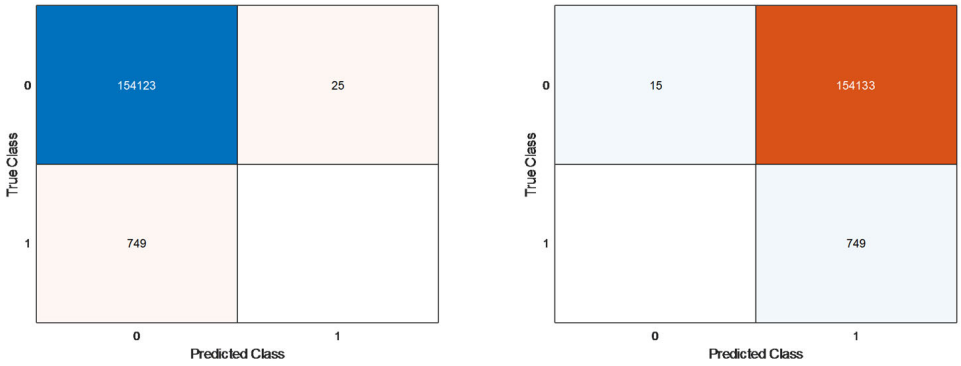
Twenty four machine learning algorithms, briefly described in Section 3 were used and their statistical parameters are reported.

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 *ij* (i, j 1...c , *ii*= 0) is the cost of misclassification of the example of the $\mathbf{i}_{th}$ class to the $\mathbf{j}_{th}$ class, where c is the number of classes. Then, the weight assigned to the $\mathbf{i}_{th}$ class after rescaling is given as [33].

$$w_i = \frac{n \text{ x } \in_i}{\sum_{k=1}^{c} (n_k \text{ x } \in_k)} \quad (24)$$

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

It uses the algorithms as described in [33–35]. 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. Figure 1a shows confusion matrix without any weights where it improperly classified all PC cases and Figure 1b shows a very high weight applied to false positives which leads to misclassification of all non-PC cases.



(a) (b)

Figure 1: Figure showing confusion matrix using classification ensemble for (a) no weight, and (b) very high weights applied to False Positive.

4.3 Feature Selection

Several feature selection algorithms [17–22, 24, 25, 36–38] have been used to extract features. These algorithms have been implemented in MATLAB. Table .2 shows the features selected by these algorithms. *D=downsampled dataset using fixed sampling rate, K=downsampled dataset using k-means clustering.* Figure 3 and figure 4 shows the varaiation of P(C—E=true) where P(E)=probability of the feature, choosing some of the most probable features for Pancreatic Cancer from Table 3.

	ILFS	ECFS	Relieff	FSV	Laplacian	UDFS	LLCFS	CFS	FSASL	Lasso	DGUFS
Education		D	D,K	K	D		D	D	D		D,K
Marital Status	D		K	K			D				D,K
Occupation	D		K	K			D	K	D		D,K
Smoked Pipe			K	K				K			D,K
Smoked Cigar			K	K				K			D,K
No of Sisters	D	D	K	K			K				D,K
No of Brothers		D	K	K			D,K	D			D,K
Use of Aspirin regularly	K	K	K				D	D			D,K
Use of Ibuprofen regularly	K	K	K	K						K	D
No of Tubal/ Ectopic pregnancies			K						D,K		D
Ever tubes tied?								D	K		
Ever had fibrocystic breast disease?					K				K		
Ever had Ovarian tumor/cyst?				D							
Ever had endometriosis?								K			
Ever had uterine fibroid tumors?											
Ever tried to become pregnant for a year or so without success?									K		
No of pregnancies?	D	D			D,K		D	D,K	D		
No of still births?				K							
Age at hysterectomy					K		K				
Age started taking birth control pills			D	D			K			D	
Taken female years hormones						K		K			
Total years taken female hormones	D	D			D,K		D				
Age at when started to urinate more than once in night	D,K	D,K	D		D		K	K			
Age when told had enlarged prostate	D,K	D,K			D		K	K	K	D,K	
Age when told had inflammed prostate	K	K					K				
Age at vasectomy	K	K	D				K				
Diagnosed with hypertension	K	K								D	
Heart Attack	K	K						D		D	
Stroke			D								
Emphysema										K	K
Bronchitis			D	D							K
Diabetes				D							
Colorectal Polyps						K				K	
Arthritis	K	K									
Osteoporosis											
Diverculitis											
Gall bladder stone or inflammation?			D	D						K	
Race					D		D		D		
Are you of Hispanic origin											
Ever had biopsy of prostrate?						D,K				D	
Ever had transurethral resection of prostate?						D,K				D	
Ever had prostatetomy of benign disease?				D,K		D,K			K		
Ever had any prostate surgeries						D,K			K		
Ever been pregnant?	D				D,K	D		D,K	D,K		
Ever had hysterectomy?					K						
Removed ovaries?					K			D	K		
Ever had enlarged prostate?						D,K			K		
Ever had inflamed prostate?											
Have problem with prostate?						K					
During past year, how many times wake up in the night to urinate?	D				D		D				
Ever had vasectomy?				D		D,K					
Ever take birth control pills?				D	K			K			
Ever take female hormones?	D				K			D			
Ever smoke regularly >6 months?			D	D		K				D,K	
Smoke regularly now?			D								
Usually filtered or not filtered?			D				K				
Colon comorbidities											
Liver comorbidities										D,K	
Family history of pancreatic cancer			D							K	
No of relatives with pancreatic cancer		D				D	K	D			
Prior history of any cancer						D				K	
Prior history of pancreatic cancer		D			D			D		D	
Gender		D			D			D			
No of cigarettes smoked daily	K	K		D		D,K		D		K	
Family history of any cancer							D				

Table 2: Features(rows)selected by the different algorithms(columns) highlighting most selected features >=6

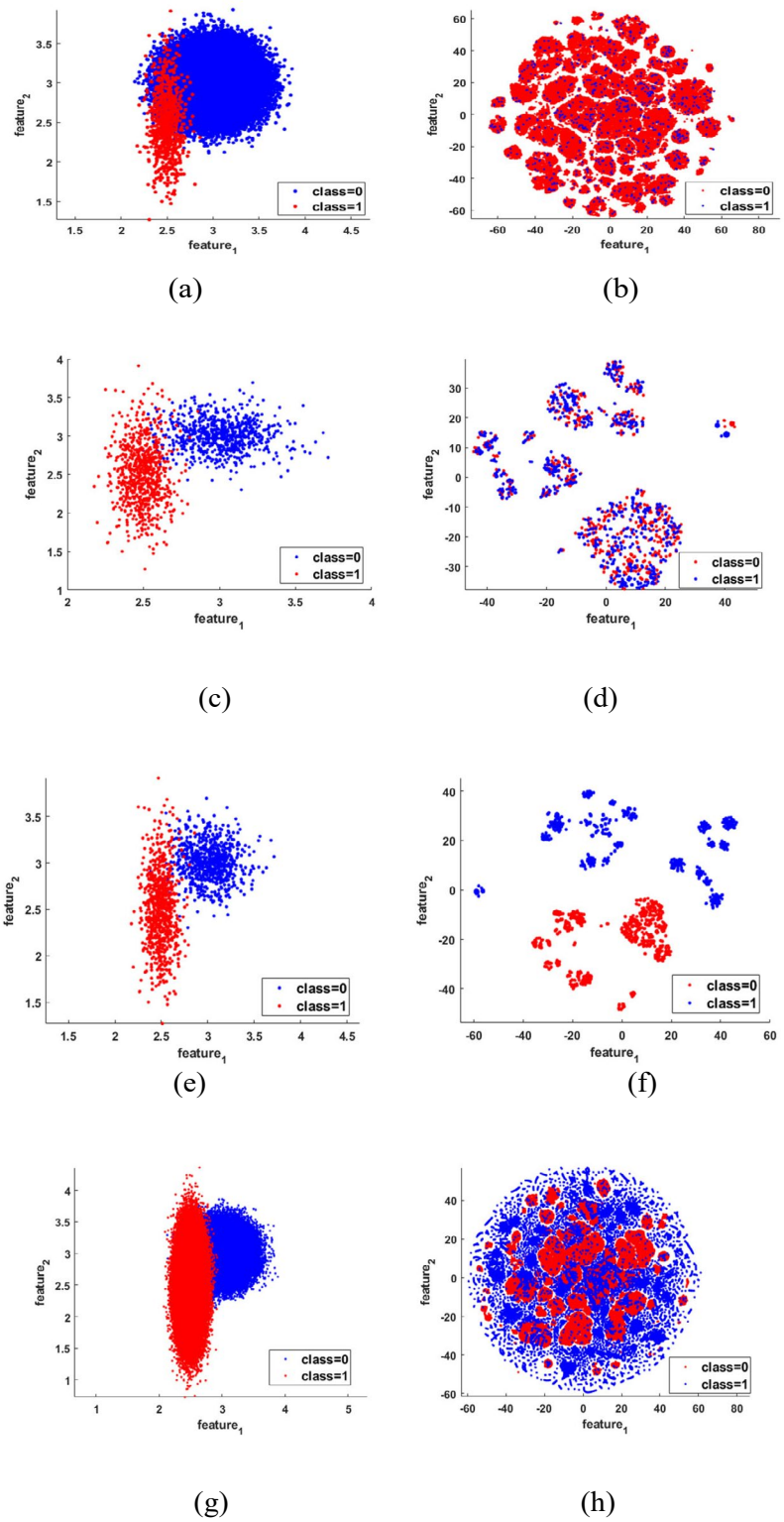


Figure 2: PLCO dataset visualized in 2D using ADASYN algorithm (left) and t-SNE algorithm (right). (First row) Original dataset (total 154,897 points; *False* class 154,148; *True* class 749). (Second row) Balanced dataset using fixed-rate down- sampling (total 1486 points; *False* class 743; *True* class 743). (Third row) Balanced dataset by downsampling using *k*-means clustering (total 1486 points; *False* class 743; *True* class 743). (Fourth row) Balanced dataset by oversampling using SMOTE algorithm (total 1486 points; *False* class 743; True class 743).

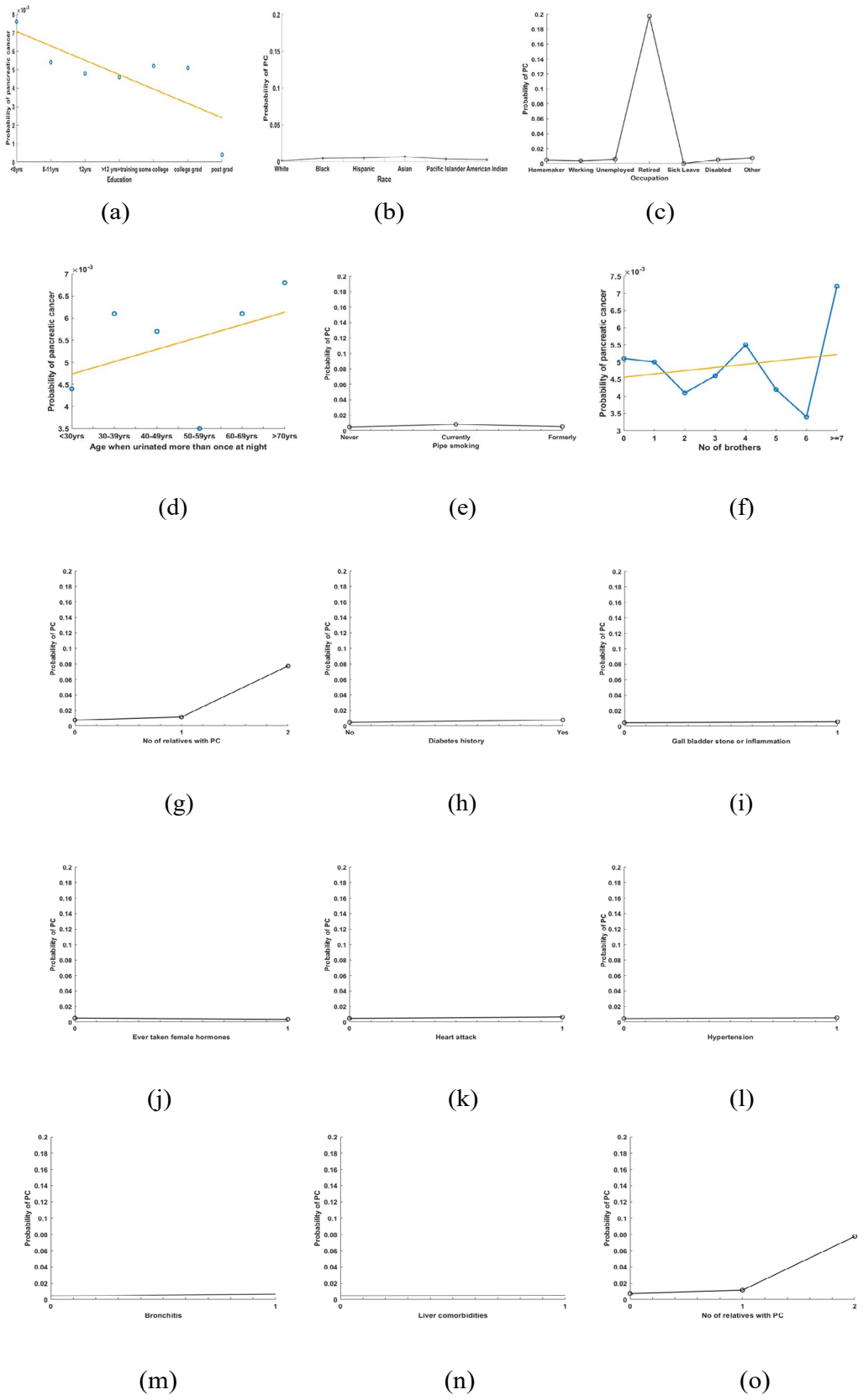


Figure 3: Figure showing variation with *P (C|E = true)* using the most frequently chosen features from Table 3.

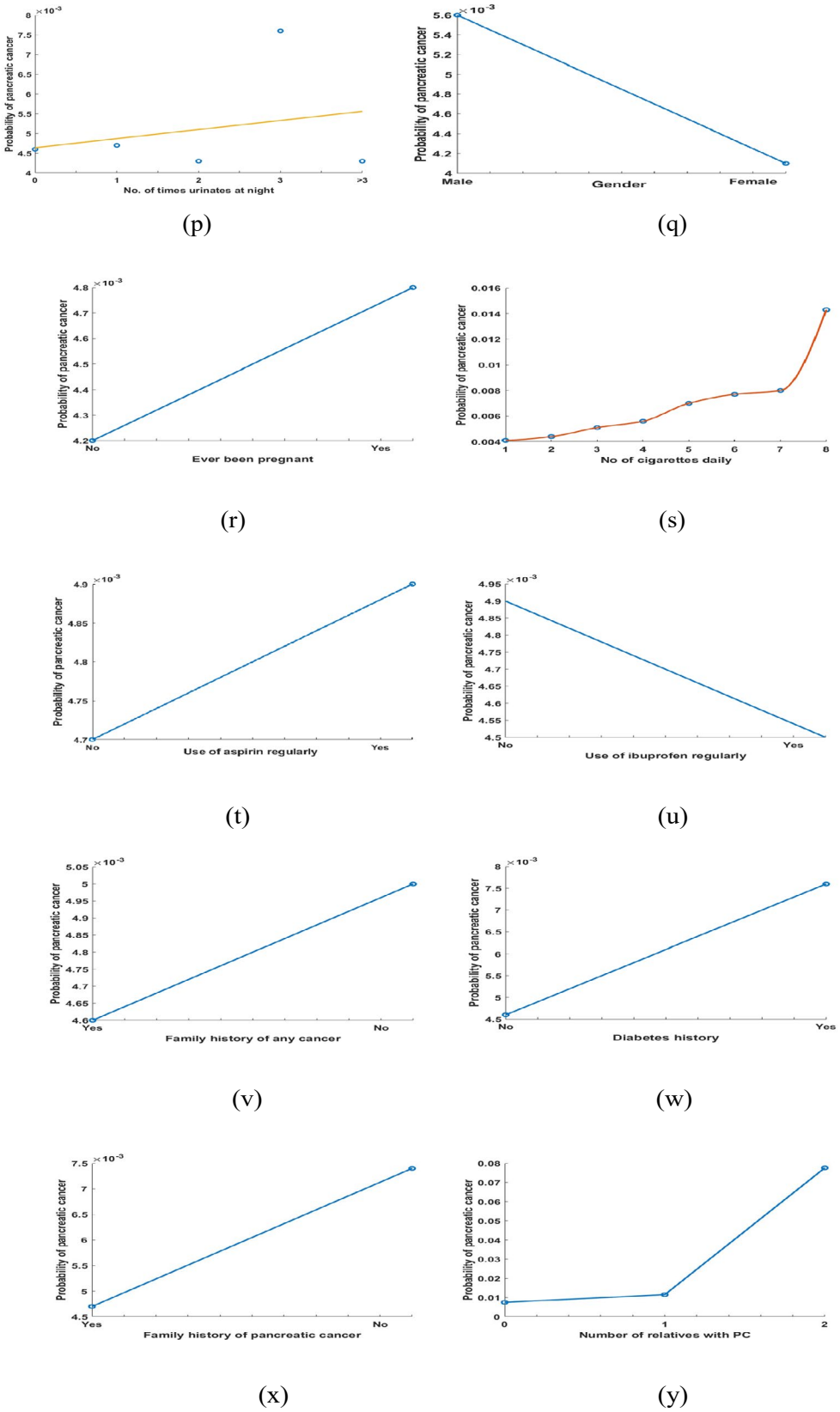


Figure 3: Figure showing variation with *P (C|E = true)* using the most frequently chosen features from Table 2. The orange line is a fitting curve

5 Conclusion

Certain features have been identified to have a direct relationship with pancreatic cancer, for example, smoking history, no. of cigarettes smoked in a day, genetics etc, whereas others have been identified by features selection algorithms and also by graphical representation to have an unproved connection with causing Pancreatic Cancer, for example, no. of brothers, total years taken female hormones [13]. Future work remains in order to find a mechanism in which a robot can predict with the highest accuracy the probability of a person having pancreatic cancer by getting answers to a set of features from the subject.

Acknowledgements

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.

References

1. Wikipedia contributors. Wikipedia, the free encyclopedia. https://en.wikipedia.org/, 2019. [Online; accessed 25-August-2019].
2. JANG Ik-Gyu. 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, January 31 2019. US Patent App. 15/833,828.
3. Huxley, R., Ansary-Moghaddam, A., Berrington de Gonzalez, A., Barzi, F., & Woodward, M. (2005). Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. British journal of cancer, 92(11), 2076-2083.
4. Everhart, J., & Wright, D. (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., Huang, W., ... & Li, Z. (2011). Diabetes mellitus and risk of pancreatic cancer: a meta-analysis of cohort studies. European journal of cancer, 47(13), 1928-1937.
6. Jones, S., Hruban, R. H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D. W., ... & Klein, A. P. (2009). Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science, 324(5924), 217-217.
7. CM, B., SL, S., CM, H., PA, H., C O'Sullivan, G., Russell, R. C. G., ... & Williamson, R. C. N. (1991). Abnormalities of the p53 tumour suppressor gene in human pancreatic. Br. J. Cancer, 64, 1076-1082.
8. Iacobuzio-Donahue, C. A., Fu, B., Yachida, S., Luo, M., Abe, H., Henderson, C. M., ... & Laheru, D. (2009). DPC4 gene status of the primary carcinoma correlates with patterns of failure in patients with pancreatic cancer. Journal of clinical oncology, 27(11), 1806.
9. Das, A., Nguyen, C. C., Li, F., & Li, B. (2008). Digital image analysis of EUS images accurately differentiates pancreatic cancer from chronic pancreatitis and normal tissue. Gastrointestinal endoscopy, 67(6), 861-867.
10. Ge, G., & Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC bioinformatics, 9, 1-12.
11. Săftoiu, A., Vilmann, P., Gorunescu, F., Gheonea, D. I., Gorunescu, M., Ciurea, T., ... & Iordache, S. (2008). Neural network analysis of dynamic sequences of EUS elastography used for the differential diagnosis of chronic pancreatitis and pancreatic cancer. Gastrointestinal endoscopy, 68(6), 1086-1094.
12. Zhang, M. M., Yang, H., Jin, Z. D., Yu, J. G., Cai, Z. Y., & Li, Z. S. (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. Gastrointestinal endoscopy, 72(5), 978-985.
13. Dutta, A. (2023). Using Machine Learning to Identify the Risk Factors of Pancreatic Cancer from the NIH PLCO Dataset.
14. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605.
15. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). Ieee.
16. Roffo, G., Melzi, S., & Cristani, M. (2015). Infinite feature selection. In Proceedings of the IEEE international conference on computer vision (pp. 4202-4210).
17. Roffo, G., & Melzi, S. (2017). Ranking to learn: Feature ranking and selection via eigenvector centrality. In New Frontiers in Mining Complex Patterns: 5th International Workshop, NFMCP 2016, Held in Conjunction with ECML-PKDD 2016, Riva del Garda, Italy, September 19, 2016, Revised Selected Papers 5 (pp. 19-35). Springer International Publishing.
18. Mangasarian, O. L., & Bradley, P. S. (1998). Feature Selection via Concave Minimization and Support Vector Machines.
19. He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection. Advances in neural information processing systems, 18. pages 507–514.
20. Yang, Y., Shen, H. T., Ma, Z., Huang, Z., & Zhou, X. (2011, December). ℓ 2, 1-norm regularized discriminative feature selection for unsupervised learning. In IJCAI international joint conference on artificial intelligence.
21. Du, L., & Shen, Y. D. (2015, August). Unsupervised feature selection with adaptive structure learning. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 209-218).
22. Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).
23. Mathworld, W. (2019). The Web's Most Extensive Mathematics Resource.
24. Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam research paper in business analytics, 30, 1-25.

25. Guo, J., & Zhu, W. (2018, April). Dependence guided unsupervised feature selection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
26. Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.
27. Mathworks.com. (2019). Train models to classify data using supervised machine learning - MATLAB. https://www.mathworks.com/help/stats/classificationlearner-app.html, 2019. [Online, accessed 25-August-2019].
28. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE transactions on systems, man, and cybernetics-part A: systems and humans, 40(1), 185-197.
29. Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.
30. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).
31. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
32. Biometry.nci.nih.gov. (2019). Pancreas-Datasets-PLCO-The Cancer Data Access System. https://biometry.nci.nih. gov/cdas/datasets/plco/10/, 2019. [Online; accessed 25-August-2019].
33. Zhou, Z. H., & Liu, X. Y. (2010). On multi-class cost-sensitive learning. Computational Intelligence, 26(3), 232-257.
34. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. belmont, ca: Wadsworth. International Group, 432:151–166.
35. Zadrozny, B., Langford, J., & Abe, N. (2003, November). Cost-sensitive learning by cost-proportionate example weighting. In Third IEEE international conference on data mining (pp. 435-442). IEEE.
36. Roffo, G. (2017). Ranking to learn and learning to rank: On the role of ranking in pattern recognition applications. arXiv preprint arXiv:1706.05933.
37. Roffo, G., Melzi, S., Castellani, U., & Vinciarelli, A. (2017). Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In Proceedings of the IEEE international conference on computer vision (pp. 1398-1406).
38. Baine, M., Sahak, F., Lin, C., Chakraborty, S., Lyden, E., & Batra, S. K. (2011). Marital status and survival in pancreatic cancer patients: a SEER based analysis. PloS one, 6(6), e21052.