*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 11 Issue 1 ǁ January 2023 ǁ PP. 157-164*

# A Machine Learning Approach to Early Detection of Pancreatic Cancer using Urinary Biomarkers

Sarvesh.E
[*1]*Department of Computer Science & Engineering, College of Engineering Guindy, Anna Universuty, Chennai, India.*

***Abstract***— *Among the deadly cancers in the world, early diagnosis of Pancreatic ductal adenocarcinoma (PDAC) is challenging due to various reasons. Early detection improves the survival rate in cancer. This paper suggests various machine learning algorithms that help in the prediction of pancreatic cancer at an early stage. Based on the level of urinary biomarkers, various machine learning techniques are used to improve the diagnosis of pancreatic cancer. The study focuses on analyzing different classification models and finding the best among them to classify the dataset. It has been identified that the Random Forest classifier and the Gradient Boosting classifier significantly outperformed other algorithms. There is a pressing need to create newer diagnostic techniques for pancreatic cancer that can address several unmet clinical requirements.*
***Keywords***— *Pancreatic ductal adenocarcinoma, urinary biomarkers, machine learning, prediction algorithms, Logistic Regression, Random forest classifier, Gradient boosting classifier.*

---

Date of Submission: 02-01-2023 Date of acceptance: 14-01-2023

---

## I. INTRODUCTION

Pancreatic cancer is an important public health issue. Due to its high mortality rate, pancreatic cancer is considered the most dreadful of all gastrointestinal cancers. It has a poor 5-year survival rate, which makes it the seventh leading cause of death of all cancer-related mortality in the world. By 2030, pancreatic ductal adenocarcinoma (PDAC) is expected to become the second leading cause of cancer deaths in the world [1,2,3]. Because of its anatomical positioning, pancreatic cancer is usually not diagnosed until later stages, when it has spread to other areas, because reliable screening tools are lacking. Even our recent imaging techniques are not sensitive enough to aid in the diagnosis. Also, the patients don't have much symptoms, which prevents the clinicians from diagnosing the cancer early. More than 90% of the pancreatic cancer cases are identified as PDAC. Early diagnosis helps the clinicians provide better treatment modalities and hence, better outcomes.

Traditionally, specific biomarkers in the blood were used to identify diseases of particular organs. Recently, serum and urinary biomarkers emerged as significant factors that aid in the diagnosis of pancreatic cancer. Identification of urinary biomarkers is a convenient alternative as well as a non invasive method of detection of pancreatic cancer. Creatinine, LYVE1, REG1A, and TFF1 are the protein biomarkers in urine that show promising results in the identification of pancreatic cancer. As the diagnosis of pancreatic cancer is very challenging, it has been tried to find the appropriate classification algorithm to help in the diagnosis of pancreatic cancer in the study [4].

## II. MATERIALS AND METHODS

### *A.* Methodology

Popular algorithms like logistic regression, K-nearest neighbor (KNN), support vector machine (SVM), decision tree, random forest, multilayer perceptron (MLP), naive bayes, and gradient boosting were used for the study. Based on the performance metrics of these algorithms, the best among them was selected.

To improve the prediction accuracy, relevant features were first selected using Chi-squared feature selection. Then, data preprocessing and feature scaling were performed. After these steps, the data set was given as input to the classification model, which gives the desired output. Fig. 1 shows the proposed flowchart of the study.

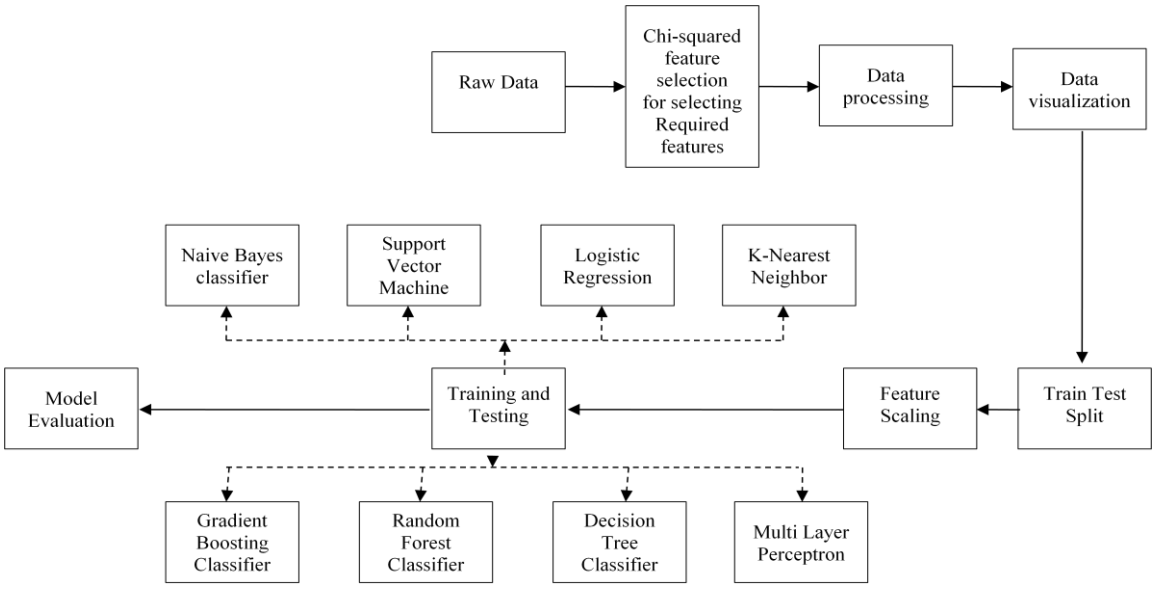


**Figure1: Flowchart of the study**

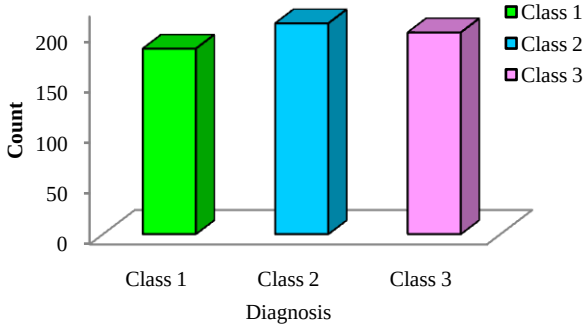


**Figure2: Frequency of data points in each class**

### *B.* Data set source

Urinary biomarkers namely Creatinine, LYVE1, REG1B, TFF1 were used in the diagnosis of pancreatic cancer. The samples were collected from 183 healthy subjects, 208 patients with non-cancerous pancreatic conditions like chronic pancreatitis, and 199 patients with PDAC. Fig. 2 shows the frequency of data points in each class. The original dataset contains 13 input features and 1 output feature [5].

### *C.* Feature Selection

Feature selection refers to the process of identifying relevant features from the entire dataset and excluding the remaining inappropriate features. It simplifies the model by reducing the dimensions of the dataset. Hence, the Chi-squared test for feature selection was used for choosing the appropriate features. The selected features from the dataset include

- Patient's Cohort.
- Sample Origin.
- Age.
- Sex.
- Plasma CA19-9 U/ml.
- Creatinine mg/ml.
- LYVE1 mg/ml.
- REG1B mg/ml.
- TFF1 mg/ml.
- REG1A mg/ml.

### *D.* Data Preprocessing

This data set contains a few missing values in a few columns. The missing value replacement has been performed for all numerical variables using the linear interpolation method. Features with more than 50% missing values was omitted.

### *E.* Prediction algorithms

Input data was provided to the classifier and the output will be 1, 2, or 3 depending on the nature of the sample. The output of the prediction algorithm represents the class to which the given input belongs. The value of 1 denotes healthy patients, 2 for non-cancerous pancreatic conditions like chronic pancreatitis, and 3 for patients with PDAC.

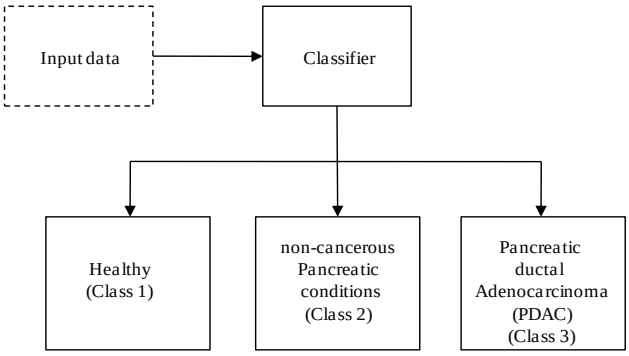


**Figure 3: Classification model**

In this study, 8 different classification algorithms were used. Classification model steps were shown as in Fig. 3.

*a)* **Logistic Regression:** Logistic regression is a supervised learning algorithm that can be used for classification tasks. It is mainly used when there is a need to predict a binary outcome based on the independent variables present in the dataset. In linear regression, based on the given inputs, the output will be a continuous variable. But in logistic regression, the predicted values are mapped to probabilities using the sigmoid function. Logistic regression shows good performance with linearly separable classes. The outcome of logistic regression lies between 0 and 1, as it is a probability. In binary logistic regression, the outcomes are restricted to two classes. In multinomial logistic regression, the outcome will be drawn from a finite set of categories. The study employs multinomial logistic regression for classifying healthy patients, patients with non-cancerous pancreatic conditions and patients with pancreatic ductal adenocarcinoma [6].

*b)* **K-nearest neighbor (KNN):** KNN is a supervised learning technique that is used for classification tasks. It is a non-parametric algorithm, as it learns from the data rather than trying to use mathematical models to establish relations between inputs and outputs like parametric algorithms. Based on the similarity between the new data point and available labeled data, the new data point is classified in an appropriate category by considering the neighboring training examples in that given region. During training, not much is done other than storing and memorizing the training dataset. Due to this reason it is called lazy learning algorithm. It is computationally expensive because it does not make any generalizations in the training phase and hence require the entire dataset for training [7].

*c)* **Support vector machine (SVM):** The principle of structural risk minimization (SRM) is used in SVM. In SVM, the main objective is to separate the different classes present in the training dataset and find a hyper plane that maximizes the margin between them. It is also called the "optimal separable hyper plane" [8].

The dimension of the higher-dimensional plane is equal to the dimension of the feature vector of the dataset. SVM can handle multiple continuous and categorical variables. This algorithm aims to minimize the generalization error. The generalization capacity of the model increases with a decrease in support vectors. Support vector machine (SVM) is a type of linear model that uses a squared L2 regularization term to prevent overfitting and promote a more generalizable solution. In cases where the number of dimensions (features) in the data is greater than the number of training examples, SVM may still be able to find a unique solution, making it effective for classification tasks in higher-dimensional spaces under this condition.

SVM can be used for classification, regression, clustering, and time series analysis. There are two kinds of classifiers used in SVM. They are: Linear SVM and Non-linear SVM. Linear SVM is used for linear classification. Non-linear SVM is used for non-linear classification, and it uses kernels to achieve this. There are a few kernel functions, namely the polynomial kernel function, the RBF (radial basis function) kernel function, and the Sigmoid kernel function [9].

*d)* **Decision Tree Classifier:** As the name suggests, it has a tree-like structure. The internal nodes in the tree represent the features of the dataset. Branches in the tree represent the outcome of the test, and the final outcomes are represented in leaf nodes. It recursively partitions the dataset such that the resulting data items belong to a particular class. Based on the impurity measures, the decision to best split is made at every internal node.

There are two phases in the decision tree classifier: tree building and tree pruning. In tree building, the dataset is partitioned in such a way that all the features belong to a particular set. Since we are traversing the training dataset repeatedly for tree building, it demands a lot of computation. In tree pruning, over fitting in the decision tree is minimized. Since we are traversing the training dataset only once for tree pruning, it requires less computational power compared to the tree building stage. It is used widely because it is more efficient and robust. It is easy to understand as it is similar to decision making by humans [10, 11].

*e)* **Random Forest:** Random forest comprises of many decision trees whose individual outputs are combined to determine the final output based on the majority votes. Its performance improvement is good compared to single tree classifiers. It is more robust to noise [12]. It is clear that the performance of the Random Forest algorithm improves with an increase in the size of the dataset. The Random Forest algorithm performs better with larger datasets compared to smaller ones [13].

*f)* **Gradient Boost:** Ensemble learning is based on the idea that, rather than building a single model and trying to improve its performance, an alternate approach can be used by combining many weak and simpler models. One such example is Random Forest. In Random forest, the averaging of the models in the ensemble takes place. Boosting is based on the idea of adding new models to the ensemble sequentially at each step.

Gradient boosting methods have connections with statistical frameworks, which were missing in the primary boosting techniques, where the entire technique was algorithm driven and complex. Gradient boosting based formulation of boosting methods provides justification for the model's hyper parameters, which were not given in the primary boosting techniques. Gradient boosting techniques are more powerful and accurate compared to other models. As it is highly customizable, it can be altered based on our needs. Gradient boosting machines are prone to over fitting. It can be solved using sub sampling, shrinkage, early stopping, and hence the generalization properties of the model is improved [14, 15].

*g)* **Multilayer Perceptron (MLP):** Unlike other machine learning algorithms, MLP classifier relies on the idea of neural networks. Artificial neural networks are made up of artificial neurons, which interact in a similar way as biological neurons do. Neurons are linked to one another to generate and share new knowledge. Each neuron has a single threshold value. In case of MLP Neural networks, each unit performs biased weighted sum of inputs and give that to the activation function to generate output. The activation of each neuron is calculated by the difference between weighted sum of its inputs and threshold of the neurons. Few examples of activation functions are logistic, hyperbolic, tangent, and sigmoid [16].

*h)* **Gaussian Naïve Bayes:** The term "Naïve Bayes classifiers" refers to a set of classification algorithms based on Bayes's theorem. It is based on the premise that each pair of features being classified is distinct from the others. Naïve Bayes assumes that each feature contributes to the outcome independently and equally. Continuous values associated with each feature are assumed to follow the normal distribution in Gaussian Naïve Bayes. When plotted graphically, it yields a symmetric bell-shaped curve [17].

### *F.* Classification using Classifiers

For training the model, 80% of the data is used for training and the remaining 20% is used for testing the model. In this stage, the testing data has to be classified by assigning the labels 1, 2, or 3.

### *G.* Evaluation Criterion

The accuracy, precision, recall, and f1-score were calculated to assess the performance of various classifiers in classifying pancreatic cancer. The confusion matrix was represented in Table 1. Accuracy is defined as the proportion of true instances obtained, which includes both positive and negative instances, out of all instances retrieved. Precision is the ratio of samples which are correctly classified as positive to the samples which are both correctly and incorrectly classified as positive. The ratio of the samples which are correctly classified as positive to the actual number of positive instances is defined as Recall.

**TABLE 1: CONFUSION MATRIX**

| | ***Actual Positive*** | ***Actual Negative*** |
|---|---|---|
| Predicted Positive | True Positives | False Positives |
| Predicted Negative | False Negatives | True Negatives |

Formulas for these performance measures can be given derived from the confusion matrix and they were given as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = \frac{2\times recall \times precision}{recall + precision} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

Consider the class of healthy patients. In this case, True positives (TP) refer to the number of samples that are actually healthy and are classified as such. False positives (FP) refer to the number of people who are unhealthy (both non-cancerous pancreatic conditions and PDAC) but are mistakenly classified as healthy. The number of people who are healthy but are classified as unhealthy by the model is referred to as false negatives (FN). True negatives (TN) denote the number of unhealthy individuals who are correctly classified as unhealthy [18].

## III. RESULT AND DISCUSSION

The number of neighbors to be considered for the classification was determined by the value of "K." The algorithm is more vulnerable to noise if the value of 'K' is very small. It was evident from Fig. 4 that the optimal value of K is 3. For the K value of 3, the error rate was minimal compared to other values of K.

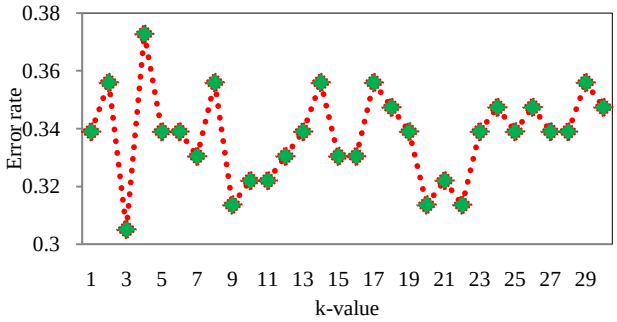


**Figure 4: Error rate vs. K value (for KNN algorithm)**

**TABLE 2: RECALL VALUES**

| Classifier | Classes | | | Overall |
|---|---|---|---|---|
| | *Healthy* | *Non-PDAC* | *PDAC* | |
| Logistic Regression | 0.94 | 0.66 | 0.74 | 0.78 |
| KNN | 0.94 | 0.66 | 0.65 | 0.749 |
| SVM | 0.94 | 0.59 | 0.72 | 0.749 |
| Decision Tree | 0.71 | 0.75 | 0.81 | 0.758 |
| Random Forest Classifier | 0.94 | 0.82 | 0.84 | 0.864 |
| MLP Classifier | 0.81 | 0.7 | 0.84 | 0.783 |
| Gradient Boosting classifier | 0.87 | 0.82 | 0.86 | 0.85 |
| Naive Bayes classifier | 1 | 0.39 | 0.65 | 0.679 |

From Table 2, it was observed that the overall highest recall value corresponds to the random forest classifier. This suggests the random forest classifier is the most successful algorithm to avoid false negatives in all three classes among the set of classifiers on which the study was conducted.

From Table 2, it can also be inferred that the maximum recall for the class of PDAC patients (class 3) is for Gradient Boosting classifier. This implies that the Gradient Boosting classifier can correctly predict the maximum number of samples with PDAC among the set of classifiers.

In this study, the main focus is to find the positive cases (samples with PDAC) in order to avoid catastrophic consequences. Naïve Bayes classifier has the lowest recall value of 0.679.

**TABLE 3: PRECISION VALUES**

| Classifier | Classes | | | *Overall* |
|---|---|---|---|---|
| | ***Healthy*** | ***Non-PDAC*** | ***PDAC*** | |
| Logistic Regression | 0.62 | 0.78 | 0.94 | 0.781 |
| KNN | 0.6 | 0.74 | 0.9 | 0.75 |
| SVM | 0.6 | 0.74 | 0.89 | 0.744 |
| Decision Tree | 0.69 | 0.69 | 0.92 | 0.765 |
| Random Forest Classifier | 0.76 | 0.84 | 0.97 | 0.858 |
| MLP Classifier | 0.71 | 0.76 | 0.86 | 0.776 |
| Gradient Boosting classifier | 0.77 | 0.8 | 0.97 | 0.848 |
| Naive Bayes classifier | 0.53 | 0.59 | 0.93 | 0.682 |

From Table 3, it can be observed that Random forest classifier has the highest overall precision score. To get the measure of patients that has been correctly identified having PDAC out of all the patients actually having it would be to consider the highest precision score for class-3. From the Table-3, has been inferred that Random forest classifier and Gradient boosting classifier have the maximum precision score for class-3. Naive Bayes classifier has the least precision value.

**TABLE 4: F1-SCORE VALUES**

| Classifier | Classes | | | *Overall* |
|---|---|---|---|---|
| | ***Healthy*** | ***Non-PDAC*** | ***PDAC*** | |
| Logistic Regression | 0.74 | 0.72 | 0.83 | 0.764 |
| KNN | 0.73 | 0.7 | 0.76 | 0.73 |
| SVM | 0.73 | 0.66 | 0.79 | 0.729 |
| Decision Tree | 0.7 | 0.72 | 0.86 | 0.76 |
| Random Forest Classifier | 0.84 | 0.83 | 0.9 | 0.856 |
| MLP Classifier | 0.76 | 0.73 | 0.85 | 0.778 |
| Gradient Boosting classifier | 0.82 | 0.81 | 0.91 | 0.847 |
| Naive Bayes classifier | 0.69 | 0.47 | 0.77 | 0.641 |

**TABLE 5: SPECIFICITY VALUES**

| Classifier | Classes | | |
|---|---|---|---|
| | ***Healthy*** | ***Non-PDAC*** | ***PDAC*** |
| Logistic Regression | 0.772 | 0.884 | 0.967 |
| KNN | 0.75 | 0.851 | 0.951 |
| SVM | 0.75 | 0.87 | 0.932 |
| Decision Tree | 0.872 | 0.792 | 0.948 |
| Random Forest Classifier | 0.889 | 0.903 | 0.985 |
| MLP Classifier | 0.87 | 0.859 | 0.903 |
| Gradient Boosting classifier | 0.901 | 0.877 | 0.984 |
| Naive Bayes classifier | 0.616 | 0.831 | 0.96 |

From Table 4, it can be observed that the Random forest classifier has the highest overall F1-score, followed closely by the Gradient boosting classifier. Also, Gradient boosting classifier has the highest F1-score for class 3 (samples with PDAC). Least F1-score corresponds to Naive Bayes classifier. From the table 5, it can be seen that specificity values for class-3 is highest for Random forest classifier and closely followed by Gradient Boosting classifier. The lowest specificity value for PDAC samples corresponds to MLP classifier.

**TABLE 6: ACCURACY**

| Classifier | Accuracy |
|---|---|
| Logistic Regression | 0.7627 |
| KNN | 0.7288 |
| SVM | 0.7288 |
| Decision Tree | 0.7627 |
| Random Forest Classifier | 0.8559 |
| MLP Classifier | 0.7797 |
| Gradient Boosting classifier | 0.8475 |
| Naive Bayes classifier | 0.6441 |

The Random Forest classifier has the highest accuracy of 85.5%, closely followed by the Gradient Boost classifier with an accuracy of 84.7%. Naive Bayes classifier has the lowest accuracy of 64.4% as evident from Table 6.

Considering the number of features as 3, 7, and 32 Artificial neural network models and Logistic regression models were created for each case, showed that corresponding artificial neural network models were better than Logistic regression models [19]. Similar to this result, a few algorithms outperformed Logistic regression in the classification task. Surprisingly, both logistic regression and the decision tree classifier have an accuracy of 76.27%, which is better compared to K-Nearest Neighbor and Support Vector Machine, Naïve Bayes classifier having accuracies 72.8%, 72.8%, and 64.4% respectively.

Out of the 262 samples, 183 of them were used for training and 79 were used as validation set showed that Random Forest outperformed other methods in the training set. In the validation set, the support vector machine (SVM) model outperformed the others in predicting 1-year relapse risk. In regard to predicting 2-year relapse risk, the K-neighbor algorithm (KNN) model had the highest accuracy and AUROC [20]. Similar to this, Here Random Forest outperformed other methods. But KNN and SVM showed low accuracy scores compared to other methods used.

## IV. CONCLUSION

Machine learning techniques could be safely used to improve early diagnosis and help in the early treatment and survival of patients. By analyzing the performance metrics carefully, it was observed that the random forest classifier outperforms other classifiers that we have considered for the study. Performance metric analysis suggests that the random forest classifier and the gradient boosting classifier were similar. Random forest classifier has the highest accuracy of 85.5%. It had a recall score of 0.864, a precision score of 0.858, and an F1 score of 0.856.

These non-invasive techniques could be used for the screening of pancreatic cancer. But to improve the performance, there is a need for more data. It is crucial to come up with new techniques that could improve the performance of the model drastically, thereby helping in accurately identifying pancreatic cancer.

## REFERENCES


[1]. F. Bray, J. Ferlay, I. Soerjomataram, RL. Siegel, LA. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J Clin, 2018, 68(6):394–424.

[2]. W. Park, A. Chawla and EM.O'Reilly, " Pancreatic Cancer: A Review," *JAMA*, 2021 Sep 7;326(9):851-862.

[3]. MCS. Wong, JY. Jiang, M. Liang, Y. Fang, MS. Yeung and JJY. Sung, "Global temporal patterns of pancreatic cancer and association with socioeconomic development," Sci Rep, 2017, 7(1):3165.

[4]. TP. Radon, NJ. Massat, R. Jones, W. Alrawashdeh, L. Dumartin, D. Ennis et al, " Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma," Clin Cancer Res, 2015 Aug 1;21(15):3512-21.

[5]. S. Debernardi , H. O'Brien , AS. Algahmdi, N. Malats, GD. Stewart, M. Plješa-Ercegovac et al, "A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study," PLoS Med, 2020 Dec 10;17(12):e1003489.

[6]. Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, *7*(4), 190-207.

[7]. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

[8]. Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, 2020, Volume 408, Pages 189-215

[9]. Ghosh, Sourish & Dasgupta, Anasuya & Swetapadma, Aleena, "A study on support vector machine based linear and non-linear pattern classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, February, (pp. 24-28)

[10]. Harsh H. Patel, & Purvi Prajapati, "Study and analysis of decision tree based classification algorithms," International Journal of Computer Sciences and Engineering, 2018, 6(10), 74-78.

[11]. MN. Anyanwu and SG. Shiva, "Comparative analysis of serial decision tree classification algorithms," International Journal of Computer Science and Security, 2009 Jun;3(3):230-40.
[12]. Breiman, L., "Random Forests," Machine Learning, 2001, 45(1), 5-32
[13]. Jehad Ali, Rehanullah Khan, Nasir Ahmad and Imran Maqsood, "Random Forests and Decision Trees." International Journal of Computer Science Issues (IJCSI), 2012, 9(5), 272.
[14]. Alexy Natekin and Alois Knoll, "Gradient boosting machines, a tutorial," Frontiers in neurorobotics, 2013, 7, 21.
[15]. Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-ilne learning and an application to boosting," Journal of computer and system sciences, 1997, 55(1), 119-139.
[16]. Gurpreet Singh and M. Sachan, "Offline Gurmukhi script recognition using knowledge based approach & Multi-Layered Perceptron neural network" 2015 International Conference on Signal Processing, Computing and Control (ISPCC), 2015, September, pp. 266-271
[17]. F. Yang, "An Implementation of Naive Bayes Classifier," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 301-306
[18]. Dalianis, Hercules, "Evaluation metrics and evaluation," Clinical text mining, 2018, pp. 45-53
[19]. Zhou Tong, Yu Liu, Hongtao Ma , Jindi Zhang , Bo Lin , Xuanwen Bao , et al, "Development, validation and comparison of artificial neural network models and logistic regression models predicting survival of unresectable pancreatic cancer" Frontiers in Bioengineering and Biotechnology, 2020, 8, 196
[20]. Li, Xiawei & Yang, Litao & Zheping, Yuan & Lou, Jianyao & Fan, Yiqun & Shi et al, "Multi-institutional development and external validation of machine learning-based models to predict relapse risk of pancreatic ductal adenocarcinoma after radical resection," Journal of translational medicine, 2021, 19(1), 1-10.