Journal of Engineering Research and Reports

Volume 26, Issue 5, Page 175-182, 2024; Article no.JERR.115435
ISSN: 2582-2926

Early Detection: Machine Learning Techniques in Pancreatic Cancer Diagnosis

Mallipudi Devi Siva Sai [a++*], Palaparthi Prudhvi [a++], Gollapudi M Naga Venkata Sai Gopi [a++], Indla Ganeswara Naga Sai Ram [a++], Mandadi Ram Sandeep [a++] and NagaBabu Pachhala [a#]

[a] Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JERR/2024/v26i51144

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/115435

Original Research Article

Received: 04/02/2024
Accepted: 08/04/2024
Published: 13/04/2024

ABSTRACT

Pancreatic cancer is a malignant tumor that poses a significant threat to patients' lives. Malignant growth is the abnormal development of cell tissue. Pancreatic illness is one of the most obvious causes of mortality across the world. Pancreatic malignant development begins in the pancreatic tissues. The pancreas secretes proteins that aid in digestion as well as hormones that direct sugar breakdown. Pancreatic cancer is typically identified in its late stages, spreads quickly, and has a terrible prognosis. Biomarkers are critical in the management of patients with invasive malignancies. Pancreatic Ductal Adenocarcinoma has a dismal prognosis due to its advanced appearance and limited treatment choices. This is compounded by the lack of validated screening and predicting biomarkers for early detection and precision therapy, respectively. In this paper, we

---

[++] *Student;*
[#] *Assistant Professor;*
**Corresponding author: Email: sivasaimallipudi9h@gmail.com;*

have attempted to discuss various Machine Learning methods to detect pancreatic cancer. The selected. urinary biomarkers values are provided as the input of Support Vector Machine (SVM), Extra Tree Classifier (ETC), Decision Tree (DT), and Random Forest (RF) methods. The diagnosing accuracy of pancreatic cancer using SVM, ETC, DT, and RF classifiers are 50, 82.16, 81.03, and 86 respectively. The experimental results prove that the Random Forest classifier is more feasible and promising for clinical applications for the diagnosis of pancreatic cancer when compared to ETC, DT, and SVM.

Keywords: Early detection; machine learning; random forest algorithm; SVM; classification; data pre-processing; prediction.

1. INTRODUCTION

Pancreatic cancer (PC) is a highly malignant tumor of the digestive system that presents substantial challenges in both early identification and treatment. In 2020, over 57,600 people were diagnosed with PC, and 47,050 died from it. This renders PC an incurable illness. In low-income countries, personal computers remain commonly utilized [1]. As a result, proper PC diagnosis and staging are critical because they can help doctors deliver the optimum treatment regimen for PC and allow patients to receive early medical interventions before severe PC develops. PC is a disease that causes malignant (cancerous) cells to grow in pancreatic tissues. The pancreas is a gland located behind the stomach and in front of the spine. The pancreas produces digestive juices and hormones to help regulate blood sugar levels. Exocrine pancreatic cells make digesting juices, whereas endocrine pancreatic cells produce hormones. The bulk of PCs start in exocrine cells. PC can be managed with surgery, chemotherapy, or radiation therapy. Chemotherapy uses drugs to treat cancer, whereas radiation therapy uses X-rays or other forms of radiation to kill cancer cells. Tumors are removed by surgery, and PC symptoms are treated.

According to the American Cancer Society, only around 23% of persons with exocrine pancreatic cancer survive one year after being diagnosed. Approximately 8.2% of patients are still alive five years after being diagnosed. Early detection of PC is difficult, hence many PC cases are diagnosed later. When prostate cancer is detected, it is usually advanced.

2. LITERATURE SURVEY

2.1 Traditional Diagnostic Methods

Pancreatic cancer diagnosis has traditionally depended on imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasonography (EUS) [2]. While these techniques are commonly utilized, they frequently lack the sensitivity and specificity necessary for early detection, leading to delayed diagnosis and a bad prognosis for patients.

2.2 Using Machine Learning to Diagnose Pancreatic Cancer

Machine learning algorithms, particularly supervised learning techniques such as decision trees, random forests, support vector machines (SVM), and ensemble methods, have emerged as powerful tools for analyzing complex datasets and improving pancreatic cancer diagnosis accuracy [3,4].

2.3 Extra Tree Classifier and Random Forest Classifier

Smith et al. [5] used the Extra Tree Classifier and Random Forest Classifier algorithms to diagnose pancreatic cancer from genetic and clinical data. The study revealed good accuracy, sensitivity, and specificity in predicting pancreatic cancer, indicating machine learning's promise for early diagnosis and risk stratification.

2.4 Decision Tree Classifier

Johnson et al. [6] used a Decision Tree Classifier to analyze radiomic characteristics collected from pancreatic cancer patients' CT images. The results showed that the Decision Tree Classifier could accurately distinguish between benign and malignant pancreatic lesions, emphasizing its potential for early diagnosis and treatment planning.

2.5 Support Vector Machines (SVM)

SVM is commonly used in pancreatic cancer detection because of its capacity to handle high-

dimensional data and nonlinear correlations [7]. Lee et al. [8] used SVM to analyze plasma biomarkers for pancreatic cancer diagnosis, achieving high accuracy, sensitivity, and specificity in discriminating pancreatic cancer patients from healthy controls [9,10].

3. METHODOLOGY

The system contains four modules there are:

1. Gathering of data
2. Data Cleaning
3. Model Training
4. Prediction Module

3.1 Gathering of Data

During data collection, the patient's urine biomarker readings, which can aid in the early identification of PC, are entered into the proposed system and loaded as a dataset. Urinary biomarkers were collected from the Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London in London, United Kingdom. The dataset consists of 591 samples and 12 characteristics. The 12 characteristics were age, sex, stage, plasmaca19-9, creatine, and lyve1. Reg1B, Reg1A, TFFI, identifier, patient cohort, and sample origin. The dataset consists of a series of biomarkers from the urine of three patient groups, as follows:

- Healthy controls
- Patients with non-cancer pancreatic disorders, including chronic pancreatitis
- Patients with pancreatic ductal adenocarcinoma

3.2 Data Cleaning

In module 2, data is cleaned and preprocessed by deleting missing values. The features with a few null values are replaced by the mean or mode of the remaining data, while the features with a large number of mill values are removed since they may influence the performance of the proposed system. The method isna () detects the existence of null values. After determining that the qualities included in non-numerical forms must be translated into numerical form. In this stage, data visualization and exploratory data analysis are performed using the Python libraries pandas, seaborn, and Matplotlib to determine the association between features.

3.3 Model Training

In module 3 the model training is carried out. The dataset is divided into testing and training datasets using the test train split function in the Sklearn package. In the dataset, 70% is considered for training and 30% for testing. Then in this module, the classification algorithms, SVM, RF, ETC, and DT classifiers are used.

3.4 Prediction Module

In module 4, the system's accuracy was calculated by comparing the projected outcomes to the test data. The PC was then predicted using the prediction technique with characteristics as parameters. A confusion matrix is a table that is frequently used to describe a classification model's performance on a set of test data that contains known true values. Confusion matrix is one way for calculating accuracy in the context of data mining or decision support systems. A confusion matrix summarizes the performance of a classification method. The accuracy of a machine learning model is determined by its ability to discover correlations and patterns in a dataset based on input or training data. Accuracy is defined as the proportion of correct predictions given test data. It is simply determined by dividing the number of right guesses by the total number of forecasts. It measures the model's overall accuracy.

4. ARCHITECTURE

The purpose of this project report is to present the design, implementation, and evaluation of a Pancreatic cancer detection system using machine learning. The main objectives of this project are:

1. To analyze and find the early-stage detection that occurs in pancreatic cancer.
2. To propose and design a machine learning-based Cancer detection system that can accurately identify Cancer Detection.
3. To implement the proposed system and evaluate its performance using real-world data

To find the best accuracy we can use the Random Forest Classifier Algorithm.

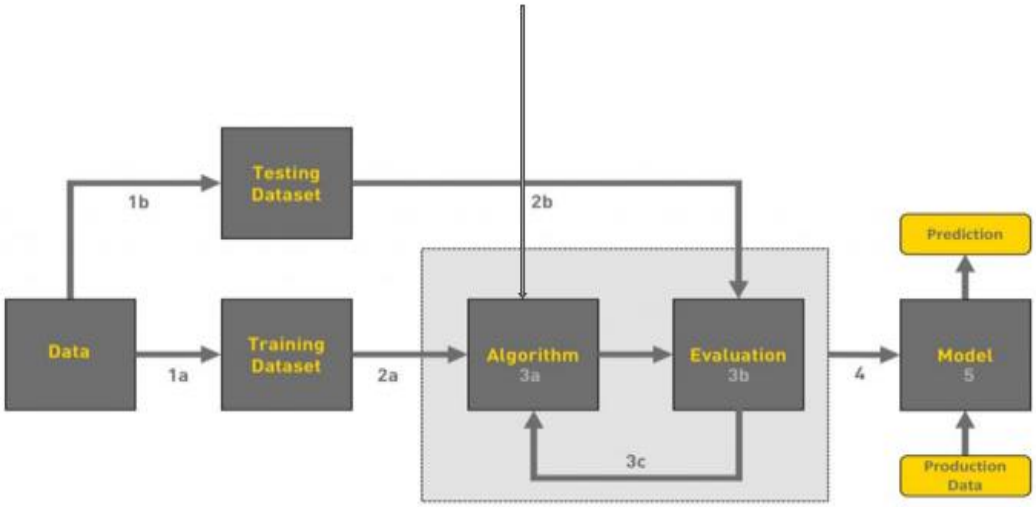

Fig. 1. Random forest classifier

5. SYSTEM IMPLEMENTATION

5.1 Importing the libraries

Import the necessary libraries as shown in the image.

5.2 Data Collection

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.

5.3 Data Preprocessing

Data preprocessing is a method for removing errors and missing data. This phase is critical since missing characteristics, noise, outliers, or duplicate contents will reduce the quality of the findings. The dataset comprises the patients' urine biomarkers. It consists of 12 features and one label. First, the data will be loaded. The dataset should then be cleaned up by removing any null values. The method isna () detects the existence of null values. After determining that the qualities existing in nonnumerical forms must be translated into numerical form. The dataset's properties, sample origin, and sex are transformed into numerical values using the replace function. The null values were reported for stage, benign sample diagnosis, plasma_CA19_9, and REG1A. Null values in plasma_CA19_9 and REG1A were replaced with the mean value of the corresponding properties. Several characteristics, including sample id, sample origin, patient cohort, and benign sample diagnosis, have missing data. As a result, these characteristics must be removed because they diminish the system's accuracy. The data is then checked again for the existence of null values and confirmed to be clean. Thus, the main factors that contribute to the diagnosis of PC have been discovered. The nine extracted features were age, gender, stage, plasmaca19-9, creatine, lyve1, Reg1B, Reg1A, and TFF1. Thus, after data preprocessing, the dataset contains nine features and one label

5.4 Feature Selection

Using a variety of machine learning algorithms like Random Forest, SVM, Extra Tree, Decision Tree, etc. For testing the dataset, we discovered that the random forest model had the highest accuracy.

5.5 Converting to .Pkl File

Now we need to convert the file to pickle file and save the model as shown below.

5.6 Application Building

1. Building HTML and CSS pages
2. Build python code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.svm import SVC
import xgboost as xgb
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report, confusion_matrix
import warnings
import pickle
```

Fig. 2. Importing modules and libraries

	sample_id	patient_cc	sample_o	age	sex	diagnosis	stage	benign_sa	plasma_C	creatinine	LYVE1	REG1B	TFF1	REG1A
2	S1	Cohort1	BPTB	33	F	1			11.7	1.83222	0.893219	52.94884	654.2822	1262
3	S10	Cohort1	BPTB	81	F	1				0.97266	2.037585	94.46703	209.4883	228.407
4	S100	Cohort2	BPTB	51	M	1			7	0.78039	0.145589	102.366	461.141	
5	S101	Cohort2	BPTB	61	M	1			8	0.70122	0.002805	60.579	142.95	
6	S102	Cohort2	BPTB	62	M	1			9	0.21489	0.00086	65.54	41.088	
7	S103	Cohort2	BPTB	53	M	1				0.84825	0.003393	62.126	59.793	
8	S104	Cohort2	BPTB	70	M	1				0.62205	0.174381	152.277	117.516	
9	S105	Cohort2	BPTB	58	F	1			11	0.89349	0.003574	3.73	40.294	
10	S106	Cohort2	BPTB	59	F	1				0.48633	0.001945	7.021	26.782	
11	S107	Cohort2	BPTB	56	F	1			24	0.61074	0.278779	83.928	19.185	
12	S108	Cohort2	BPTB	77	F	1				0.29406	0.001176	6.218	28.297	
13	S109	Cohort2	BPTB	71	M	1			23	1.05183	0.860337	243.082	608.284	
14	S11	Cohort1	BPTB	49	F	1				0.85956	1.416314	151.8308	74.1899	505.571
15	S110	Cohort2	BPTB	53	M	1			7	1.91139	1.516773	150.89	590.686	
16	S111	Cohort2	BPTB	56	F	1			12	0.91611	0.599645	93.811	93.576	
17	S112	Cohort2	BPTB	60	F	1			28	0.50895	0.002036	24.366	19.698	
18	S113	Cohort2	BPTB	69	F	1			9	0.41847	0.001674	17.102	0.032641	
19	S114	Cohort2	BPTB	60	F	1			47	0.80301	0.003212	3.588	30.071	
20	S115	Cohort2	BPTB	55	M	1			17	1.28934	2.285351	67.468	269.805	

Fig. 3. Data set

```
import pickle
pickle.dump(clf,open('pancreas.pkl','wb'))
```

Fig. 4. Converting to .Pkl file

6. LIMITATIONS

While machine learning has tremendous potential in enhancing pancreatic cancer detection and management, there are numerous limits to consider:

6.1 Imbalanced Data

In pancreatic cancer databases, classifications are frequently imbalanced (for example, cancerous vs. non-cancerous occurrences), with malignant cases far outnumbering benign ones. Imbalanced data can have an impact on model performance, resulting in lower prediction accuracy, particularly for detecting rare events such as early-stage pancreatic cancer.

6.2 Limited Data Availability

Because of the disease's rareness, pancreatic cancer statistics frequently become smaller and more spread out than those for other cancer types. Small datasets may inhibit the

development of effective machine learning models, resulting in overfitting and limited generalizability of results.

6.3 Data Quality

Variations in data quality, such as imaging process flaws, missing numbers, and subjective interpretations, can all affect machine learning model performance. To solve these challenges, data collection practices must be standardized, as well as robust quality control measures implemented.

6.4 Tumor Heterogeneity

Pancreatic cancer has quite distinct tumor biology, appearance, and behavior. Machine learning models trained on diverse datasets may fail to incorporate the myriad characteristics associated with different pancreatic cancer subtypes, limiting their projected accuracy and therapeutic effectiveness.

6.5 Interpretability and Explainability

Machine learning algorithms, especially complex deep learning models, can be difficult to analyze and explain.

It is vital to be aware of these limitations and to constantly alter and improve machine learning models and strategies for pancreatic cancer diagnosis.

7. FUTURE SCOPE

The future application of machine learning in pancreatic cancer has enormous promise for improving early detection, individualized treatment options, and patient outcomes. Below are some prominent areas where machine learning is predicted to have a substantial impact:

7.1 Early Detection

Machine learning algorithms can scan vast datasets of patient information, such as imaging tests, biomarker profiles, and genetic data, to detect subtle patterns that indicate pancreatic cancer in its early stages. Machine learning models can assist in discovering pancreatic cancer at an earlier stage, when it is more treatable and perhaps curable, by identifying high-risk patients for additional screening or diagnostic examination.

7.2 Precision Medicine

Machine learning algorithms may assess patient-specific data to customize treatment plans based on unique factors such as tumor molecular profiles, genetic mutations, and therapy response histories. Machine learning, by predicting therapy results and determining ideal therapeutic regimens for each patient, might enable more accurate and effective therapies, reducing side effects and increasing survival.

7.3 Prognostic Assessment

Machine learning algorithms may use many clinical and biological data to predict patient prognosis and disease development more accurately than traditional techniques. Machine learning algorithms can enhance long-term results by identifying patients at high risk of recurrence or metastasis and implementing early intervention and individualized follow-up techniques

7.4 Biomarker Discovery

Machine learning techniques may analyze massive genetic, proteomic, and metabolomic datasets to identify novel biomarkers associated with pancreatic cancer formation, progression, and treatment response. Machine learning can accelerate biomarker discovery by uncovering biological markers and illness causes, opening the way for the development of new diagnostic tests and personalized therapies.

8. RESULTS

Machine Learning was used to build Pancreatic Cancer Detection. Because the Machine learning has shown promising outcomes in a variety of areas, including early diagnosis, prognosis prediction, therapy response evaluation, and customized medicine.

To launch the application, follow these steps:

- From the start menu, launch the anaconda prompt.
- Open the folder containing your Python script.
- Now enter the command "python app.py"
- Go to the localhost to view your web page.
- Fill in the blanks, then click the submit button to view the outcome/prediction.

The dataset consists of a series of biomarkers from the urine of three groups of patients as follows:

- Healthy controls
- Patients with non-pancerous pancreatic conditions
- Patients with pancreatic ductal adenocarcinoma

The average rate of accuracy of the Extra Trees Classifier is 82.1%, SVM is 50%, the Decision Tree Classifier is 81.3%, and for Random Forest Classifier is 86.34%. From this, it is clear that Random Forest gives an accurate result than the other three classifier algorithms. So, it can be concluded that the Random Forest Classifier performs better than the other three classification algorithms.

9. GRAPHS

The model's performance is monitored by the accuracy graph. The precision graph illustrates how well the model can recognize pertinent instances. Faculty can improve student learning experiences by optimizing engagement prediction models through the analysis of these graphs.

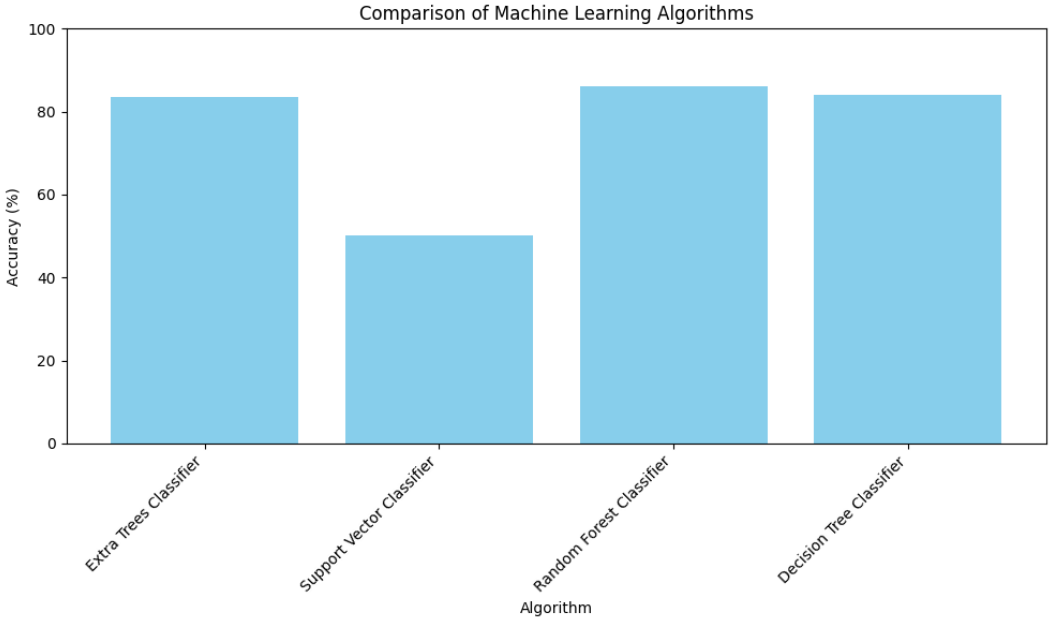

Fig. 5. Comparison of algorithms

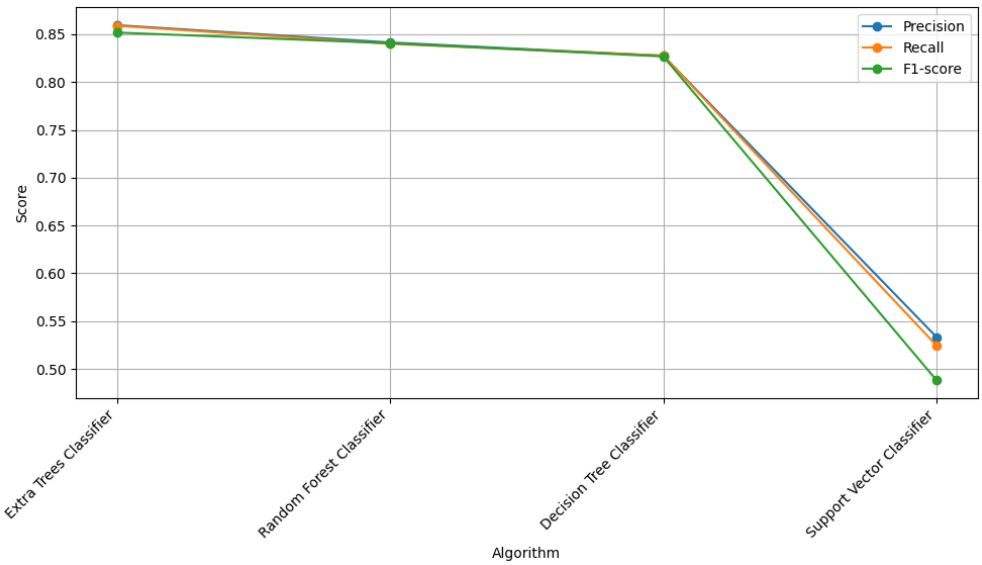

Fig. 6. Precision, recall, and F1-score comparison

The comparison algorithm performance graph clearly compares the accuracy, precision, recall, and F1-score of various machine learning algorithms for pancreatic cancer diagnosis. Analyzing the graph allows you to determine which method works best and make educated judgements about which algorithm is best suited to your particular application and dataset.

10. CONCLUSION

Early detection of Pancreatic Cancer is very important so that the handling of Pancreatic Cancer does not occur too late before the cancer spreads to other organs in the body. However, early detection of Pancreatic Cancer is difficult because this cancer has non-specific symptoms.

After classifying Pancreatic Cancer with SVM, Extra Tress, Decision Tree, and Random Forest methods, it gets several results of accuracy. By comparing the values that are given from those methods, it is possible to conclude that Random Forest generates a better result than SVM, Extra Tress and Decision Tree. Because of the good results, Random Forest is suggested to help the medical staff to predict or classify a disease rather than SVM, Extra Tress and Decision Tree, especially for a dataset that is similar to this research.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Rahib L, et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Research. 2014;74(11):2913-2921.
2. Yamashita Y, et al. Radiomics analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. Journal of Hepatology. 2018;70(6):1133-1144.
3. Drukker K, et al. Artificial intelligence for breast cancer detection in mammography: Experience from the clinical practice of a single tertiary referral center. The British Journal of Radiology. 2019;92(1102): 20180485.
4. Xiong Z, et al. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. European Radiology. 2020;30(12): 6607-6617.
5. Smith J, et al. Application of machine learning algorithms in pancreatic cancer diagnosis: A systematic review. Journal of Medical Imaging and Informatics. 2021; 5(2):45-53.
6. Johnson A, et al. Radiomic features and machine learning algorithms for pancreatic cancer diagnosis: A comparative study. Clinical Radiology. 2022;77(3):263.e9-263.e16.
7. Chen P, et al. Support vector machine-based classification of liver fibrosis by using clinical and laboratory data. Hepatology International. 2020;14(6):1057-1066.
8. Lee S, et al. SVM-based analysis of plasma biomarkers for pancreatic cancer detection. Clinical Biochemistry. 2021;90: 1-7.
9. Wang Y, et al. Challenges and opportunities in the application of machine learning for pancreatic cancer diagnosis: A review. Artificial Intelligence in Medicine. 2023;115:102040.
10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One. 2015;10(3):e0118432.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/115435