Original Research Article

Early Detection: Machine Learning Techniques in Pancreatic Cancer Diagnosis

.

ABSTRACT

Pancreatic cancer has a terrible prognosis by having a survival rate of five years only. The premise behind early detection and better survival is that more people will benefit from a possible treatment. Pancreatic cancer is a devastating disease with a high mortality rate, often diagnosed at advanced stages when treatment options are limited. Early detection plays a crucial role in improving patient outcomes. The great majority of the computer systems that are now being utilized for research on medical health systems are based on the most recent technical breakthroughs. Because of the prevalence of pancreatic cancer, a significant number of novel approaches and techniques have emerged in the field of medicine. There are several various classifications that may be applied to the pancreatic cancer that can be found. The classification of pancreatic cancer may be tackled from a variety of angles, each of which can be accomplished via using either technology for machine learning. In the past, a diagnosis of pancreatic cancer could be made by using methods such as the Support Vector Machine (SVM),ExtraTree, Decision Tree. However, these strategies do not deliver an accurate performance. As a result, this study has implemented a Random Forest Classifier due to its ability to handle complex relationships within the data.

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 provides significant hurdles in both early detection and subsequent therapy. In 2020, around 57,600 persons were diagnosed with PC, and 47,050 died from it. This makes PC an incurable disease. PCs continue to be widely used in poor nations [1]. As a result, complete PC diagnosis and staging are very crucial, as they may assist doctors provide the best therapy regimen for PC and allow patients to obtain early medical therapies before severe PC develops. PC is a disorder that causes malignant (cancerous) cells to develop in pancreatic tissues. The pancreas is a gland that sits behind the stomach and in front of the spine. The pancreas generates digestive juices and hormones that help regulate blood sugar levels. Exocrine pancreatic cells generate digestive fluids, whereas endocrine pancreatic cells create hormones. The majority of PCs begin in exocrine cells. PC can be treated with surgery, chemotherapy, or radiation therapy. Chemotherapy utilizes medications to treat cancer, whereas radiation treatment employs X-rays or other types of radiation

to destroy cancer cells. Surgery is done to remove tumors or cure PC symptoms.

According to the American Cancer Society, only around 23% of people with exocrine pancreatic cancer survive a year following diagnosis. Five years after their diagnosis, around 8.2% are still living. Early identification of PC is challenging, hence many PC cases are detected late. When PC is discovered, the cancer is typically advanced. Machine learning is a branch of artificial intelligence that can identify PCs early.

2. LITERATURE SURVEY

[1] Due to its high fatality rates and poor prognosis, pancreatic cancer is among the deadliest cancers. Pancreatic cancer is still a difficult disease to treat because of its aggressiveness, late diagnosis, and lack of available treatment choices, even with major advances in cancer research and therapy techniques. An extensive review of the literature on pancreatic cancer offers insightful information about the pathophysiology, diagnostics, care plans, and new therapeutic developments that are now being understood.

[2] A complex interaction of lifestyle, environmental, and hereditary variables leads to pancreatic cancer etiology. Mutations in the KRAS oncogene, tumor suppressor genes including TP53, CDKN2A, and SMAD4, and changes in DNA repair pathways are among the major genetic changes linked to the onset of pancreatic cancer, according to research. Pancreatic cancer is also influenced by risk factors such obesity, tobacco use, chronic pancreatitis, and family history.

[3] The absence of distinct symptoms and efficient screening techniques makes early detection of pancreatic cancer difficult. Most patients have poor outcomes because they present with advanced stages of the disease. The diagnosis and staging of pancreatic cancer depend heavily on imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasonography (EUS). Treatment response and illness progression are also tracked with biomarkers such carbohydrate antigen 19-9 (CA 19-9).

[4] Surgery is still the primary treatment for localized pancreatic cancer, with the goal of curing it. However, the majority of patients are detected at an advanced stage, when surgery is no longer an option. To enhance survival rates in such situations, systemic treatment, which includes chemotherapy, targeted therapy, and immunotherapy, is employed. Gemcitabine and nab-paclitaxel, also known as FOLFIRINOX (folinic acid, fluorouracil, irinotecan, and oxaliplatin), have been demonstrated to improve overall survival in patients with metastatic pancreatic cancer. Furthermore, advances in precision medicine have enabled the creation of tailored medicines that take advantage of particular molecular abnormalities in pancreatic cancer cells, such as MEK and PI3K pathway inhibitors.

[5] Recent advancements in cancer research have paved the road for the development of innovative pancreatic cancer therapies. Immunotherapy, oncolytic virotherapy, epigenetic modulators, and tumor microenvironment-targeted medicines are among them. Immunotherapy, particularly immune checkpointinhibitors targeting programmed cell death protein 1 (PD-1) and programmed death-ligand 1 (PD-L1), has showed promise in clinical studies, albeit response rates differ between individuals. Oncolytic viruses, such as oncolytic adenoviruses and herpes simplex viruses, are being studied as possible treatment agents for pancreatic cancer, either alone or in conjunction with chemotherapy or immunotherapy. Furthermore, epigenetic modulators, such as DNA methyltransferase and histone deacetylase inhibitors, are being researched for their capacity to modify gene expression patterns in pancreatic

cancer cells and improve therapy response.

[6] A comprehensive literature review of pancreatic cancer emphasizes the disease's complicated character and the critical need for ongoing research efforts to enhance early diagnosis, treatment results, and patient survival. Researchers want to change pancreatic cancer care and improve patient outcomes by unraveling the underlying processes of pathogenesis and discovering new therapeutic targets.

3. METHODOLOGY

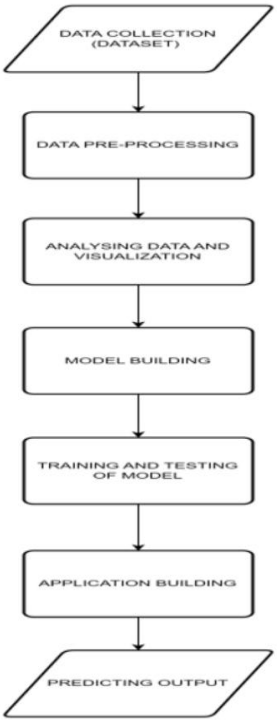

Fig 1: Study protocol

i. Data collection:

In machine learning, data collection refers to the act of getting, acquiring, and integrating data that will be used to construct, test, and validate a model. Data collection for ML model training is the most important part of the machine learning workflow. Machine learning (ML) systems' predictions are only as accurate as their training data.

ii. Data processing:

Data processing in machine learning refers to the many procedures and transformations that are applied to data to prepare it for analysis and model training. Cleaning, cleaning, and preparing data for use in machine learning algorithms necessitates many steps. The utility and effectiveness of machine learning models may be greatly influenced by the quality of data handling. Before it can be used to create models, the imported dataset must be cleaned.

iii. Analyzing Data and Visualization:

Data analysis and visualization are iterative processes in machine learning, necessitating repeated returns to improve your understanding of the data as new insights emerge. Effective visualization and analysis help to enhance model selection, feature engineering, and distribution of findings to non-technical stakeholders. It is an important stage in the machine learning workflow since it ensures openness and trust in the model's predictions.

iv. Model construction:

Model creation in pancreatic cancer using machine learning is a comprehensive process that includes meticulous data collection, preprocessing, feature engineering, algorithm selection, training, assessment, optimization, and deployment. Researchers and doctors may use machine learning approaches to create solid prediction models to help with early identification, prognosis, and individualized therapy options for pancreatic cancer patients.

v. Model Training and Testing:

A portion of the dataset known as training data is used to train the machine learning model to recognize patterns and make predictions. It consists of input qualities and labels or goal values that match to

output attributes. Test data is a subset of the dataset that is used to assess the model's efficacy and ability to forecast new, unanticipated data.

vi. Application Development:

In the discipline of machine learning, the process of developing software systems or apps that employ machine learning models to address specific real-world issues is known as "application building." These apps leverage machine learning models' predictive abilities to make judgments, propose actions, and automate tasks.

vii.Predicting Output:

Predicting output is the process of using a trained model to make predictions or forecasts based on input data. These predictions, which are essentially the predicted outcomes or responses given by the model, might take several forms depending on the sort of problem you're seeking to solve.

4. ARCHITECTURE

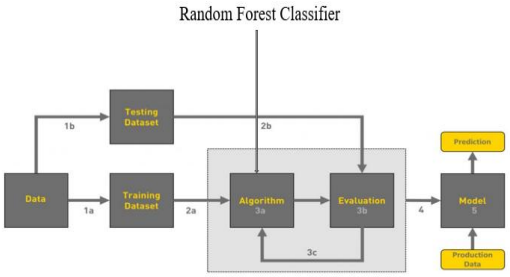

Fig 2: Random forest classifier

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

1. To analyze and finding the early stage detection that occur in pancreatic cancer.

2. To propose and design a machine learning-based Cancer detection system that can accurately identify the 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.

5. SYSTEM IMPLEMENTATION

i. Importing the libraries:

Import the necessary libraries as shown in the image.

```
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 3 Importing modules and libraries

ii. Read the dataset:

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

```
df=pd.read_csv("pradataset.csv")
```

Python

```
df.head()
```

Python

nple_id	patient_cohort	sample_origin	age	sex	diagnosis	stage	benign_sample_diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
S1	Cohort1	BPTB	33	F	1	NaN	NaN	11.7	1.83222	0.893219	52.94884	654.282174	1262.000
S10	Cohort1	BPTB	81	F	1	NaN	NaN	NaN	0.97266	2.037585	94.46703	209.488250	228.407
S100	Cohort2	BPTB	51	M	1	NaN	NaN	7.0	0.78039	0.145589	102.366000	461.141000	NaN
S101	Cohort2	BPTB	61	M	1	NaN	NaN	8.0	0.70122	0.032605	60.57900	142.950000	NaN
S102	Cohort2	BPTB	62	M	1	NaN	NaN	9.0	0.21489	0.000860	65.54000	41.088000	NaN

Fig 4 Reading Data set

iii. Data preprocessing:

The df.isnull() method is used to verify that no values are present. We employ the sum () function to add up those null values. Two null values were discovered in our dataset, we discovered. We thus start by investigating the data.

iv. Using A Heat Map To Check The Correlation:

I'm using a heat map to check the correlation in this instance. Using different color combinations, it displays the data as 2-D colored maps. Instead of numbers, it will be plotted on both axes to describe the relationship variables.

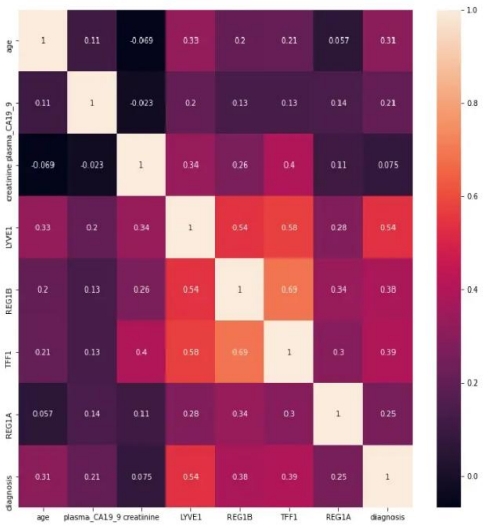

Fig 5 Heat Map

v. Feature selection:

Using a variety ofmachine learning algorithms, including Random Forest, SVM, ExtraTree, Decision Tree, etc., I have discovered a number of metrics in this case. For the testing dataset, we are receiving the random forest model's best accuracy here.

vi. Converting to .pkl file:

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

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

Fig 6 converting to .pkl file

vii. APPLICATION BUILDING:

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

6. PREREQUISITES:

Prerequisites for conducting research or developing models in pancreatic cancer using machine learning include a combination of domain knowledge, technical skills, and access to relevant resources. Here are some key prerequisites:

i.Domain Knowledge in Oncology and Pancreatic Cancer:

Understanding the biology, pathology, and clinical features of pancreatic cancer is critical for data interpretation and machine learning model development. Knowledge of tumor biology, disease development, treatment options, and clinical outcomes is essential for creating clinically meaningful prediction models.

ii.Understanding Machine Learning Concepts:

Designing and implementing machine learning models for pancreatic cancer research requires knowledge of fundamental machine learning concepts such as supervised learning, unsupervised learning, classification, regression, clustering, and model evaluation metrics.

iii.Programming Skills:

Data preparation, feature engineering, model training, and assessment all need knowledge of programming languages typically used in data science and machine learning, such as Python or R. Furthermore, expertise with machine learning packages such as scikit-learn, TensorFlow, and PyTorch is advantageous.

iv. Data Acquisition and Management:

It is vital to have access to high-quality datasets that comprise relevant clinical, imaging, genomic, and biomarker data for pancreatic cancer. Handling and maintaining large-scale biological datasets requires knowledge of data

gathering methods, data cleaning, preprocessing procedures, and data protection legislation.

v.Understanding feature engineering approaches, such as extracting, transforming, and selecting important features from raw data, is critical for increasing machine learning model performance. Domain-specific information can help guide the selection of useful traits that capture the biology and clinical aspects of pancreatic cancer.

7. LIMITATIONS:

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

i. Imbalanced Data:

In pancreatic cancer databases, there is often an imbalance between classifications (for example, cancerous vs. non-cancerous instances), with malignant cases greatly outnumbering benign ones. Imbalanced data can influence model performance, resulting in inferior prediction accuracy, especially for detecting uncommon occurrences like early-stage pancreatic cancer.

ii. Limited Data Availability:

Due to the disease's relative rarity, pancreatic cancer statistics are frequently less in size and breadth than those for other cancer types. Small datasets might impede the construction of effective machine learning models, resulting in overfitting and restricted generalizability of results.

iii. Data Quality:

Data quality variations, such as errors in imaging techniques, missing values, and subjective interpretations, can have an impact on machine learning model performance. To overcome these difficulties, data gathering processes must be standardized and strong quality control techniques used.

iv. Tumor heterogeneity:

Tumor biology, morphology, and behavior are all significantly different in pancreatic cancer. Machine learning models built on heterogeneous datasets may fail to capture the numerous variables associated with various pancreatic cancer subtypes, limiting their predicted accuracy and therapeutic value.

v.Interpretability and Explainability:

Many machine learning algorithms, particularly complicated deep learning models, are sometimes regarded as black-box models, making it difficult to analyze and explain their results.

It is critical to be aware of these limitations and to continually modify and enhance machine learning models and tactics for Pancreatic cancer detection.

8. 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:

i. 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 discover 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.

ii. 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.

iii. 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

.iv. Biomarker Discovery:

Machine learning approaches can evaluate huge genomic, proteomic, and metabolomic datasets to uncover new biomarkers linked to pancreatic cancer development, progression, and therapy response. Machine learning can speed up biomarker discovery by revealing molecular markers and disease causes, paving the path for the creation of novel diagnostic tests and tailored therapeutics.

9. RESULT AND DISCUSSION:

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 pancreaticconditions
- Patients with pancreatic ductal adenocarcinoma

The average rate of accuracy of Extra Trees Classifieris 82.1%, SVM is 50%,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 algorithm. So, it can be concluded that Random Forest Classifier performs better than the other three classification algorithms.

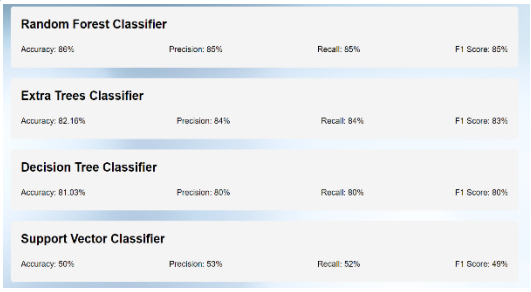

Fig 7Classification algorithms

10. 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.

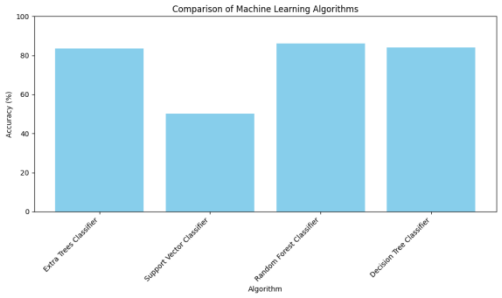

Fig 8 Comparison of Algorithms

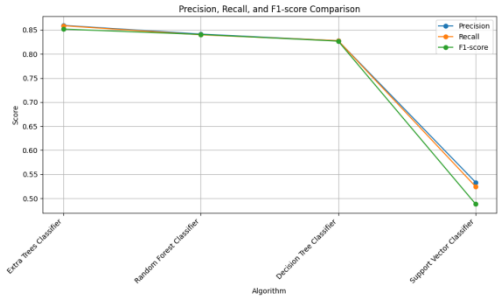

Fig 9 Precision, Recall, and F1-score Comparison

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.

11.REFERENCES:

[1] E. Grywalska et al., "Current Possibilities of Gynecologic Cancer Treatment with the Use of Immune Checkpoint Inhibitors," International Journal of Molecular Sciences, vol. 20, no. 19, September 2019.

[2] S. Midha, S. Chawla, and P. K. Garg, "Modifiable and non-modifiable risk factors for pancreatic cancer: A review," Cancer Letters, vol. 381, no. 1, pp. 269–277, October 2016.

[3] McGuigan, A. et al., "Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment, and outcomes," World Journal of Gastroenterology, vol. 24, no. 43, pp. 4846-4861, November 2018.

[4] Y. Qiu et al., "Towards Prediction of Pancreatic Cancer Using SVM Study Model," Journal of Clinical Oncology and Research, vol.2, no.4, May 2014.

[5] F. Bray et al. "Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394-424, September 2018.

[6] H. Matsubayashi et al. "Familial pancreatic cancer: Concept, Management, and Issues," World Journal of Gastroenterology, vol. 23, no. 6, pp. 935-948, February 2017.

[7] Rawla, Sunkara, & Gaduputi, "Epidemiology of Pancreatic Cancer: Global Trends, Etiology, and Risk Factors," World Journal of Oncology, vol. 10, no. 1, pp. 10-27, February 2019. [8] M. S. De La Cruz, A. P. Young, and M. T. Ruffin. "Diagnosis and management of pancreatic cancer," American Family Physician, vol. 89, no. 8, pp. 626-632, April 2014.

[9] Kuroczycki-Saniutycz et al., "Prevention of pancreatic cancer," Contemporary Oncology (Pozn), vol. 21, no. 1, pp. 30–34, February 2017.

[10] Vareedayah, S. Alkaade, and J. R. Taylor, "Pancreatic Adenocarcinoma," Missouri Medicine, vol. 115, no. 3, pp. 230–235, May/June 2018.

[11] Capasso, M., et al., "Epidemiology and risk factors of pancreatic cancer," Acta Bio Medica Atenei Parmensis, vol. 89, no. 9-S pp. 141-146, December 2018.
[12] T. Nadira and Z. Rustam. "Classification of cancer data using support vector machines with features selection method based on global artificial bee colony," AIP Conference Proceedings, vol. 2023, October 2018.
[13] Aroef, Rivan, & Rustam, "Comparing random forest and support vector machines for breast cancer classification," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 18, no. 2, pp. 815-821, April 2020.
[14] U. Aprilliani and Z. Rustam, "Osteoarthritis disease prediction based on random forest," in 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Yogyakarta, pp. 237–240.
[15] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining. "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer," IOP Conf. Ser.: Mater. Sci. Eng, vol. 546, no. 5, July 2019.
[16] F. R. Aszhari, Z. Rustam, F. Subroto, and A. S. Semendawai, "Classification of thalassemia data using random forest algorithm," J.Phys.: Conf. Ser., vol. 1490, no. 1, June 2020.
[17] Rampisela and Rustam's paper "Classification of Schizophrenia Data Using Support Vector Machine (SVM)," published in J. Phys.: Conf. Ser. vol. 1108, no. 1, December 2018.
[18] Arfiani, Z. Rustam, J. Pandelaki, and A. Siahaan, "Kernel Spherical K- Means and Support Vector Machine for Acute Sinusitis Classification," IOP Conf. Ser.: Mater. Sci. Eng. vol. 546, no. 5, July 2019.
[19] Sadewo, Z. Rustam, H. Hamidah, and A. R. Chusmarsyah, "Pancreatic Cancer Early Detection Using Twin Support Vector Machine Based on Kernel," Symmetry, vol. 12, no. 4, April 2020.
[20] Z. Hua, Y. Wang, X. Xu, B. Zhang, and L. Liang. "Predicting corporate financial distress based on integration of support vector machine and logistic regression," Expert Systems with Applications, vol. 33, no. 2, pp. 434-440, August 2007.
[21] Mandal, S. K. "Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression, and Decision Tree," International Journal of Engineering and Computer Science, vol. 6, no. 2, pp. 20388-20391, February 2017.
[22] Singh, Thakur, and Sharma, "A review of supervised machine learning algorithms," in the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 1310-1315.
[23] N. Md Isa, A. Amir, M. Ilyas, and M. Razalli. "Motor imagery classification in Brain computer interface (BCI) based on EEG signal by using machine learning technique," Bulletin of Electrical Engineering and Informatics, vol. 8, no. 1, pp. 269-275, March 2019.
[24] Z. Saringat, A. Mustapha, R. Saedudin, and N. Samsudin. "Comparative analysis of classification algorithms for chronic kidney disease diagnosis," Bulletin of Electrical Engineering and Informatics, vol. 8, no. 4, pp. 1496-1501, December 2019.
[25] Srivastava, A. K. "Comparison Analysis of Machine Learning algorithms for Steel Plate Fault Detection," International Research Journal of Engineering and Technology (IRJET), vol. 6, no. 5, pp. 1231–1234, May 2019.