Early Detection: Machine Learning Techniques in Pancreatic Cancer Diagnosis

.

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 have made an attempt 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 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 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

Several experiments and research studies on illness diagnosis using machine learning techniques have been conducted in recent years.

2.1 Combining SVM with Magnetic Resonance Imaging:

This study examined the diagnosis and application of PC using Support Vector Machine (SVM) and Magnetic Resonance Imaging (MRI). Simultaneously, the classical SVM classification model is optimizeto increase classification accuracy, while the Quantum Genetic Algorithm (QGA) is utilize to optimize its parameters. The OGA-SVM classification model is built upon this foundation.In the PC detection approach based on the SVM classification model, the parameters of the kernel function and the penalty factor Care are significant elements influencing recognition, therefore accurate parameter selection is critical for improving recognition rate. This statistic represents the RBF kernel function in SVM. Overall, the purpose of this study is to apply MIRI pictures for clinical auxiliary diagnosis research, assisting imaging specialists in identifying PC lesions and providing opinions and references for PC diagnosis. The study's main goal is to find the best approach for extracting crucial PC properties.

2.2 Image Classification using Random Forest:

The goal of this project is to identify a picture based on item category using Random Forest (RF) and ferns. Caltech-101 and Caltech-256 were utilized as Datasets. Caltech-101 contains photos from 101 object categories, while Caltech-256 has images from 256 object categories. This study utilized Image Representation and Matching, with a focus on spatial pyramid matching. Spatial pyramid representation is done by employing appearance and shape descriptors along with the picture spatial arrangement to produce two representations. Pyramid Histogram of Visual Words (PHOW) and Pyramid HOG (PHOG) descriptors for appearance and form, respectively.

2.3Framework for Tumor Detection in Pancreatic Cancer

This study attempts to provide a unique and efficient pancreatic tumor detection system that completely utilizes context information. Computed Tomography (CT) scans are used to represent numerous scales. Deep Convolutional Neural Networks (DCNNs) have demonstrated strong performance and outcomes in medical image processing, prompting the development of many deep-learning-based tumor detection systems in recent years. Nowadays, automated identification of pancreatic tumors utilizing contrast-enhanced CT is commonly used for PC diagnosis and staging. Traditional handcrafted methods can only extract low-level features.Normal convolutional neural networks, however, fail to make full use of effective contest information, which causes inferior detection results. In this paper, a novel and efficient pancreatic tumor detection framework aiming at fully exploiting the context information at multiple scales is designed.

2.4 HEART DISEASE CLASSIFICATION WITH MACHINE LEARNING TECHNIQUES

The goal of this project is to classify cardiac disease using data mining and machine learning techniques. The dataset is sourced from the University of California. The dataset has 13 characteristics, one target variable, and 303 occurrences. This work uses six data mining tools: Orange, Weka, RapidMiner, Knime, Matlab, and Scikit-learn, as well

as six machine learning techniques: Logistics regression, k-NearestNeighbour, ANN, SVM, IRF, and NB. The method estimates accuracy, sensitivity, and specificity, and ANN is shown to be the best model for heart disease classification among the comparison tools when tested on the University of California dataset.

Author	Title	Methodology	Remarks
Zhang, et al.	Support vector machine combined with magnetic resonance imaging for accurate diagnosis of paediatric pancreatic cancer	Classification – SVM Multi-fold cross-validation	SVM was found to be accurate for diagnosing paediatric PC
Bosch, et al.	Image Classification using Random Forest and Ferns	Random Forest classifier Random Fern classifier	Without optimization-38.7% With optimization -43.7%
Zhang, et al.	A Novel and Efficient Tumor Detection Framework for Pancreatic Cancer via CT Images	Augmented Feature pyramid network Self-adaptive feature fusion Dependencies computation module	Results shows slight improvements in accuracy
Tougui, et al.	Heart disease classification using data mining tools and machine learning techniques	Data mining tools Machine Learning	ANN gives better results than the compared tools – KNN, SVM, NB, RF, Logistic regression
Arslan, et al.	Diagnosis Of Pancreatic Cancer By Pattern Recognition Methods using Gene Fade Profiles	KNN ANN	KNN – 82.7% ANN – 84.6%
Maliha, et al.	Cancer Disease Prediction Using Naive Bayes, K Nearest Neighbor and J48 algorithm	NB KNN J48	NB – 98.2% KNN –98.8% J48 – 98.5%

Table 2.1: Literature survey summary

3. METHODOLOGY

The system contains of four modules there are:

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

3.1Gathering 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 :

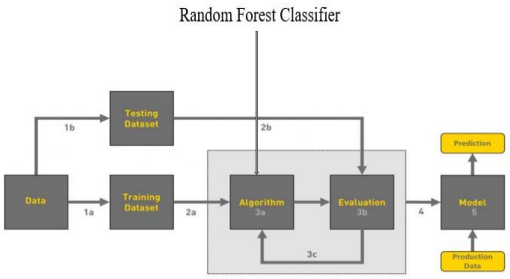
In module 3 the model training are carried out. The dataset is divided into testing and training datasets using the test train split function in Sklearn package. In dataset 70% is considered for training and 30% for testing. Then in this module the classification algorithms, SVM, RF, ETC, DT classifiers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

5.1 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 5.1 Importing modules and libraries

5.2 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("predataset.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	0.032805	60.57900	142.950000	NaN
S102	Cohort2	BPTB	62	M	1	NaN	NaN	9.0	0.21489	0.030360	65.54000	41.088000	NaN

Fig 5.2 Reading Data set

5.3 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.

5.4 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.

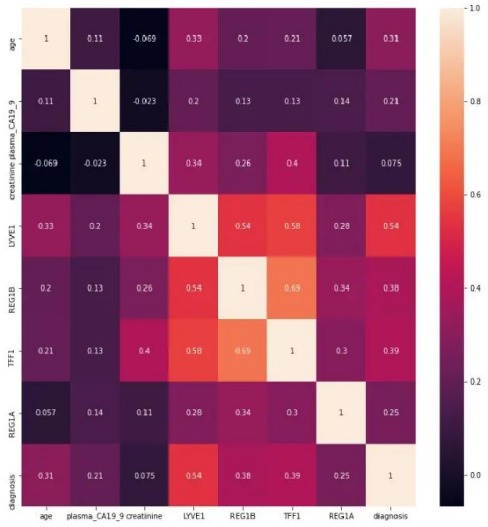


Fig 5.3 Heat Map

5.5 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.

5.6 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 5.5 converting to .pkl file

5.7 APPLICATION BUILDING:

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

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, 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.

6.2 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.

6.3 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.

6.4 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.

6.5 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.

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 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.

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

7.4Biomarker 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.

8. RESULT:

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:

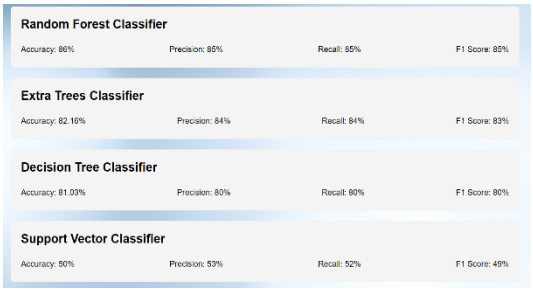
• From the start menu, launch the anaconda prompt.
• Open the folder containing your Python script.
• Now enter the command "python app.py"
• Go to the localhost to view your web page.

• 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.



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.

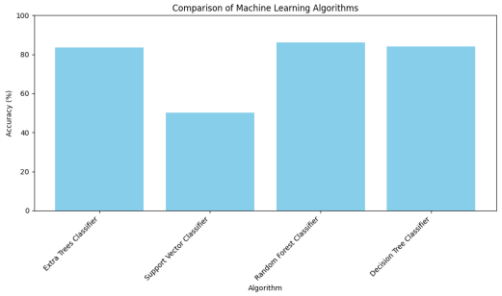


Fig 9.1 Comparison of Algorithms

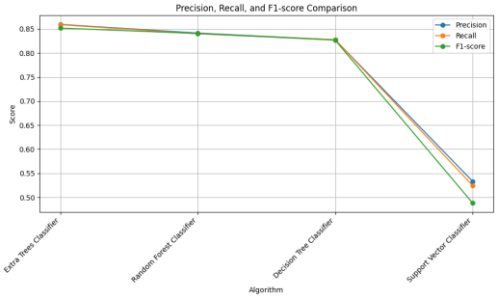


Fig 9.2 Precision, Recall, and F1-score Comparison

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 AteneiParmensis, 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.