IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Cancer Cell Classification using Scikit-Learn

Mrs. Harshada Suraj Dandage[1] and Mrs. Snehal Mahesh Katke[2]
Lecturer, Computer Technology, Solapur Education Society's Polytechnic, Solapur, India[1]
Lecturer, Computer Technology, Solapur Education Society's Polytechnic, Solapur, India[2]

Abstract: *Timely detection and precise diagnosis of breast cancer are crucial for enhancing treatment success and increasing survival chances. This project aims to develop a machine learning model utilizing the Scikit-learn library to classify cancer cells as benign or malignant, based on the Breast Cancer Wisconsin dataset. The dataset comprises various features derived from digitized images of breast cancer cells, including measurements of cell size, shape, and texture*

Keywords: Scikit-learn

I. INTRODUCTION

Cancer remains one of the leading causes of mortality worldwide, underscoring the need for innovative approaches in diagnosis, treatment, and management of the disease. Effective cancer classification is a critical step towards personalized medicine, enabling healthcare professionals to tailor treatment plans based on the specific characteristics of a patient's cancer. Machine learning, particularly supervised learning techniques, has shown great promise in improving the accuracy and efficiency of cancer classification. This project utilizes Scikit-learn, a robust machine learning library in Python, to build a model that classifies cancer cells using a variety of biological and clinical attributes. By leveraging various classification algorithms, we aim to develop a model that is not only efficient but also interpretable, enabling better understanding and insights into cancer pathology. To achieve this, we utilize a dataset consisting of cancer cell characteristics, such as cell size, shape, and texture, along with their corresponding classifications as benign or malignant. Through data preprocessing, feature selection, and model training, we aim to identify the most effective classification techniques for distinguishing between different cancer types. The insights gained from this project will not only contribute to the advancement of cancer classification techniques but also demonstrate the practical application of machine learning in the healthcare domain. By combining computational power with medical knowledge, this project aspires to pave the way for more accurate and timely diagnosis, ultimately enhancing patient outcomes and advancing cancer research.

II. OBJECTIVE

To build a machine learning model using Scikit-learn to classify cancer cells as benign or malignant using the Breast Cancer Wisconsin dataset. The aim is to assist in early detection and diagnosis of breast cancer.

III. DATASET DESCRIPTION

Dataset Used:
Breast Cancer Wisconsin Diagnostic Dataset – Available in Scikit-learn's datasets module.
Features:
Total samples: 569
Classes:
0 = Malignant
1 = Benign
Features: 30 real-valued input features such as:
Radius
Texture (standard deviation of gray-scale values)
Perimeter

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Area
Smoothness
etc.
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()

IV. METHODOLOGY

STEP 1: IMPORT LIBRARIES

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

STEP 2: LOAD AND EXPLORE DATA

```
data = load_breast_cancer()
X = data.data
y = data.target
```

STEP 3: DATA PREPROCESSING

Train-test split (80-20)
Feature scaling using StandardScaler

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42,
shuffle=True)scalerStandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

STEP 4: MODEL BUILDING

Using Random Forest Classifier:

```
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

V.Model Evaluation

Metrics Used:
Accuracy
Confusion Matrix
Precision, Recall, F1-Score

```
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Sample Output:

lua
CopyEdit
Accuracy: 0.9649
Confusion Matrix:
[[40 3]
[1 70]]
Precision: 0.96, Recall: 0.97, F1-score: 0.96

V. CONCLUSION

The Random Forest model achieved over 96% accuracy on the test set.
High precision and recall values for both classes indicate the model is robust.
The project demonstrates the feasibility of using machine learning for early cancer diagnosis using structured medical data.

VI. TOOLS & TECHNOLOGIES USED

Tool	Purpose
Python	Programming language
Scikit-learn	ML library for modeling
Pandas, NumPy	Data manipulation
Matplotlib, Seaborn (optional)	Data visualization
Jupyter Notebook	Development environment

VII. FUTURE WORK

Implement cross-validation and hyperparameter tuning
Compare different ML algorithms (SVM, Logistic Regression, XGBoost)
Deploy the model using a web interface (e.g., Flask or Streamlit)
Use deep learning (e.g., TensorFlow/Keras) for comparison

REFERENCES

[1]. Scikit-learn documentation
[2]. UCI Machine Learning Repository
[3]. Dua, D., & Graff, C. (2019). The Breast Cancer Wisconsin (Diagnostic) dataset is available through the UCI Machine Learning Repository and is commonly used for evaluating classification algorithms.University of California, Irvine. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[4]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://www.statlearning.com/
[5]. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.