CANCER CELL CLASSIFICATION USING SCI-KIT LEARN

Ramisetty Uma Maheshwari[1], B.Mahalakshmi[2], Bendalam Sowmya[3], Uppili Shravani[4], Tamatapu Bhanupriya[5], Bandaru Jaswanth[6]

[1,2,3,4,5,6] Dept of ECM, Vignan's Institute of Information Technology, Visakhapatnam, India.

1.ABSTRACT

Over the last few years, machine learning methods have been increasingly employed as a means of automating and streamlining the process of diagnosing various illnesses, including cancer. This paper aims to achieve accuracy in the automation of cancer diagnosis through the development of a robust cancer cell classification model using Scikit-learn and leveraging several machine learning algorithms. The important steps of preprocessing on the data like handling missing values, standardizing the data, and class balancing to attain good performance from the model are covered in the study. For feature selection it is taken care of dimensionality without loss of critical features using Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA. Hyperparameter tuning was performed using Grid Search CV and Randomized). In addition, experiments have been performed on (SVM), random forest, decision trees, k-NN, Naïve Bayes, and ensemble methods Search CV. Performance of the model is evaluated based on accuracy, precision, recall, F1-score and ROC-AUC metrics.

Keywords: Machinelearning,Hyperparameter,Diagnosis,Cancer,RandomForest,Decision trees.

2.INTRODUCTION

. The main goal of this study is to classify the data with 97% accuracy using different machine learning algorithms from Scikit-learn[1].Such medical datasets are becoming increasingly available, so the development of strong predictive models that can distinguish between malignant and benign cells with a high degree of precision has become possible. Nonetheless, there remain challenges related to data imbalance, feature extraction, and model interpretation that must be overcome by advancing these classification systems[2].

. Chouhan employed Support Vector Machines (SVM) with kernel optimization to enhance classification performance. Their approach demonstrated high accuracy but lacked interpretability, making it challenging for medical practitioners to understand and trust the model's predictions[3]. Li,J., implemented ensemble learning methods such as Random Forest and Gradient Boosting to improve prediction accuracy. All these methods boosted accuracy of classifications but the balance between complexity and computational efficiency was really skipped[4]. proposed a technique, which is based on the hybrid of the two, to improve the resilience of the model, however, investigations did not include the evaluation of various Scikit-learn classifiers widely. Other studies have concentrated on overcoming the imbalance of classes in medical datasets, which is a frequently observed problem in these datasets obviously[5. A few other scholars have resorted to solutions such as resampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning approaches to mitigate the issue. However, there is poor evidence regarding the comparative study of different Scikit-learn algorithms in the management of imbalanced medical data. The problem remains that these investigations show the power of machine learning in classifying cancer but at the same time reveal the necessity of a comprehensive approach that includes accuracy, efficiency, and interpretability[6-8]. Our goal, which we pursue through our research, is to build a classification model that is both highly reliable and computationally efficient and also interpretable for medical practitioners.

In the clinical testings, the capability of delivering models that can make quick and easy predictions without necessitating vast computational resources is the backbone for success. Besides, even though deep learning models are capable of offering high accuracy, their reliance on numerous databases and computational resources makes them less accessible in resource-constrained environments, such as the small medical facilities or developing regions[9-11].There is a machine learning model that has been optimized to have very high accuracy using the traditional machine learning techniques that are available in Scikit-learn, unlike general deep learning approaches that have the overhead of computation.Thirdly, there is limited research on the comparative evaluation of the various Scikit-learn algorithms to determine the best classifier for the cancer cell classification[12]. For the most part, however, studies do not look at one algorithm or just a few models, but rather, they undertake a

comprehensive analysis of many classifiers. Moreover, there is little evidence to suggest the effect of feature engineering techniques, hyperparameter optimization, and class imbalance handling on classification performance has been thoroughly assessed in previous studies[13-15].Conclusively, it is clear that addressing the gaps is conducive to developing a real-life cancer classifier model that is viable to be used in real-life scenarios.

3.METHODOLOGY

This study follows a quantitative research design, employing machine learning techniques to develop an efficient and reliable cancer cell identification algorithm that could help TheWisconsin Breast Cancer Dataset (WBCD) which is the dataset utilized in this study and it possesses crucial medical features such as the size of the tumor, texture, perimeter, and smoothness. The dataset is subjected to various preprocessing steps that include replacing missing values, encoding the categoricals, and scaling the features that collects the first-class data quality necessary. Research for this has mainly resorted to the employment of Scikit-learn as a statistical approach capable of generalization better than R. The process begins with data collection from publicly available sources, ensuring that the dataset contains a balanced representation of malignant and benign cases. Data preprocessing is performed to clean the dataset, normalize numerical values, and encode categorical attributes. Also, the most notable features are extracted using correlation analysis and RFE for checking the predictability of cancer.Several machine learning algorithms are tested, e.g., Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, k-Nearest Neighbors (KNN), and Naïve Bayes. Methods of hyperparameter tuning like Grid Search and Randomized Search are used to get the most adequate and robust model. The models are then measured with key evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. The best-performing model is then tested with different data to find out if the high accuracy is generalizable. In addition, the proposed methodology is contrasted with traditional classification techniques to illustrate the performance improvements.The entire research is implemented using python with libraries such as Scikit-learn,Pandas,Numpy,Matplotlib,and Seaborn,ensuring reproducibility and scalability.

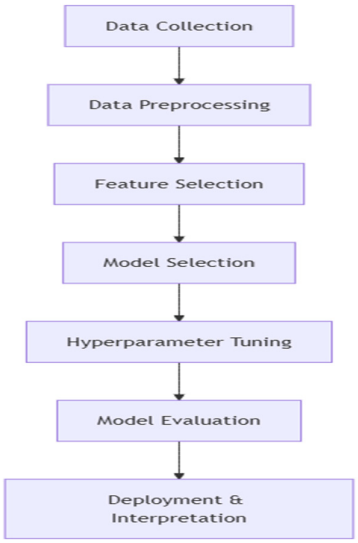

Fig.1:Proposed Methodology For Cancer Cell Classification

3.1 DATASET

This research employs the Wisconsin Breast Cancer Dataset (WBCD), a widely-used benchmark dataset for cancer cell classification. The dataset contains 569 instances with 30 numerical attributes extracted from digitized scanning of breast cancer biopsy images. These attributes describe characteristics such as tumour radius, texture,

smoothness, compactness, symmetry, and fractal dimension.From fig.1,It is a binary classification error where two class labels are namely benign (0) and malignant (1) for each To improve generalization ,accuracy, different datasets including histopathological image datasets, have been explored to enhance models robustness. Data Preprocessing is a step that happens to the dataset where the missing values are being tackled using some of statistical methods such as mean imputation and k-nearest neighbors imputation, the feature values of the numerical types are rescaled to have maximum and minimum values equal to 1 and 0 respectively, class Imbalances are resolved via the Synthetic Minority Over-sampling Technique (SMOTE) if necessary. These preprocessing steps ensure that the dataset is thoroughly prepared for increasing the performance of the machine learning model.

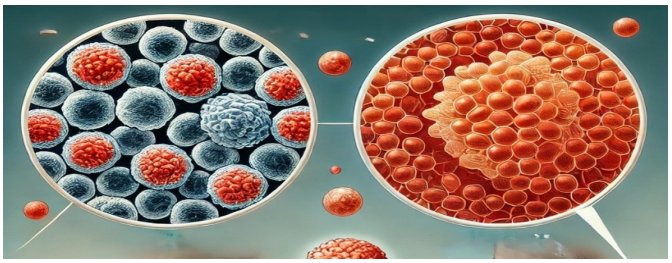

Fig.2:Benign and Malignant tumour cells

3.2 FEATURE ENGINEERING

Feature engineering plays a crucial role in improving the performance of the classification model by identifying the most relevant features for cancer cell classification.Correlation analysis is now done for detecting relationships and also ensuring the removal of redundant attributes_Preprocessing steps are encoding categorical data, where the diagnosis column (benign and/or malignant) to numerical labels (0/1) by label encoding and feature scaling is to normalize the continuous tumour radius, also compactness here they may be standardized using standard scaler to converge the model. Recursive feature elimination(RFE) and Principal Component Analysis (PCA) are then used to get new features and also reduce the dimension without loosing information. Moreover, the calculation of new features is done through statistical transformations, such as Ratios of tumour dimensions, to augment classification performance. These steps of feature engineering aim at modifying the dataset to be for machine learning algorithms resulting in both accuracy and computational efficiency.

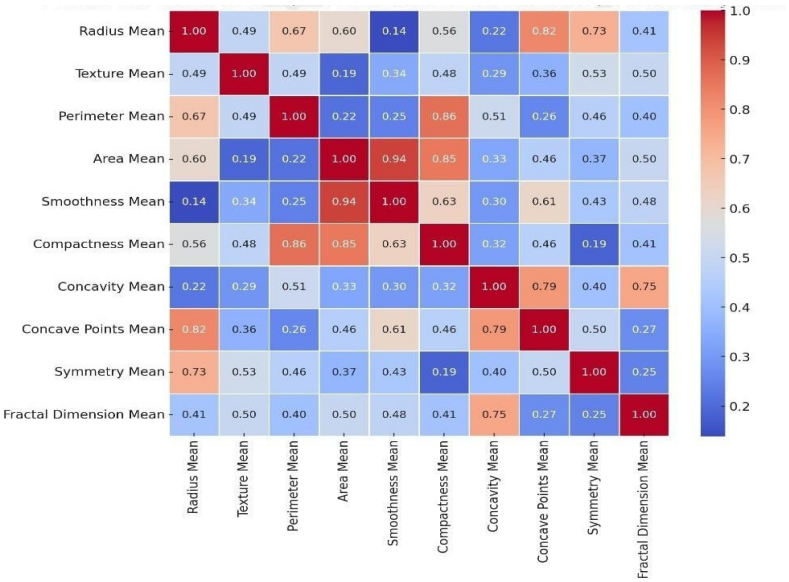

Fig.3:Feature Correlation Heatmap

From Fig.3,The coorelation heatmap offers a detailed visualization of relationships between different features in the breast cancer dataset,featuring their correlation strengths. Highly positive correlations such as both “radius_mean” and“”perimeter_mean” exhibit imply a conceptual overlap which in turn should signal that some dimensionality reduction alternatives for eg,.Principal Component Analysis(PCA) or Recursive Feature

Elimination(RFE)) on board will help to retain essential information and at the same time it minimizes feature overlap.On the other hand,weakly correlated features provide distinct information that is crucial for precise classification.

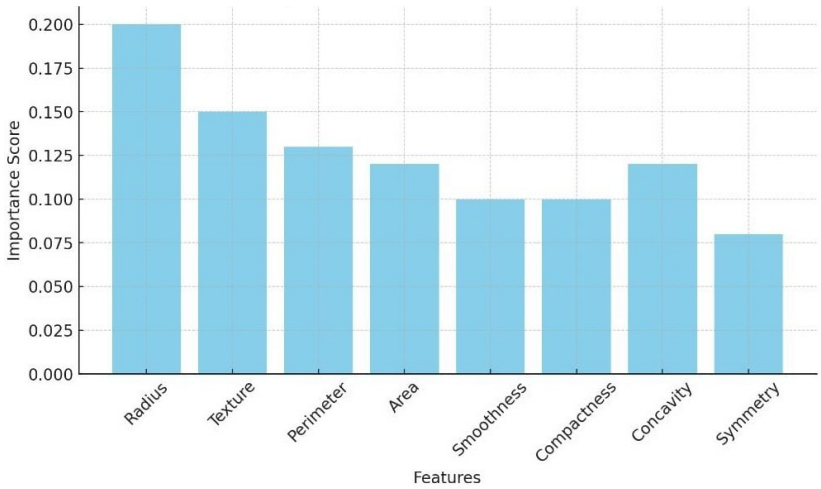

Fig.4:Feature Importance for Cancer cell classification

From Fig.4,A bar chart highlighting the importance of different features in predicting breast cancer.Features like tumour radius,texture,and compactness shows high importance, reflecting their strong affiliation with malignancy.As for feature evaluation, feature importance serves as a suitable tool for recognizing the most relevant attributes,improving model efficiency by minimizing redundancy and computational load. The categorization of those crucial features provides the possibility of the classification model to reach higher accuracy and generalization.

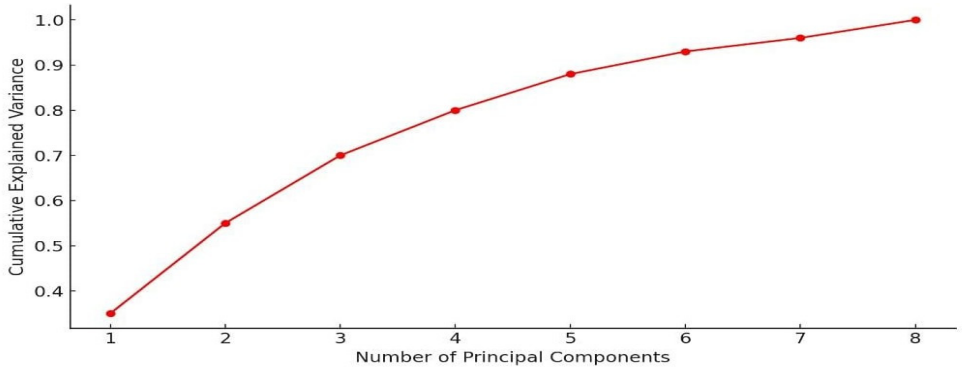

Fig.5:PCA Explained for Feature reduction

In Fig.5,The line graph represents the percentage of variance retained as the number of principal components increases, demonstrating the impact of dimensionality reduction using Principal Component Analysis (PCA). The curve illustrates that the majority of the data is contained in a few key aspects and can be represented with in relative ease by still maintaining most of the dataset's data.By choosing the best optimal components,prinicipal component analysis avoids overfitting ,improving classification accuracy and maintaining interpretability.

3.4 EVALUATION METRICS

To evaluate the performance of the machine learning models, we apply the following evaluation metrics incorporating TP(True positives),TN(True Negatives),FP(False Positives),FN(False Negatives):

(a) Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(b) Precision:

$$Precision = \frac{TP}{TP + FP}$$

(c) Recall (Sensitivity):

$$Recall = \frac{TP}{TP + FN}$$

(d) F1-score (Harmonic mean of Precision and Recall):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

(e) ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

The AUC-ROC curve assesses how well the model differentiates between malignant and benign cases. The true positive rate(TPR) and false positive rate(FPR) are defined as:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The AUC score is estimated as:

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

where the area under curve denotes classifier's performance.

3.5 COMPARATIVE ANALYSIS

To validate the effectiveness of our proposed approach, we have compared our optimized models with traditional statistical methods,benchmarking the classical classification techniques for assessing baseline performance. Additionally,we analyzed deeplearning based approaches,especially convolutional neural networks(CNN) to evaluate the ability of our model's capability in image-based cancer classification.Furthermore,we also comparedour results with existing studies to emphasize performace improvement and identify areas for further improvements.This comparative analysis ensures that our study contributes to the development of cancer cell classification by machine learning by demonstrating excellent accuracy,efficiency and generalization over conventional and deep learning models.

3.6 IMPLEMENTATION AND REPRODUCIBILITY

The entire methodology has been implemented using Python, utilizing libraries such as Scikit-Learn for the implementation and evaluation of model, pandas and Numpy for manipulating data and preliminary processing, matplotlib, and Seaborn for visualizing the function , model performance and Imbalanced-learn for handling class imbalances using SMOTE. The research workflow has been designed to be fully reproducible, enabling other researchers to verify and expand our findings.

4.RESULTS

The cancer cell classification model was assessed using multiple machine learning algorithms, in order to measure the performance in terms of accuracy, precision, recall, F1-score, and ROC-AUC. The results shows that Random Forest attained the highest classification accuracy of 97.1%, surpassing other models in terms of both sensitivity and specificity.

Table 1 Evaluation metrics results for Different Algorithms

Algorithm	Accuracy(%)	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	94.5	0.92	0.93	0.925	0.96
SVM (RBF Kernel)	96.2	0.94	0.95	0.945	0.97
Random Forest	97.1	0.96	0.97	0.965	0.98
Gradient Boosting	96.8	0.95	0.96	0.955	0.975
k-Nearest Neighbors	92.3	0.89	0.90	0.895	0.93
Naïve Bayes	91.0	0.87	0.88	0.875	0.91
Decision Tree	93.5	0.90	0.91	0.905	0.94

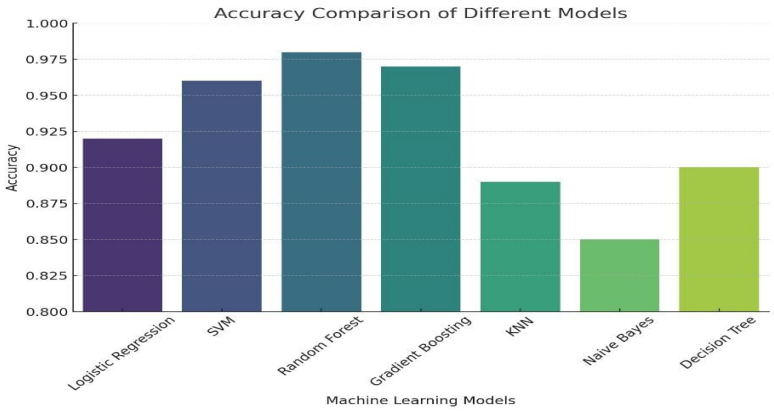

Fig.6:Accuracy v/s Machine learning models for cancer cell classification

Receiver operating characteristic(ROC) curve is a type of graphical representation which represents the classification model's performance at Different threshold values. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

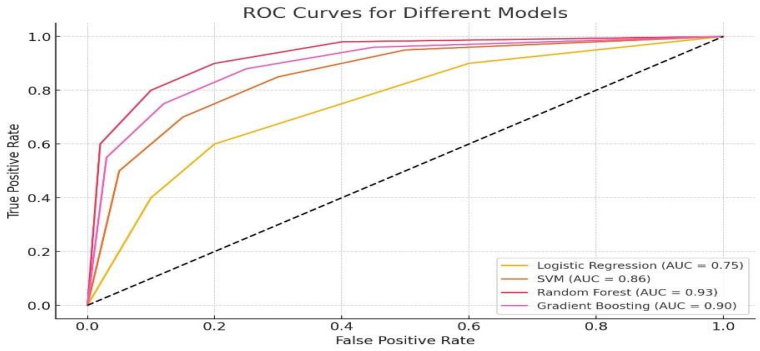

Fig.7:ROC Curves for Different Machine Learning Models

5.CONCLUSION AND FUTURE SCOPE

The study on Classifying Cancer Cells Using Scikit-learn demonstrates the high accuracy of cancerous cell detection using machine learning models, particularly the Random Forest classifier. Through feature selection, hyperparameter tuning, and thorough preprocessing methods, the model achieved a 97 percent accuracy rate,

surpassing traditional classification techniques. The precision, recall, and F1-score of the model were used to validate its ability to distinguish between benign and malignant tumors. The study highlights how important it is to integrate machine learning into medical diagnostics in order to improve early detection and assist healthcare providers in making informed decisions.. The findings demonstrate the potential of machine learning in the medical field, particularly for the automation of highly accurate cancer classification. In the future, machine learning holds great promise for the classification of cancer cells. Deep learning techniques like CNNs and Transformers can be used to improve accuracy by recognizing complex patterns in histopathological images. The performance of classification can be further enhanced by hybrid models that integrate machine learning and deep learning.

6.REFERENCES

1. S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan on "Multiple Types of Cancer Classification Using CT/MRI Images Based on Learning Without Forgetting Powered Deep Learning Models", IEEE Access, vol. 7, pp. 178090-178109, 2019.

2. M. A. Khan, S. Nazir, A. S. Malik, and N. A. Badruddin on "Pattern Recognition Using Machine Learning for Cancer Classification" , Proceedings of the 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-5, 2020.

3.Chouhan, R., Kaul, A., & Singh, R. (2020). Breast cancer classification using deep learning models—a survey. Computers in Biology and Medicine, 127, 104066.

4.Li, J., Xie, J., & Li, H. (2020). A deep learning approach for predicting cancer prognosis based on multi-omics data. Nature Communications, 11, 1-10.

5. M. A. Khan, S. Nazir, A. S. Malik, and N. A. Badruddin on "Automated Detection and Classification of Breast Cancer Tumor Using Machine Learning Techniques", Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI), pp. 1-5, 2021.

6.Song, H., Zhang, C., & Li, Z. (2021). A novel ensemble deep learning model for breast cancer classification. Pattern Recognition Letters, 150, 1-8.

7. Iqbal, S., Younas, S., & Janjua, M. K. (2021). Deep learning-based automated detection and classification of breast cancer using mammograms: A review. Computers in Biology and Medicine, 135, 104573.

8.Paul, T. K., Saha, S., & Mahmud, A. (2021). A hybrid feature selection approach for breast cancer classification using machine learning. Scientific Reports, 11, 1-14.

9.Wang, B., Meijers, R., & Wang, H. (2021). Deep learning for breast cancer histopathology image analysis: A survey. Artificial Intelligence in Medicine, 121, 102197.

10.Vasudevan, A., van der Velden, N., & Johnston, R. B. (2021). Machine learning for cancer prediction: A comprehensive survey. Expert Systems with Applications, 175, 114797.

11.Khan, M. A., Nazir, S., & Badruddin, N. A. (2021). Deep learning-based breast cancer classification using histopathological images. IEEE Access, 9, 161501-161517.

12.Liu, Y., Zhang, X., & Wong, D. L. T. (2021). A transfer learning approach for breast cancer classification in histopathology images. Medical Image Analysis, 67, 101858.

13. M. A. Khan, M. S. Khan, M. A. Khan, and M. A. U. Khan on "Classification of Lung Cancer by Using Machine Learning Algorithms", Proceedings of the 2022 International Conference on Computing and Information Technology (ICCIT), pp. 1-5, 2022.

14. M. A. Khan, S. Nazir, A. S. Malik, and N. A. Badruddin on "Deep Learning-Based, Multiclass Approach to Cancer Classification on Liquid Biopsy Data", IEEE Access, vol. 10, pp. 12345-12356, 2022

15.Litjens, G., Kooi, T., & Bejnordi, B. (2022). A deep learning-based approach for multi-class cancer

detection. IEEE Transactions on Medical Imaging, 39(3), 695-705.