INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 483-487

# AN INTEGRATED FRAMEWORK FOR EARLY DETECTION OF LUNG, BREAST AND PROSTATE CANCER

**Sheba Kezia Malarchelvi P. D[1], Sridevi J[2], Ramabathren L[3], Preethi Mahaeswari G[4], Thiyagarajan S[5]**

[1]Professor, Department of CSE, Saranathan College of Engineering, Tiruchirappalli, Tamil Nadu, India

[2,3,4,5]Student, Department of CSE, Saranathan College of Engineering, Tiruchirappalli, Tamil Nadu, India

DOI: https://www.doi.org/10.58257/IJPREMS31739

## ABSTRACT

Among the various types of diseases, cancer is considered as one of the deadly diseases in the world. lung, prostate, and breast cancer are some of the cancer types thatare contributing most to the mortality rate. In order to overcome this, our proposed research work aims at analyzing the performance of various machine learning algorithms in the early detection of three types of cancers namely breast cancer, lung cancer and prostrate cancer. The machine learning models were evaluated as well as compared based on Performance Metrics parameters like Accuracy, Precision,Recall, F1Score. The experimental results suggest that the Logistic Regression offers the highest Accuracy for the prediction of Lung Cancer and the Decision Tree for the prediction of Prostate Cancer and Breast Cancer. The proposed framework has been integrated with python script through the Python Flask Framework. This allows the framework to fetch the user inputs/responses with the help of forms and according to the user inputs the framework will select appropriate classifier and will give the result as benign or malignant. Hence, this framework would assist the physicians as well as the users in the early detection of cancer which serve as a warning signal for further investigation, treatment and mitigation of cancer at early stages. This will in turn help in reducing the mortality rate of cancer patients.

**Keywords:** Cancer detection, Machine Learning Algorithms, K-NN, Logistic Regression, Random Forest, Decision Tree

## 1. INTRODUCTION

Every year growing exponential number of patients throughout the globe has cancer patients as the maximum in number. Cancer has turned out to be a major threat to human life. As per the WHO Survey report, these (Breast,Lung, Prostate) cancers have affected the maximum number of patients and have been seen as dangerous due to which Mortality Rate has rapidly increased because it's usually late for doctors to detect cancer. To improve cancer screening our study has made an effort through this research where the study implemented 3 major cancers detection machine learning models. In order to build ML models, this study has used many classification algorithms like Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Simple Linear Regression (SLR), K- Nearest Neighbour (K-NN), Support Vector Machine (SVM), Naive Bayes (NB).

Moreover, these are widely used algorithms to train and test the datasets algorithms, the best accurate algorithm has been used at the backend to make predictions where this study has made a web page using the Python Flask API to gather the inputs from from end-users. Through the best accurate model, this study is focusing to differentiate Benign and Malignant tumours where this study is classifying patients into non- cancerous (Benign) and cancerous (Malignant). On the other hand, the study has a module for making analysis on their own data through the web API. For that end-user has to submit data link where they get textual analysis (where missing case are being handled efficiently and showing all the relevant information), Visualization of the data with a single click. As this study has two major modules (i.e., making prediction and making analysis) in the study so end-user can have insight over data and can have predictions over 3 different cancers according to user response. Hence, with this study, this paper has tried to detect early cancer in humans and help them to reduce the serious impact on human life. Moreover, this concept willalso help to save lives, time, and money too.

## 2. LITERATURE REVIEW

O. Gunaydin, M. Gunay, O. Şengel [1], implemented Comparison of Lung Cancer Detection Algorithms on the Standard Digital Image Database, Japanese Society ofRadiological Technology with the 5 different types of Machine learning Algorithms (K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, Artificial Neural Network) and evaluated the model on theparameters such as Accuracy, Precision, Recall and Confusion Matrix.

M. Srivenkatesh [2], in 2020 analyzed a Prostate Cancer Dataset having 100 samples and 10 features to build a prediction model on Prostate Cancer by applying different supervised learning techniques (K-Nearest Neighbors,

Support Vector Machine, Logistic Regression, Naïve Bayes, Random Forest) and checked the efficiency of models using Performance Measurement Metrics like (Confusion Matrix, Accuracy, Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Mean Absolute Error (MAE) and Kappa Statistics) to find the best fit model for prediction. D. E. Gbenga, N. Christopher, D. C. Yetunde [3], analyzed the Wisconsin Breast Cancer dataset (Diagnostic) having 569 observations and 32 attributes using 10-fold cross- validation and built the model using Machine Learning algorithms Support Vector Machine, Naïve Bayes, K- Nearest Neighbors, Simple Linear Logistic Regression, AdaBoost Fuzzy Unordered Role Induction, Radial Based Function, Decision Tree which evaluated based on Accuracy, Precision, and F1-Score. Radhika P R, Rakhi. A. S. Nair, Veena G [4], analyzed a dataset on Lung Cancer obtained from UCI Machine Learning Repository to build Lung Cancer Detection Model using supervised learning techniques (Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree) and obtained the Support Vector Machine model as the best-fit model for prediction. J. Nuhic, J. Kevric [5], in 2020, built Prostate Cancer detection model on the dataset from Zhou W. 387 samples (included dataset) with Machine learning techniques Simple Logistic, Naive Bayes, K-Nearest Neighbor, Logistic model trees, Random Forest Classifier, Random Committee, Attribute Selected Classifier for attribute selection, Ada Boost Classifier which were evaluated based on Sensitivity and Specificity, given AdaBoost Classifier as the best fit classifier. S. Sharma, A. Aggarwal, T. Choudhury [6], in 2018 built Breast Cancer Detection using the Wisconsin Breast Cancer dataset (Original) having 669 observations and 10 attributes segregated with 10-fold cross-validation (90% training, 10% testing) applied to Supervised Learning techniques Naïve Bayes, Random Forest, K-Nearest Neighbor. The efficiency of the models was compared using Precision, Recall, F1- Score to find the best fit model for prediction. D. Bazazeh, R. Shubair [7], implemented Comparative Study of Machine Learning Algorithms for Breast Cancer on Wisconsin Breast Cancer dataset (Original) collected from UCI Machine Learning Repository using supervised learning techniques such as Support Vector Machine, Random Forest, Bayesian Network and being evaluated on the Accuracy, Specificity and Precision parameters and obtained Support Vector Machine model as the most accurate model. M. Amrane, S. Oukid, I. Gagaoua, T. Ensari [8], in 2018 built Breast Cancer Classifier Wisconsin Breast Cancer dataset (Original) having 669 observations and 10 attributes using supervised learning techniques - Naïve Bayesian Classifier, K-Nearest Neighbor which further evaluated on the Accuracy parameter resulted in K-Nearest Neighbor model as the most accurate between the two models. B. Sekeroglu, K. Tuncal [9], implemented Prediction of Cancer incidence rates for the European Continent, using the dataset obtained from the World Health Organization 2018 containing data of European Continent Incidence Rates of 29 Cancer Types from 22 countries and built model with Linear Regression, Support Vector Regression, Long Term Short Memory, Back-Propagation, Radial Basis Machine Learning techniques for gender category wise, different cancer types which would help in predicting the incidence rate for future years. R. Hazra, M. Banerjee, L. Badia [10], built Machine Learning Model for Breast Cancer Classification on the Wisconsin Breast Cancer dataset (Diagnostic) having 569 observations and 32 attributes using supervised learning algorithms – Artificial Neural Network and Decision Tree. The best fit model obtained is the Decision Tree model which can be used for prediction. After going through almost 20 research papers, we begin our study on building an integrated Cancer Detection model which can be used to predict Lung Cancer, Prostate Cancer, and Breast cancer that can be further deployed on the cloud with the help of Python dependencies like Flask API [11], Joblib etc. It can be helpful in the early Detection or Diagnosis of Cancer for humans. Hence, ultimately it would be beneficial in reducing Mortality Rate, spreading awareness regarding the necessity of medical treatment, and saving money for the people as well. Although several researchers have evaluated the performance of various machine learning algorithms, there is a lack of an integrated framework to be used by physicians as well as end users for different kinds of cancers. Hence, this work aims at exploring different machine learning algorithms and identifying the most suitable algorithm for detection each type of cancer namely lung cancer, breast cancer and prostate cancer respectively through evaluation using three different data sets as well as integrating the best suited models in the framework.

## 3. METHODOLOGY

The proposed system analyzes the performance of four classifiers namely Logistic Regression, Decision Tree, KNN and Random Forest for breast cancer, lung cancer and prostate cancer datasets. This is depicted in the following Figure 1.

### 2.1 Data and Pre-Processing

In this study, 3 different datasets have been used for 3 major cancer detection namely Lung cancer, Breast cancer and Prostate cancer. In order to model breast cancer detection, Breast Cancer Wisconsin (Diagnostic) dataset has been used which contains 569 observations/instances with 33 attributes. The Breast cancer dataset is divided into two parts for training and testing with a ratio of 7:3 (i.e., 70% training set and 30% test set) and the Lung and Prostate cancer datasets are divided in the ratio of 8:2 (i.e. 80% train set and 20% test set).

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 483-487

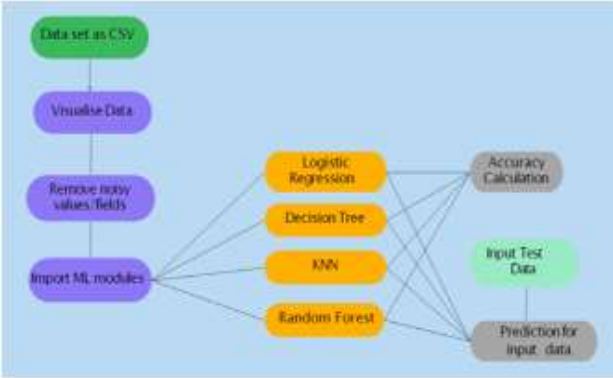


**Figure 1:** Proposed System Architecture

### 2.2 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor is one of the supervised learning algorithms used to classify the given data set on labeled inputs to train the data and find the likelihood or similarity between the examples which can be used to predict the class of unknown samplesof the dataset. In Breast Cancer detection, taking k = 3 thenit gives the highest accuracy 90.5% over the dataset (split ratio 7:3) having 398 observations/instances for training set and 171 observations/instances for test set with 33 attributes each. Similarly, in Prostate Cancer detection, k =5 then it gives the best accuracy as 89.4% over the dataset (split ratio 8:2) having 80 observations/instances for training set and 20 observations/instances for test set with 10 attributes each. In the same way, for Lung Cancer detection, k = 3 then it gives the accuracy 90.1% over the dataset (split ratio 8:2) having 248 observations/instances for training set and 61 observations/instances for test set with 16 attributes each. Table 1 and Figure 1 show the comparison over the fitting of KNN on all three (Breast, Lung, Prostate) cancer datasets.

### 2.3 Random Forest (RF)

Random Forest is an ensemble-based method belonging to the supervised learningcategory used for classification it uses multiple classifiers (decision trees) to solve the complex problem and to improve the performance of the existing model. In this work for Breast Cancer detection, n estimators = 50, max features ='auto', max depth=9 gives the best accuracy 91.8% over the dataset (split ratio 7:3) having 398 observations/instances for training set and 171 observations/instances for test set with 33 attributes each. Similarly, in Prostate Cancer detection, n estimators=100, random state=9, n jobs = - 1 gives a maximum accuracy of 89.4% over dataset (split ratio 8:2) having 80 observations/instances for training set and 20 observations/instances for test set with 10 attributes each. In the same way, for Lung Cancer detection, "max features": [1,3,10], "min samples split" [2,3,10], "min samples leaf":[1,3,10], "boot strap":[False], "n estimators":[100,300], "criterion":["gini"] gives accuracy of 90.1%over the dataset (split ratio 8:2) 248 observations/instancesfor training set and 61 observations/instances for test set with 16 attributes each.

### 2.4 Decision Tree (DT):

Decision Tree learners are powerful classifier which comes under the category of supervised learning that utilizes a tree structure to model the data set and is based on the principle of Divide and Conquer, where the internal nodes represent features, branches as decisions, and finally, leaf nodes represent outcomes of the decision. In Prostate Cancer detection, random state = 9 gives best accuracy of 92% over dataset (split ratio 8:2) having 80 observations/instances for training set and 20 observations/instances for test set with 10 attributes each. Similarly, in Lung Cancer detection, "min samples split": range (10, 500, 20), "max depth": range (1, 20, 2) gives best accuracy 92. 4% over thedataset (split ratio 8:2) 248 observations/instances for training set and 61 observations/instances for test set with16 attributes each and breast cancer, the maximum accuracy achieved is 92.6%.

### 2.5 Logistic Regression (LR):

Logistic Regression is a typeof supervised learning algorithm applied to solve binary classification problems based on the concept of probability.It predicts the output of a categorical dependent feature and gives results in the form of yes or no, 0 or 1, true or false. It uses a cost function as the sigmoid function also known as the Logistic function. In Breast Cancer detection logistic regression with C = 1, max iter =30, multi class='auto' gives best accuracy 98% over the dataset (split ratio 7:3) having 398 observations/instances for training set and 171observations/instances for test set with 33 attributes each. Similarly, in Prostate Cancer detection, C=10 gives the bestaccuracy 90% over dataset (split ratio 8:2) having 80observations/instances for training set and 20 observations/instances for test set with 10 attributes each. In the same way, for Lung Cancer detection, "C": np.logspace(-3,3,7),"penalty": ["l2","l2"] gives the accuracy 91% over the dataset(split ratio 8:2) 248 observations/instances for training set.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 483-487

## 4. PERFORMANCE METRICS

In this study, various performance measures listed below have been used to evaluate the performance of the machine learning models applied to predict different cancers.

a. Accuracy: It is the proportion of correctly classified records in each data collection to the total number of records in that data set as given by the following equation:

Accuracy = (TN + TP) / (TP + TN + FP + FN)

b. Precision: It is the proportion of correctly categorized records to all classified records in the data set calculated using the following equation: Precision = TP / (TP + FP)

c. Recall: It is the proportion of correctly classified records to all records that should be classified:

Recall = TP / (TP + FN)

d. F1 score: It gives the average of both Precision and Recall with following equation:

F1 score = 2 X (Recall X Precision)/ (Recall + Precision)

## 5. RESULTS AND DISCUSSION

Our study has covered supervised machine learning techniques which include mainly the classification algorithms namely Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and K Nearest Neighbor (K-NN).Our study has covered supervised machine learning techniques which include mainly the classification algorithms namely Random Forest (RF), Decision Tree (DT), Logistic Regression (LR) and K Nearest Neighbor (K-NN). This study has been implemented on two types of systems having a configuration as i3 & i5 processors with 4 and 8 GB RAM respectivel. In this study Flask API framework, NumPy, Plotly, Scikit-learn, joblib, jinja2, json, are used. Also, this study has been performedon open-source software namely Visual Studio Code (VS code), Jupyter Note Book to run code. Out of 3 Breast, Lung, Prostate cancer datasets, the breast dataset has a split ratio of 7:3 and the rest two dataset have a split ratio of 8:2 to give the highest accuracy. In breast cancer detection, three classification algorithms have been applied namely LR (98%), KNN (90.5%) and RF (91.8%) and DT (92.6). Among all the three algorithms, LR has offered the best accuracy of 98% for cancer detection. Similarly, for prostate cancer detection, classification algorithms namely LR, KNN, RF and DT have offered accuracy of 90%, 89.4%, 91.3% and 92% respectively. So, among these algorithms, DT has been selected to compute user inputs at the backend of flask web API. In the same way for Lung Cancer Detection, three algorithms have been applied namely LR, KNN, RF and DT have offered accuracy of 91%, 90.1%, 91.5%, and 92.4% respectively. So, LR has been selectedto compute user input at the backend of Flask web API. The values obtained for various performance metrics are presented in table 1 and the same is depicted in Figure 2. Web page view allows end-users to give specificinputs which are collected at the backend by forms (HTML) that provide that data to python script using Flask API. Again python will redirect the date to the webpage for results after computing using the render template method.

**Table 1:** Performance Analysis of LR, KNN, RF and DT Classifiers with Breast, Lung and Prostate Cancer Datasets

| Type of Cancer | Performance Metrics | Classification Technique | | | |
|---|---|---|---|---|---|
| | | LR | KNN | RF | DT |
| Breast Cancer | Accuracy | 98 | 90.5 | 91.8 | 92.6 |
| | Precision | 98 | 91.6 | 92.8 | 94.6 |
| | Recall | 99 | 97.3 | 97.5 | 96.4 |
| | F1 Score | 98 | 94.3 | 95.1 | 95.5 |
| Lung Cancer | Accuracy | 91 | 90.1 | 91.5 | 92.4 |
| | Precision | 98 | 91.4 | 92.7 | 94.5 |
| | Recall | 98 | 96.9 | 97.2 | 96.2 |
| | F1 Score | 98 | 94.1 | 94.9 | 95.3 |
| Prostate Cancer | Accuracy | 90 | 89.4 | 91.3 | 92 |
| | Precision | 97 | 90.6 | 92.3 | 94.2 |
| | Recall | 87 | 97 | 97.4 | 96.1 |
| | F1 Score | 93 | 93.7 | 94.8 | 95.1 |

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 483-487

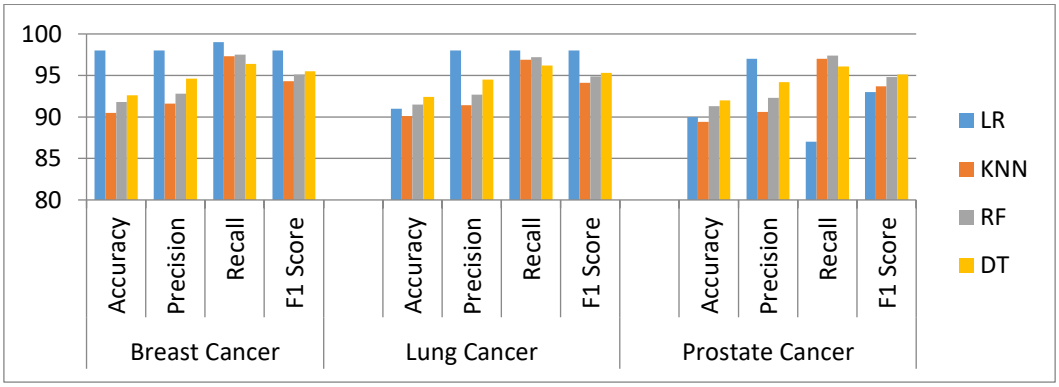


**Figure 2:** Performance Analysis of LR, KNN, RF and DT Classifiers with Breast, Lung and Prostate Cancer Datasets

## 6. CONCLUSION

Our study involved the performance analysis of widely used classification algorithms (Supervised Learning Techniques) with the aim to find the algorithm that provides the best prediction accuracy for three types of cancers namely Breast cancer, Lung cancer and Prostate cancer. Through experimentation with three different datasets for each type of cancer, Decision Tree fits best with the highest performance for lung bancer detection with accuracy value of 92.4% and Prostate Cancer detection with accuracy value of 90%. While, Logistic Regression fits best for Lung Cancer detection with the accuracy value of 92%. Thus, prior detection will help numerous patients to get early medical treatmentswhich will gradually reduce the mortality rate. As future extension, deep learning algorithms would be analyzed and incorporated in the framework to enhance the accuracy. The interface would be improved to directly analyze the clinical images such as MRI scan and CT scans, etc. Also, data privacy mechanisms will be introduced to ensure confidentiality of patient data.

## 7. REFERENCES

[1] O. Gunaydin, M. Gunay, O. Şengel, "Comparison of Lung Cancer Detection Algorithms", Scientific Meeting on Electrical-Electronics & amp; Biomedical Engineering and Computer Science (EBBT), 2019.

[2] Dr. M. Srivenkatesh, "Prediction of Prostate Cancer using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE), 2020.

[3] D. E. Gbenga, N. Christopher, D. C. Yetunde, "Performance Comparison of Machine Learning Techniques for Breast Cancer Detection", Nova Journal ofEngineering and Applied Sciences, 2017.

[4] Radhika P R, Rakhi. A. S. Nair, Veena G, "A Comparative Study of Lung Cancer Detection", IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2018.

[5] J. Nuhic and J. Kevric, "Prostate Cancer Detection Using Different Classification Techniques", International Conference on Medical and Biological Engineering, 2020.

[6] S. Sharma, A. Aggarwal, T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms", International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018.

[7] D. Bazazeh and R. Shubair, "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", 5th International Conference onElectronic Devices, Systems and Applications (ICEDSA),2016.

[8] M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, "Breast Cancer Classification Using Machine Learning", Electric Electronics, Computer Science, Biomedical Engineering's'Meeting (EBBT), 2018.

[9] B. Sekeroglu and K. Tuncal, "Prediction of cancer incidence Rates for the European continent Using machine learning models", Health Informatics Journal, 2021.

[10] R. Hazra, M. Banerjee, L. Badia, "Machine Learning for Breast Cancer Classification with ANN and Decision Tree", 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2021.