International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Review on Cancer Prediction Using Machine Learning Techniques

Srikanth R[1], Tamil Priya D.[2*], Jagadeesan S.[3], Savita P. Patil[4], Anupama K. Ingale[5], Manojkumar Vivekanandan[6], Venkadeshan Ramalingam[7]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract: This comprehensive study aims to conduct a thorough analysis of machine learning methods and applications in cancer prediction. Breast cancer, lung cancer, and colorectal cancer are the three distinct categories of cancer that impact individuals on a global scale. We focus on machine learning (ML) algorithms to predict cancer which would be influenced by various performance measures. Using the most common ML techniques, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Regression, Decision Tree and Naive Bayes we investigate the accuracy of cancer prediction. Our study can serve as an analysis and recommendations regarding the use of machine learning techniques in clinical settings to improve cancer detection and care.

Keywords: *Breast cancer, colorectical cancer, Lung cancer, Machine Learning, Prediction*

1. Introduction

Breast, lung, and colorectal cancer are the top three most prevalent kinds of cancer in worldwide. The most prevalent cancers in universal are lung and breast cancer of about 12.5% and 12.2% of new cases respectively [1]. Colorectal cancer ranks as the third most prevalent malignant neoplasm. of about 1.9 million new cases, making up to 10.7% of all new cases. Fig.1.shows the overall cancer cases in the world wide [1]. Unfortunately, doctors finds difficult to anticipate cancer cases precisely. The algorithms used for machine learning are currently developed to aid medical researchers in determining the type of cancer or illness by leveraging decisive instruments.

1.1. Global Cancer Prevalence in Men

About 15.4% of new cases of cancer taking place in males, lung cancer is the most predominant case occurs in men worldwide. Fig.2. represents number of top three cancers that affecting men most frequently [1].

[1] *SCORE, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA*
[2] *SCO RE, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA*
ORCID ID: 0000-0002-6768-6639
[3] *SCORE, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA*
ORCID ID: 0000-0002-1689-2328
[4]*Dept. of IT, Rajarambapu Institute of Technology, Shivaji University, Sangli, Maharashtra, INDIA.*
ORCID ID: 0000-0003-0171-9239
[5]*Dept. of CSE, Rajarambapu Institute of Technology, Shivaji University, Sangli, Maharashtra, INDIA.*
ORCID ID: 0000-0003-0378-2477
[6]*Dept. of CSE, School of Engineering and Applied Sciences (SEAS) SRM University-AP, INDIA*
ORCID ID: 0009-0005-4874-4994
[7]*Dept. of IT, University of Technology and Applied Sciences-Shinas Branch, Sultanate of Oman.*
ORCID ID: 0000-0001-7631-2156
** Corresponding Author Email: tamilpriya.d@vit.ac.in*

1.2 Global Cancer Prevalence in Women

The three most prevalent types of cancer globally are breast cancer, as well as colorectal cancer, along with lung cancer. Breast cancer is widely recognised as the predominant form of cancer affecting women globally, accounting for around 25.8% of newly diagnosed cases. The total number of breast cancer cases in women given in the Fig.3.

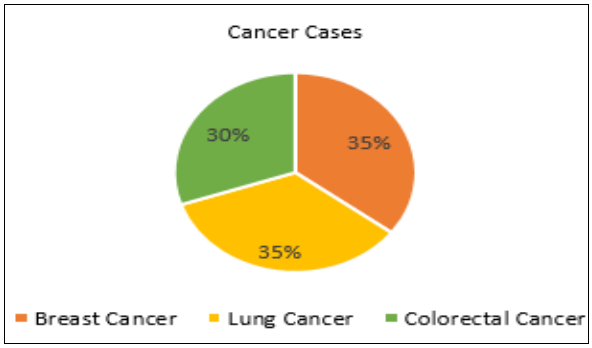

Fig 1. Number of Cancer cases in world wide

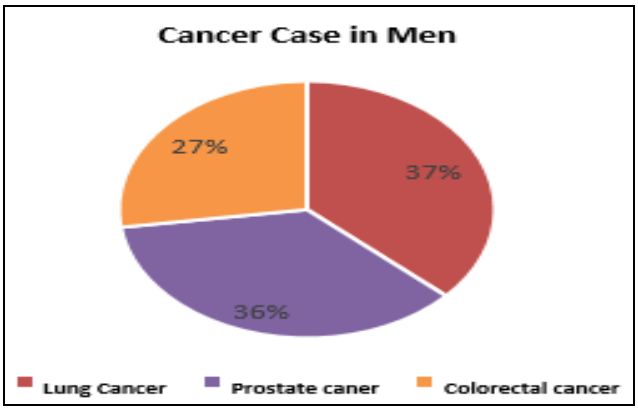

Fig 2. Number of cancer cases in men

Remaining section of this paper is discuss as follows. Part 2 discusses and organizes the methodology for gathering data, cleaning data, analysing data, training and testing algorithms, and obtaining results. Classifiers such as Support Vector Machine (SVM), K-nearest Neighbour (KNN), a naive Bayes approach, Decision Tree as well as Logistic Regression are commonly employed in various academic and research domains. is discussed

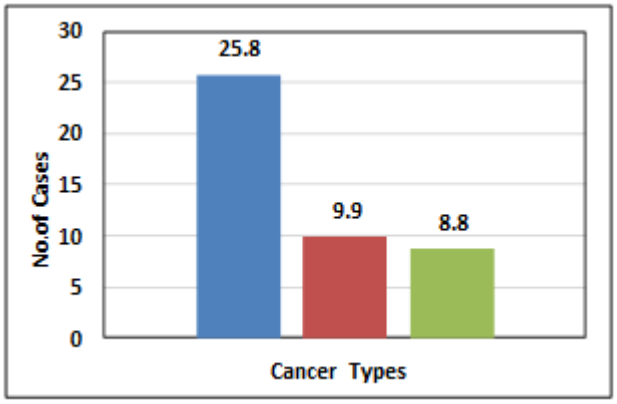

Fig 3. Number of Cancer cases in Women

in part 3 of this article. The discussion of the experimental outcomes is presented in Section 4, focusing on the performance of the model, the performance of the features, and the assessment and analysis of the learning models. The presentation of the results and the discussion is located in Section 5, while the conclusion is provided in Section 6.

2. Methodology

This section provides an overview of several methodologies and approaches employed in the training and testing of data. The paper also examines the diverse array of machine learning methods employed in the prediction of cancer in humans.

The steps involved in making a machine learning model by training and testing them with the help of the algorithm and the dataset is show below in figure 4[23, 28].

2.1 Methods for collecting data

This section discusses various method used for data collection, data cleaning, analyzing data, training and testing data, and in obtaining results.

2.1.1 Collecting the dataset

Kaggle is widely regarded as a highly valuable resource for data specialists and machine learning practitioners seeking datasets. This makes it simple for users to locate, download, and publish datasets [8, 9]. Also, it offers the chance to collaborate with many other machine learning specialists on challenging data science related subjects. Depending on the measures, data can be acquired from a range of sources, including medical records, patient questionnaires, and management databases used to handle payments or provide therapy [14, 30].

2.1.2 Cleaning the dataset

Data cleaning refers to the systematic procedure of eliminating or altering data that is deemed unreliable, incomplete, irrelevant, redundant, or incorrectly formatted, with the objective of rendering it suitable for subsequent analysis. [26]

Step 1**:** Remove irrelevant material
Step 2: Fix structural problems
Step 3: Address any incomplete data from step 3
Step 4: Clean up the data by removing outliers
Step 5: Validate your data

Fig 4. Methodologies used in machine learning

2.1.3 Analysing the data

In the field of machine learning, data interpretation refers to the systematic process of manipulating and modelling data with the objective of drawing meaningful conclusions and facilitating informed decision-making. [22, 29]

In order to analyse cancer in the healthcare system, precise and high-quality data must be provided [27]. In order perform the above process in efficient manner, there should to be a continuous proliferation in healthcare quality and a drop in treatment costs, healthcare must use data in the global economy. [33]

2.1.4 Training and testing the algorithm

Two subsets of datasets are used in machine learning. A piece of our real dataset, referred to as the initial subset's data sets is supplied further into machine learning algorithm to assist it find and understand patterns. The testing data is the name given to the other subset [18, 23, 31].

Typically, the volume of training data exceeds that of testing data. The purpose of this is to furnish the model with an extensive amount of information in order to facilitate its discovery and acquisition of valuable patterns in accordance with user specifications [22]. After being given data from our datasets, a machine learning system extracts patterns and decides what to do with them.

Training data will differ depending on whether you're using supervised or unsupervised machine learning [28].

2.1.5 Obtaining the result

Medical professionals can properly gather a patient's medical history using ML, and they can utilize healthcare management to choose the most pertinent inquiries to ask a patient depending on a number of different parameters. It assists in gathering crucial data and offers a forecast of the most probable scenarios [25].

Frequently, but not universally, the primary aim of machine learning involves the training of a model utilising past data that has been labelled or tagged (i.e., data for which the outcome is already known). This training process enables the model to ascertain the value of a certain variable or classification for a novel data item, where the value of the variable or categorization is now unknown.

3. Classification Algorithms/Classifiers

Various machine learning approaches, such as Support Vector Machine (SVM), K-Nearest Neighbour, Naive Bayes, Decision Tree and Logistic Regression, and Random Forest are employed for disease prediction.

3.1 Support Vector Machine

The supervised machine learning approach applied for both classification and regression problems is the Support Vector Machine (SVM). Nevertheless, the utilisation of categorization difficulties is most commonly observed in this context. In our analysis, it is common practise to graphically display individual data points in an n-dimensional spatial framework, where the magnitude of each feature corresponds to the value of a specific coordinate. The subsequent classification is conducted through the identification of the super-level which firmly differentiates the two groups [19]. With the exception of the linear kernel parameter and the 40 setting for the random state parameter, we typically leave all important parameters at their default settings in order to get the desired outcome [5].

3.2 K-Nearest Neighbour

The K-nearest neighbour method is commonly employed in the fields of pattern detection and clustering. Predictive analysis is commonly utilised in various applications. The K-Nearest Neighbours (K-NN) algorithm is employed to determine the closest neighbouring data points when novel data is provided. [18,27] The K-nearest neighbours (KNN) technique is employed to classify a novel data point by determining its resemblance to the previous data points that have been recorded. The fundamental principle behind the functioning of the K-Nearest Neighbours (KNN) algorithm is the identification of data points in close proximity to the newly inserted point within the machine. The algorithm then separates out the closest points on the basis of the arrival point's distance [1, 6].

3.3 Naive Bayes

The Naive Bayes (NB) classifier is a predictable machine learning method commonly employed for solving classification problems.[26] The classifier is based on the Bayes Theorem, a mathematical principle used to determine the probability of an event occurring given that it has already happened. The method in question is widely recognised as a highly efficient and impactful machine learning technique that is currently employed in various industries and domains. The Naive Bayes algorithm operates under the assumption of independence among all indicators, which is unusual and lacks empirical support [5, 6].

3.4 Decision Tree

The decision tree (DT) is often regarded as the most efficient and popular approach for prediction and classification tasks. This tool exhibits diverse applicability across multiple disciplines. The proposed methodology utilises a flow diagram, resembling a tree structure, to depict the predictions derived from a series of splits depending on features. The process commences with the establishment of a primary node and culminates in the determination derived from the outcomes generated by the terminal nodes [5].

3.5 Random Forest

The Random Forest technique is employed at the regularisation step, where the model's quality is maximised and the trade-off between variance and bias issues is addressed. To address the decision tree challenge, the Random Forest algorithm generates many Decision Trees by utilising random samples. Observations are categorised by each individual tree, and the final result is made based on a majority vote. The unsupervised model utilises the Random Forest algorithm to assess the proximity of various data points.

3.6 Logistic Regression

The linear regression hyperplane generated is not suitable for predicting the variable that is dependent in linear regression using only the independent variable. Logistic regression is utilised in cases when the data is discontinuous [4]. This method is employed to forecast the outcome of a dataset including several distinct variables that determine a particular result [2]. Logistic regression, as opposed to predicting continuous variables, is utilised to predict the binary outcome of whether an event is true or false. The tool functions as a means of categorization [4].

4. Cancer Prediction Methods

This section discusses the various prediction method to classify cancer based dataset and cancer types.

4.1 Prediction method based on various dataset

This section describe the various types of cancer classification based different dataset. Table.1 represents cancer prediction based on various dataset.

Table 1. Cancer prediction based on various Dataset

References	Dataset	Description	Link
[2,6]	UCI machine learning repository	Consists of 569 patients ,212 have an outcome of malignancy and 357 are Benign	https://archive.ics.uci.edu/ml/index.php
[3,4,5,6]	Wisconsin Breast Cancer Dataset repository	Has 699 instances, 2 classes, 11 integer-valued properties, and 458 benign and 241 malignant instances (Benign: 458; Malignant: 241)	https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[7]	Kaagle	Consist of cancer cell data	https://www.kaggle.com/
[8]	Lung Image Database consortium	CT Scan Images	https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
[11]	Data World Site	1000 instances total 25 attributes total (2 predictive, 1 class characteristic) Details about the attribute: Attribute 25 is the class label.	https://data.world/
[13]	PIMA dataset	The dataset has a total of 768 samples, each containing 8 numerical-valued attributes. Among these samples, 500 cases were tested negative, whereas 268 occurrences were tested affirmative.	https://www.kaggle.com/
[12]	Cancerdata.org	The statistical analysis was conducted using a sample size of 509 patients.	https://cancerdata.org/
[14]	mRTarbase and TarBase	There are a total of 11 microRNAs (miRNAs) and 41 messenger RNAs (mRNAs) identified in this particular cancer entity.	https://www.ncbi.nlm.nih.gov/
[15]	Stanford Microarray Database	Five microarray datasets of cancer data.	http://smd.princeton.edu/
[16]	NCBI Website	89 healthy individuals and 92 CRC patients.	https://www.ncbi.nlm.nih.gov/
[17]	University Medical Center Manheim Database	The dataset comprises a collection of 5000 histology pictures that have been categorised into eight distinct classes.	https://www.umm.uni-heidelberg.de/medical-faculty-mannheim/home/
[18]	The Cancer Genome Atlas(TCGA)	The dataset consist of cancer cell metrics.	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[19]	Electronic Medical Record	The dataset comprises a collection of 5000 histology pictures that have been categorised into eight distinct classes.	https://digital.ahrq.gov/electronic-medical-record-systems
[9,10]	CT Image	CT Scan Images	https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

4.2 Comparison of different machine learning algorithms for each type of cancer

This section discusses briefly about various types of cancer based on different machine learning algorithms.

4.2.1 Breast Cancer

This section describes the various machine learning algorithms used for classifying the breast cancer based its accuracy and its limitations. Breast cancer is characterised by the uncontrolled proliferation of cells within the breast tissue. The classification of breast cancer is contingent upon the specific cellular transformation occurring within the breast tissue. Metastasis refers to the process by which cancer cells disseminate from the primary breast tumour to distant sites inside the body via blood arteries and lymphatic vessels.[18,21] Table.2 represents classification of breast cancer using different machine learning algorithms.

Table 2. Various machine learning techniques used for breast cancer classification

References	Year	Algorithm	Accuracy	Advantage	Disadvantage
[2]	2020	Support Vector Machine (SVM)	97.14%	Construction of a separating hyperplane facilitates ease in the process.	Dataset exhibits a non-linear distribution that cannot be effectively separated by a hyperplane.
		K-Nearest Neighbors (KNN)	95.28%	Widely used in predictive analysis	The given factor exerts a significant impact on the temporal spacing of data points.
		Random Forests (RF)	95.7%	Solves problem of computational intelligence	Low interpretability
		Logistic Regression (LR)	95.7%	Analyzes dataset in which independent variables decides the result	Very expensive and time-consuming
[3]	2021	Support vector machine (SVM)	96.48%	Outliers have less influence	Tricky and complex
		K-Nearest Neighbors (KNN)	95.83%	Simple to understand and implement	Computationally expensive
		Decision Tree (DT)	96.91%	Normalization is not required	Leads to overfitting of the data
		Naïve Bayes (NB)	95.12 %	Mainly used for classification tasks.	It needs large amount of data to achieve its best accuracy.
[4]	2016	Support Vector Machine (SVM)	97.13%	Better accuracy	Under performs if the data has more noise.
		Decision Tree (DT)	95.13%	Quick and appropriate	Small variations in the data results complex different tree
		Naive Bayes (NB)	95.99%	Easy to implement	Set of predictors which are completely independent
		k Nearest Neighbors (KNN)	95.27%	Robust to noisy data	It uses all the training data at the runtime
[5]	2017	K-Nearest Neighbors (KNN)	97.56%	Can be updated at a very little cost	Distance can be dominated by irrelevant attributes
		SVM with different	99.44%	Input data can be converted	Under performs if the data has more noise.

		kernels Logistic Regression (LR) Naive Bayes (NB)	99.4% 97.75%	into High dimensional data updated easily to reflect new data Training period is less	Difficult to capture complex relationships Zero-frequency problem.
[6]	2022	Logistic Regression (LR) Decision Tree (DT) Support Vector Machine (SVM) Random Forest (RF) Naive Bayes (NB) K- Nearest Neighbors (KNN)	97.9% 95.8% 97.4% 98.2% 92.1% 96.2%	Very efficient and linearly separable Classify non-linearly separable data Inbuilt functionality Normalizing of data is not required Used for classification tasks. Decision boundaries can be of arbitrary shapes	Very expensive and time-consuming. Take more time for training-time complexity Takes a long training time on large datasets. It has low interpretability Not great for imbalanced data Complexity is O(n) for each instance to be classified
[7]	2021	K- Nearest Neighbors (KNN) Support Vector Machine (SVM) Logistic regression (LR) Naïve Bayes (NB) Random Forest (RF)	96.2% 97.4% 97.9% 92.1% 98.2%	Used in predictive Analysis Convex optimization Inference about the importance of each feature Don’t require more training dataset It is flexible to both classification and regression problems	Costs will be high Difficult to interpret Moderate multi-collinearity between independent variables Due to low training data accuracy is not perfect. Requires much computational power

4.2.2 Lung Cancer

This section describes the various machine learning algorithms used for classifying the lung cancer based its accuracy and its limitation. The onset of this condition typically originates within the pulmonary system and has the potential to disseminate to other lymph nodes or just other bodily organs. Additionally, it is noteworthy that malignancies originating from other organs can also metastasize to the lungs. The two categories are classified as small cell as well as non-small cell. Different forms of lung cancer have distinct growth patterns and require varying treatment approaches [6, 11]. Table.3 represents classification lung cancer using different machine learning algorithms.

Table.3. Various machine learning techniques used for lung cancer

References	Year	Algorithm	Accuracy	Advantage	Disadvantage
[8]	2019	Support Vector Machine (SVM)	86.6%	Early detection of cancer plays a pivotal role in achieving comprehensive remission of the disease.	Highly expensive
		K- Nearest Neighbors (KNN)	89%	The ability to detect the presence of malignant cells at an early stage is enhanced.	Survival rate is poor
		Random Forest (RF)	89.9%	The study demonstrates the highest level of categorization and achieves commendable outcomes in terms of competitiveness.	Need for the development of a reliable and effective methodology to detect lung cancer at its advanced stages.
[9]	2019	Random Forest (RF)	66%	Classify various image data	Classification was not predicted accurately
		Support Vector Machine (SVM)	94.5%	Categorizes negative and positive specimen of cancerous lung picture	Classifiers have low accuracy
[10]	2020	Random Forest (RF)	79%	Handle many numbers of variable without deleting any variable	Low contrast of intensity values
		Support Vector Machine (SVM)	86%	Separates the dataset into two classes	Miss few hidden patterns
[11]	2018	Logistic Regression (LR)	66.7%	Analysis of epidemiologic datasets	Time consuming process
		Decision Tree (DT)	90%	Used to predict the result.	Overfitting
		Naïve Bayes (NB)	87.8%	used in the area of Data Mining and Machine Learning	Conditional independence assumption does not always hold
		Support Vector Machine (SVM)	99.2%	reduces the misclassification rate	Choosing an appropriate kernel function is difficult consuming process
[12]	2020	Support Vector Machine (SVM)	60.6%	Solves both classification and regression problems	Zero probability problem
		Naïve Bayes (NB)	56.7%	Simple to implement	Small change in the data causes instability
		Decision Tree (DT)	50.4%	Non-parametric method	High variance
		Random Forest (RF)	96.8%	Automates missing values present in the data	Fails to determine the significance of each variable
[13]	2017	K- Nearest Neighbors (KNN)	76.96%	New data can be added seamlessly	Does not work well with large dataset
		Decision Tree (DT)	90.43%	Handle both numerical and categorical variables	Complexity levels are greater

4.2.3 Colorectal Cancer

This section describes the various machine learning algorithms used for classifying the colorectal cancer based its accuracy and its limitation. Table.4 represents classification colorectal of cancer using different machine learning algorithms. Colorectal cancer, alternatively referred to as bowel cancer, colon cancer, as well as rectal cancer, is characterised by the malignant growth of cells originating from the colon or rectum. Common clinical manifestations of this condition encompass the presence of hematochezia, alterations in defecation patterns, unintentional weight reduction, and feelings of exhaustion [30, 34].

Abbreviations and Acronyms

Table 4. Various machine learning techniques used for colorectal cancer

References	Year	Algorithms	Accuracy	Advantages	Disadvantages
[14]	2015	CRCmiRTar Naïve Bayes classifier	96% 92%	It is more sensitive than other related tools Improves the structural framework	Not suitable for protecting against intentional alteration of data Data driven so insufficient data would create problem
[15]	2018	Support Vector Machine (SVM)	60%	Applied in bioinformatics because of its high exactness	Not 100% efficient due to lack deep information on CRC gene
[16]	2018	SVM model with kernel Logistic Regression (LR)	90.1(when k=5) 91.2%	Exactness is clearly higher than different models Extend to multiple classes	It is not very efficient Requires no multicollinearity between independent variables
[17]	2020	Naive Bayes (NB) Support Vector Machine (SVM) Random Forest (RF) K- Nearest Neighbors (KNN)	92.83% 88.9% 73.3% 70.5%	Handles productivity of characterizing The effectiveness of the approach is enhanced when the number of dimensions exceeds the number of samples. Works well with both categorical and continuous variable. Non-linear data is particularly well-suited for analysis, as it does not require any underlying assumptions.	Eliminates the picture which can't be determined Doesn't perform well when the target classes are overlapping Complexity is high as it creates lots of trees The prediction of N in KNN could be slow if its value is high
[18]	2019	Naïve Bayes Classifier	95.24%	Deals with information with high aspects.	Creates problem for imbalanced data
[19]	2021	Logistic Regression (LR)	83.5% 98.2%	Helps in analyzing millions of complex CRC data Automatically handle the missing	Not very effective when number of observations is less than number of features

		Random Forest (RF) K- Nearest Neighbors (KNN) Support Vector Machine (SVM)	95.28% 97.13%	values New data can be added anytime and it won't affect the model relatively memory efficient	Requires much more time to train Sensible to noisy and missing data Not suitable for large datasets

5. The Computation of Performance Metrics for Various Cancer Types

The performance of the fore mentioned algorithms is evaluated using measures including accuracy, sensitivity, specificity, and precision. The metrics are derived using a confusion matrix, which includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$Sensitivity = \frac{TP}{(TN + FP)} \quad (3)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (4)$$

5.1 Breast Cancer

This section discusses the performance measure of breast cancer based on various machine learning algorithms. The accuracy and other performance metrics are been shown in the graphical representation in figure 5 and 6 respectively.

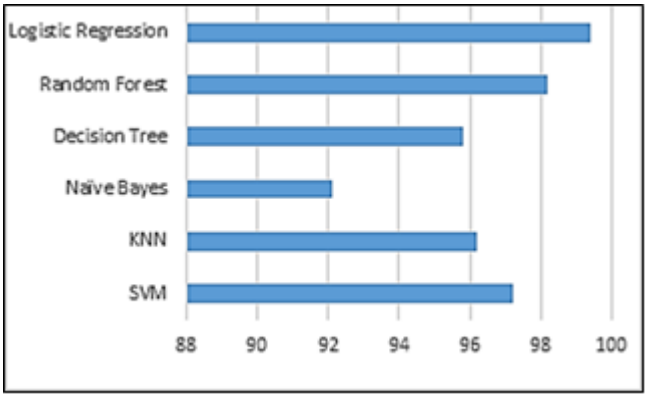

Fig 5. Accuracy for breast cancer

Fig 6. Performance for breast cancer

5.2 Lung Cancer

This section discusses the performance measure of lung cancer based on various machine learning algorithms. The accuracy and other performance metrics are given in the graphical representation in figure 7 and 8 respectively.

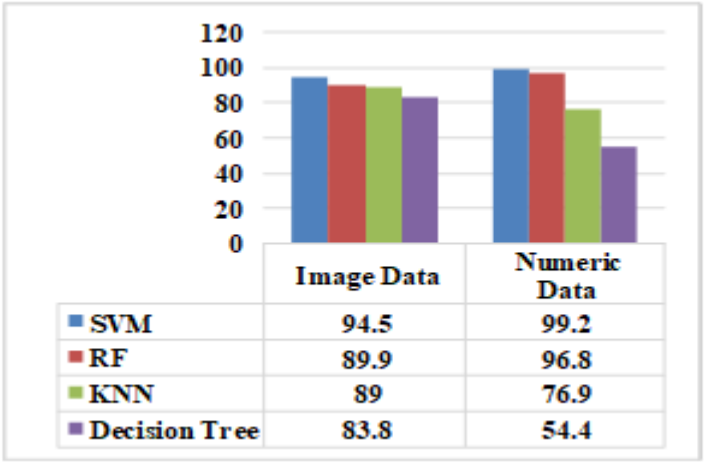

Fig 7. Accuracy for lung cancer

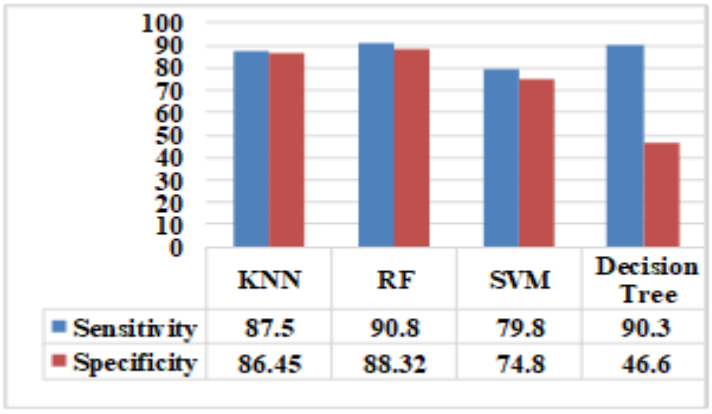

Fig 8. Performance of lung cancer

5.3 Colorectal Cancer

This section discusses the performance measure of colorectal cancer based on various machine learning algorithms. The accuracy and other performance metrics are given in the graphical representation in figure 9 and 10 respectively.

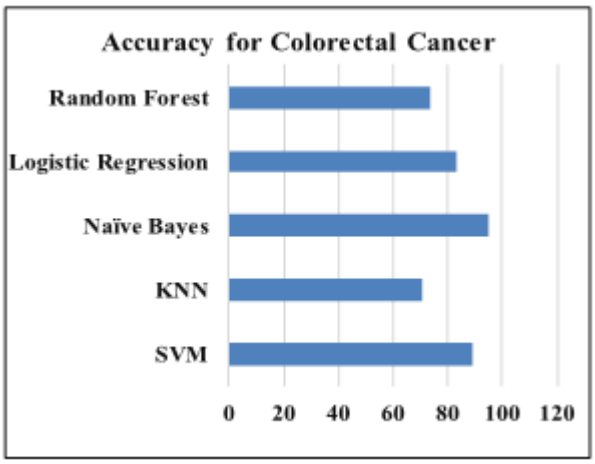

Fig 8. Accuracy for colorectal cancer

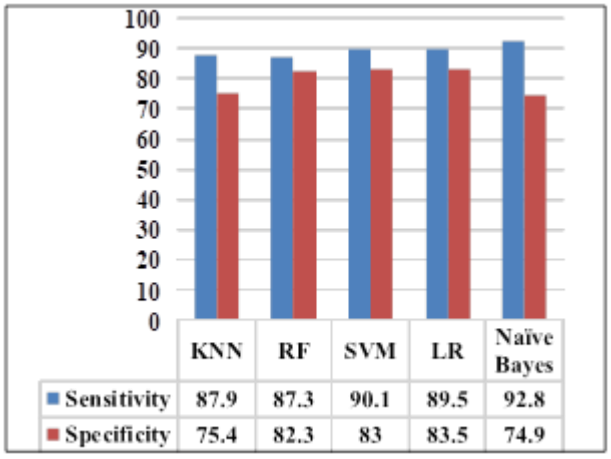

Fig 10. Performance for colorectal cancer

6. Conclusion

This research discussed a comprehensive comparison study on the most popular machine learning method for cancer prediction discussed this study. This study provides in a justification for the utilization of the prevailing machine learning models in the prediction of lung, breast, and colorectal cancer. After careful analysis, it has been shown that the logistic regression technique exhibits the highest level of reliability in predicting breast cancer. Subsequently, both the support vector machine as well as decision tree algorithms have been identified as the next most dependable methods for predicting lung cancer, while the Nave Bayes algorithm has been found to be the most suitable for predicting colorectal cancer. By utilizing this method, a cancer prediction method may be developed, which holds potential benefits for patients as well as doctors. Disease prediction and the implementation of suitable preventive measures can be achieved by the utilization of machine learning techniques such as K-Nearest neighbor, random forests, Support Vector Machines, logistic regression, as well as Naive Bayes.

Acknowledgements

The author's thanks their university/Institute, friends and colleagues to carry out this research work.

Author contributions

Srikanth Raman: Visualization, Investigation, Writing-Reviewing and Editing.

Tamil Priya Dhandapani: Visualization, Investigation, Writing-Reviewing and Editing.

Jagadeesan Srinivasan: Visualization, Investigation, Writing-Reviewing and Editing.

Savita P Patil: Visualization, Investigation, Writing-Reviewing and Editing.

Anupama K Ingale: Visualization, Investigation, Writing-Reviewing and Editing.

Manojkumar Vivekanandan: Visualization, Investigation, Writing-Reviewing and Editing.

Venkadeshan Ramalingam: Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] World Cancer Reasearch Fund International (WCRFI).

[2] Md. Milon Islam,Md. Rezwanul Haque, Hasib Iqbal Md. Munirul Hasan, Mahmudul Hasan, Muhammad Nomani Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques" SN Computer Science. 290, 2020. https://doi.org/10.1007/s42979-020-00305-w.

[3] [Apoorva V, Yogish H K, Chayadevi M L "Breast Cancer Prediction Using Machine Learning Techniques" ICIIC 2021.

[4] Sweta Bhise, Simran Bepari, Shrutika Gadekar, Deepmala Kale, Aishwarya Singh Gaur, Dr. Shailendra Aswale " Breast Cancer Detection using Machine Learning Techniques" (IJERT) 07, July-2021.

[5] [Mengjie Yu, B.S "Breast Cancer Prediction Using Machine Learning Algorithm" The University of Texas at Austin, May 2017.

[6] Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 1064 – 1069.

[7] Yousif A. Alhaj, Marwan M. Al-Falah, Abdullah M. Al-Arshy, Khadeja M. Al Nashad, Zain Alabedeen Ali Al Nomi, Badr A. Al Badawi and Mustafa S. Al

Khayat “An Efficient Machine Learning Algorithm for Breast Cancer Prediction” .

[8] D. Jayaraj, S. Sathiamoorthy “Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography Images ” IEEE Xplore Part Number: CFP19P17-ART; ISBN:978-1-7281-2119-2.

[9] Kyamelia Roy, Sheli Sinha Chaudhury, Madhurima Burman, Ahana Ganguly,, Chandrima Dutta, Sayani Banik, Rayna Banik,” A Comparative study of Lung Cancer detection using supervised neural network” IEEE Xplore.

[10] ikita Banerjee, Subhalaxmi Das, ” Prediction Lung Cancer– In Machine Learning Perspective”IEEE Xplore

[11] Radhika P R, Rakhi.A.S.Nair, Veena G,” A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms”,IEEE Xplore.

[12] Pranamita Nanda, Dr. N. Duraipandian,” Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest”, IEEE Xplore Part Number:CFP20F70-ART; ISBN:978-1-7281-4685-0.

[13] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan,” An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques”, (ECCE), February 16-18, 2017.

[14] Raheleh Amirkhah,a Ali Farazmand, Shailendra K. Gupta,bc Hamed Ahmadi,d Olaf Wolkenhauerbe and Ulf Schmi,” Naive Bayes classifier predicts functional microRNA target interactions in colorectal cancer”, Mol. BioSyst., 2015.

[15] Jiajun Zhi, Jiwei Sun, Zhongchuan Wang, Wenjun Ding ,“Support vector machine classifier for prediction of the metastasis of colorectal cancer”.Volume 41 Issue 3, March-2018.

[16] Dandan Zhao1,2 & Hong Liu1,2 & Yuanjie Zheng1,2 & Yanlin He1,2 & Dianjie Lu1,2 & Chen Lyu,” A reliable method for colorectal cancer prediction based on feature selection and support vector machine” Medical & Biological Engineering & Computing .57:901–912.

[17] Elene Firmeza Ohata1,João Victor Souza das Chagas, Gabriel Maia Bezerra,Mohammad Mehedi Hassan, Victor Hugo Costa de Albuquerque, Pedro Pedrosa Rebouças Filho,”A novel transfer learning approach for the classifcation of histological images of colorectal cancer”, https://doi.org/10.1007/s11227-020-03575-6.

[18] Nafizatus Salmi1 and Zuherman Rustam1*,” Naïve Bayes Classifier Models for Predicting the Colon Cancer”, .052068 IOP Publishing, 2019. doi:10.1088/1757-899X/546/5/052068

[19] Hui Li, MS1,#, Jianmei Lin, BS1,#, Yanhong Xiao, MS1 , Wenwen Zheng, MS1 , Lu Zhao, PhD1 , Xiangling Yang, PhD1,2, Minsheng Zhong, MS3 , and Huanliang Liu,” Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data” Volume 20: 1-9 © The Author(s) 2021.

[20] Kumar, B. S., Daniya, T., & Ajayan, J. Breast cancer prediction using machine learning algorithms. International Journal of Advanced Science and Technology, 29(3),2020

[21] Sharma, S., Aggarwal, A., & Choudhury, T. Breast cancer detection using machine learning algorithms. In 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 114-118). IEEE, 2018.December

[22] Alarabeyyat, A., & Alhanahnah, M. Breast cancer detection using k-nearest neighbor machine learning algorithm. In 2016 9th International Conference on Developments in eSystems Engineering (DeSE) (pp. 35-39). IEEE, 2016, August

[23] Harinishree, M. S., Aditya, C. R., & Sachin, D. N. Detection of Breast Cancer using Machine Learning Algorithms–A Survey. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1598-1601). IEEE, 2021, April.

[24] Junaid Malik, Serkan Kiranyaz, Suchitra Kunhoth, Turker Ince, Somaya Al-Maadeed, Ridha Hamila, Moncef Gabbouj,” Colorectal cancer diagnosis from histology images: A comparative study”, https://arxiv.org/pdf/1903.11210.

[25] OliverKenniona,StuartMaitlandb,Richard Bradyc,” Machine learning as a new horizon for colorectal cancer risk prediction? A systematic review” Volume 4, 100041, September 2022

[26] Harinishree, M. S., Aditya, C. R., & Sachin, D. N. Detection of Breast Cancer using Machine Learning Algorithms–A Survey. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1598-1601). IEEE, 2021, April.

[27] Shravya, C., Pravalika, K., & Subhani, S. Prediction of breast cancer using supervised machine learning techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6), 1106-1110, 2019.

[28] Bazazeh, D., & Shubair, R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th international conference on electronic devices, systems and applications (ICEDSA) (pp. 1-4). IEEE, 2016, December.

[29] Takamatsu, M., Yamamoto, N., Kawachi, H., Chino, A., Saito, S., Ueno, M., & Takeuchi, K. Prediction of early colorectal cancer metastasis by machine learning using digital slide images. Computer methods and programs in biomedicine, 178, 155-161, 2019.

[30] Nartowt, B. J., Hart, G. R., Muhammad, W., Liang, Y., Stark, G. F., & Deng, J. Robust machine learning for colorectal cancer risk prediction and stratification. Frontiers in big Data, 3, 6, 2020.

[31] Lu, W., Fu, D., Kong, X., Huang, Z., Hwang, M., Zhu, Y., & Ding, K. FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. Cancer Medicine, 9(4), 1419-1429, 2020.

[32] Gupta, P., Chiang, S. F., Sahoo, P. K., Mohapatra, S. K., You, J. F., Onthoni, D. D., ... & Tsai, W. S. Prediction of colon cancer stages and survival period with machine learning approach. Cancers, 11(12), 2007, 2019.

[33] Zheng, L., Eniola, E., & Wang, J. Machine Learning for Colorectal Cancer Risk Prediction. In 2021 International Conference on Cyber-Physical Social Intelligence (ICCSI) (pp. 1-6). IEEE, 2021, December.

[34] Xu, Y., Ju, L., Tong, J., Zhou, C. M., & Yang, J. J. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. Scientific reports, 10(1), 1-9, 2020.