Scope
Volume 14 Number 01 March 2024

Recent Advancements in Cancer Diagnosis Using Machine Learning Techniques: A Systematic Review of Decades of Research, Comparisons and Problems

Sulekha Das[1], Avijit Kumar Chaudhuri[2], Partha Ghosh[3], Prithwish Raymahapatra[4]

[1]Research Scholar, Information Technology , GCECT, Kolkata, West Bengal, India
ORCID ID: 0000-0002-6641-3268
[2]Associate Professor, Computer Science and Engineering, Brainware University, Barasat, West Bengal, India, ORCID ID: 0000-0002-5310-3180
[3]Assistant Professor Computer Science Engineering ,GCECT, Kolkata, West Bengal, India
[4]UG - Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata, India, ORCID ID: 0000-0001-7147-9491

Abstract : Cancer is a non-communicable disease that spreads throughout the body through uncontrolled cell growth. The malignant cell grows into a tumor, which weakens the immune system and disrupts other biological processes. The most frequent types of cancer are breast, lung, and cervical cancer. Several screening methods are available to detect the presence of cancer at various stages. Misdiagnosis can occur in some circumstances owing to human mistakes or incorrect data interpretation, resulting in the loss of human lives. To address these issues, this research study proposes an effective machine learning-based review and diagnosis technique backed by intelligence learning models. Artificial intelligence-based feature selection and classification techniques are used to detect cancer at an earlier stage, improve prediction accuracy, and save lives. In this research study, breast, cervical, and lung cancer datasets from the University of California, Irvine repository was used in these experimental investigations. To train and validate the optimal features minimized by the proposed system, the authors used supervised machine learning approaches. There could be numerous features that may contribute to the occurrence of cancer, it is difficult to pinpoint the specific environmental and other diagnostic features that contribute to it, but it still plays a role in determining cancer occurrence. We can achieve our goal of estimating the probability of cancer occurrences by using machine learning algorithms and frequent diagnostic data. Cancer data sets contain a variety of patient information features, but not all of them are useful in cancer prognosis. In such cases, a feature selection approach plays a crucial role in identifying the relevant feature set. In this research, we compare the effects of feature selection approaches on the accuracy provided by existing machine learning algorithms. We investigated the following machine learning methods for this purpose: Logistic Regression(LR), Naive Bayes(NB), Random Forest(RF), Hoeffding Tree(HT), and Multi-Layer Perceptron(MLP). Information Gain(IF), Gain Ratio(GR), Relief-F(R-F), and One-R(OR) were all evaluated as feature selection strategies.The training and performance models are validated using various accuracy matrices such as accuracy, sensitivity, specificity, f-measure, kappa score, and area under the ROC curve(AUC) using the 10-fold cross-validation approach. The accuracy of the proposed framework was 100%, 100%, and 91.30% on breast, cervical, and lung cancer datasets, respectively. Furthermore, this approach may serve as a versatile tool for extracting patterns from several clinical trials for various forms of cancer conditions.

Keywords : Lung Cancer, Cervical Cancer, Breast Cancer, Logistic Regression , Naïve Bayes , Random Forest

1. Introduction

Cancer is a major cause of death that is frequently caused by the accumulation of hereditary disorders and a variety of pathological alterations. Cancerous cells are abnormal growths that can develop in any part of the human body and are potentially fatal. Cancer, also known as malignancy, must be detected early and accurately to determine what treatments may be effective. Even though each modality has its own sets of problem, the most common causes of mortality are convoluted histories, poor diagnosis, and inappropriate treatments. The goal of this research is to examine, assess, categorize, and address current advances in human body cancer detection utilizing supervised machine learning approaches for breast, cervical and lung cancers.Cancer research has changed dramatically over the last few decades [1]. Scientists used a variety of methods, including early-stage screening, to detect cancer types before they developed symptoms. Furthermore, they have developed novel methods for predicting cancer to treatment early on. As a result of the introduction of new medical technologies, large amounts of cancer data have been collected and made available to the medical research community. However, one of the most exciting and difficult jobs for doctors is disease prediction and proper treatment. As a result, machine learning technologies are becoming increasingly popular among medical researchers. These methods may find and identify patterns and links in complicated datasets, as well as accurately forecast future outcomes of particular cancer.

Considering the importance of personalized medicine and the growing use of machine-learning approaches in cancer prediction and prognosis, we present a review of papers that use these methods. These studies address prognostic and predictive factors, which may be independent of a specific treatment or are incorporated to advise a therapy for cancer patients [2]. Furthermore, we discuss the various ML approaches used, the types of data they integrate, and the overall performance of each proposed scheme along with its benefits and drawbacks.

The suggested works clearly show a trend toward the integration of mixed data, such as clinical and genetic data. But we found a common problem in many existing research works: the absence of validation methods for existing ML techniques or testing of their models' predictive power. Applying ML techniques could improve the accuracy of predictions for cancer susceptibility, recurrence, and survival. According to [3] the use of ML approaches has increased the accuracy of cancer prediction outcomes by 15%–20% over the past few years.

In the present work, only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

2. Research Gap

A precise, error-free diagnosis is necessary for cancer disease detection. Any incorrect diagnosis will result in losses that cannot be recovered. Tumors are very common in cancer disorders, and the number of patients has increased yearly. As a result, the workload for medical professionals in this field has increased somewhat. It is urgent to present a tumor image segmentation method that is accurate and effective in order to meet the growing demand.

According to [4] Coccia (2020) research on deep learning technology can result in a paradigm shift in the diagnostic assessment of any cancer type and disease. This new technology can also benefit poor regions by allowing them to send digital images to labs in other developed regions for cancer type diagnosis, narrowing the current healthcare gap as much as possible.Bhinder (2021) [5] stated that AI has the ability to drastically impact nearly all aspects of oncology—from improving diagnosis to personalizing therapy and finding new anticancer drugs. They analyze the recent tremendous success in the application of AI to oncology, emphasize constraints and dangers, and outline a route for AI acceptance in the cancer clinic.

A complicated global health issue with a high fatality rate is cancer. The fast advancement of high-throughput sequencing technology and the use of various machine learning techniques that have appeared in recent years have made it possible to make cancer therapy decisions on gene expression, which has allowed for success in this field. In their study, [6] Xiao (2018) demonstrated the present interest in developing machine learning techniques that can effectively distinguish cancer patients from healthy individuals. They also discovered that none of the classification techniques used to date for cancer prognosis performed better than the others.

Cruz &Wishart (2006) [7] demonstrate that machine learning approaches may be used to significantly (15-25%) enhance the accuracy of predicting cancer susceptibility, recurrence, and death in better planned and validated research.In recent years, deep learning has exhibited exceptional accuracy when processing images for cancer diagnosis applications. The accuracy attained rivals that of radiologists and is acceptable for use as a clinical tool. Nevertheless, one big issue is that these models are black-box algorithms, which means they are inherently inexplicable. Because of the lack of confidence and accountability that characterizes black box algorithms; this provides a barrier to clinical deployment. Furthermore, contemporary rules prohibit the use of inexplicable models in therapeutic settings, demonstrating the need of explainability [8] Abreu(2016). Recurrence is a critical component of breast and cervical cancer behavior, and it is closely linked to mortality. Despite its importance, the majority of breast and cervical cancer databases rarely include it, making research into its prediction more difficult.

[8] Abreu (2016) ,utilized contemporary machine learning methodologies in their research, as these approaches are recognized for their effectiveness in providing unbiased insights for addressing the recurrent cervical cancer inferential challenge. In the past, a clinical diagnosis of recurrent cervical cancer was made using a doctor's clinical knowledge of various risk factors. Because of the vast categories of risk variables, years of clinical research and experience have attempted to pinpoint important risk factors for recurrence. In order to assess the efficacy of adjuvant therapy, clinical trials should randomly assign patients stratified by these prognostic characteristics. Furthermore, improved post-treatment surveillance may aid in detecting relapses earlier, and more accurate recurrent status assessment may improve outcomes.

Despite the use of various approaches, the issue of breast and cervical cancer recurrence prediction remains unresolved. The combination of various machine learning techniques and the establishment of standard predictors for breast cancer recurrence appear to be the key future avenues for improved outcomes [9] Ahmad (2013).

Despite the use of several approaches, predicting cancer recurrence remains a challenge. The integration of multiple machine learning approaches, as well as the establishment of standard predictors for various cancer recurrences; appear to be the primary future paths to achieve improved outcomes.

Phan (2023) [10] use deep learning techniques and variations in the density of the Hounsfield Units on computed tomography scans to develop an improved method for the automatic detection and classification of common liver lesions. In their research, they have not focused on detecting lesions from all parts of the human body, not just the liver. The authors did not focus on big data analysis when developing real-time processing systems.

Based on the previous studies, the authors in this article present the classification and detection of probability of cancer using feature selection-enabled machine learning techniques. First features are selected using IG, GR, RF, and OR feature selection methods. These feature selection methods are filters that select relevant features for the classification. It results in improving the accuracy of the classification models. Then, classification is performed using LR, NB, RF, HT, and MLP algorithms. Detecting the probability of cancer using machine learning techniques, especially when combined with feature selection, can be a powerful approach for improving the accuracy and interpretability of cancer diagnosis.

Researcher should remember that the features, models, and feature selection strategies used will differ based on the kind of cancer, the dataset, and the specific aims. It is critical to work closely with domain experts and medical specialists to ensure that the model adheres to medical standards and norms. Furthermore, while managing patient data, compliance with regulations regarding privacy and ethical issues is critical.

3. Relevant Literature

Breast cancer

Breast cancer is one of the leading causes of death in women around the world. Although breast cancer is now the leading cause of death in India, cervical cancer was previously the most common cancer among Indian women. It primarily affects women between the ages of 30 and 69, with younger age groups being more affected (in their thirties and forties). Breast cancer must be confirmed based on a number of factors, including biopsy results, family history, and a slew of others. If any of these factors change, the likelihood of developing breast cancer changes. A consistent diagnosis enables us to gather variations and their effects on a specific patient (patient history of health). They would provide insight into the patient's medical situation, allowing us to more accurately predict the risk factor for the instance (patient). To classify the incidence of breast cancer in a specific patient, we developed basic classifiers using the LR, NB, RF, HT, and MLP algorithms. Following the accuracy of the aforementioned classifiers, we use feature selection techniques such as RLF, IG, GR, and OR approaches to determine the accuracy of each of the basic classifiers on the smaller feature set provided by these feature selection procedures. We then perform a comparative analysis to determine which set of feature selection strategies and machine learning algorithms provides the best results. The datasets used in this study are the Breast Cancer Wisconsin (Diagnostic) Data Sets from the UCI Repository.

Table 1.Wisconsin (Diagnostic) Data Set for Breast Cancer Performance Comparison

Year	**Method**	**Results**
Sridevi&Murugan, 2014 [11]	Multilayer perceptron(MLP)	Accuray : 100 %
Alickovic&Subasi, 2017 [12]	Rotation Forest model classifies using GA	Accuracy : 99.48% AUC :0.993
Hamsagayathri&Sampath, 2017 [13]	Priority based decision tree classifier	Accuracy : 93.63% Sensitivity : 0.936 Specificity : 0.982 Auc :0.929
Zheng, et al., (2014) [14]	K-SVM	Accuracy : 97.38%
Sewak (2007) [15]	Ensemble SVM	Accuracy : 99.29%

		Sensitivity :1 Specificity :0.981
Obaid, (2018) [16]	Quadratic Kernel Based SVM	Accuracy : 98.1% Auc (benign) : 0.984305 Auc (malignant):0.988352
Kumari&Arumugam, 2015 [17]	Hybrid Krill Herd	Accuracy :87.89 % Sensitivity :0.975 Specificity :0.718
Chaudhuri, et.al (2021) [18]	DCA	Accuracy :97% Sensitivity :0.99 Specificity :0.96 Auc:1
This study		**Accuracy :100%** **Sensitivity :1** **Specificity :1** **Auc : 1**

Cervical Cancer

Cervical cancer takes place when malignant tumor cells grow in the cervix which is located in the lower part of the uterus of a female's reproductive system. Commonly, women over the age of 30 experience a higher risk of cervical cancer. The main cause of cervical cancer is the infection of certain types of human papillomavirus (HPV), specifically HPV16 and HPV 18. Although cervical cancer sounds prevalent, it can be easily prevented with HPV vaccinations and regular screening tests. Though HPV vaccinations seem to have promising effects, it is still safer to take regular screening tests as HPV vaccines are not recommended for people older than 26 . Screening tests include "cervical cytology (also called the Pap test or Pap smear) and, for some women, testing for human papillomavirus (HPV)". Cervical cancer is highly treatable if found early through screenings. However, on the patient side, screenings cost time (at least one office visit) and money (the cost for the test and the visit). Also, screening tests are inefficient considering the limited hospital resources and the large populations that need the screenings. Such a traditional method of screening cannot deal with large amounts of patients at once. Furthermore, the Pap test, "a test in which cells are taken from the cervix and vagina and examined under a microscope", can be highly dependent on the doctors' experience and be rather subjective. There are inaccuracies in human decisions after all.The datasets used in this study are the Breast Cancer Wisconsin (Diagnostic) Data Sets from the UCI Repository.

Table 2. Wisconsin (Diagnostic) Data Set for Cervical Cancer Performance Comparison

Reference	**Risk Factors Used**	**Machine Learning Technique**	**Results**
Ahishakiye& Emmanuel (2020) [19]	5	Ensemble of {KNN, CART, NB, SVM} with Voting Classifier	Accuracy (%) - 87.21
Choudhury, et al. (2018) [20]	6	DT	Accuracy (%) - 97.52 Sensitivity – 100, Specificity – 95.03, Precision – 95.27, F-measure – 97.58
Lu, Jiayi, et al. (2020) [21]	14	Ensemble of {LR, DT, SVM, MLP, KNN}	Accuracy (%) - 83.16 Recall – 28.35, Precision – 51.73, F1 score – 32.80
Nasution, et al. (2018, March) [22]	12	PCA + C4.5 DT	Accuracy (%) - 90.70 Particularity – 100, Precision - 100
Nithya&Ilango (2019) [23]	14	C5.0	Accuracy (%) - 100 AUC – 0.91
		RLF	Accuracy (%) - 100 AUC – 0.91
		RPART	Accuracy (%) - 97 AUC – 0.81
		SVM	Accuracy (%) - 93 AUC – 0.8
		KNN	Accuracy (%) - 89 AUC – 0.5
Priya&Karthikeyan (2020) [24]	10	DT	Accuracy (%) - 91.03
		Rotation Tree	Accuracy (%) - 88.52
		RF	Accuracy (%) - 92.63
		SVM – Linear	Accuracy (%) - 93.82

		Backpropagation	Accuracy (%) - 97.25
Sawhney, et al. (2018) [25]	4.6	Binary Firefly Algorithm (BFA) + RLF	Accuracy (%) - 97.36
Singh, H. D. (2018) [26]	7	SVM – Linear	Accuracy (%) - 65.47 Sensitivity – 55.56, Specificity – 66.15,
		RLF	Accuracy (%) - 70 Sensitivity – 44.4, Specificity – 71.53
		GBM	Accuracy (%) - 40.31 Sensitivity – 77.8, Specificity – 41.8,
Tripathi, et al. (2020) [27]	15	Chicken Swarm Optimization (CSO) + KNN	Accuracy (%) - 97.82
		CSO + RLF	Accuracy (%) - 99.53
Chaudhuri, et al. (2021) [18]	12	LR, NB, SVM, ET, RLF, GDB (Values are given for GDB method.)	Accuracy (%) - 96 Sensitivity – 96 Specificity – 97 f1-score – 96 Precision – 97 Kappa – 0.71 AUC – 89
	5	LR, NB, SVM, ET, RLF, GDB (Values are given for LR method.)	Accuracy (%) - 96 Sensitivity – 96 Specificity – 95 f1-score – 97 Precision – 97 Kappa – 0.74 AUC – 94
This study	**6**	**LR, NB, RLF, HT, MLP**	**Accuracy (%) - 100** **Sensitivity – 100** **Specificity – 100** **f1-score – 100** **Precision – 100Kappa – 1** **AUC – 100**

Lung Cancer

Lung cancer occurs when malignant tumor cells develop in the lungs, which are vital organs of the respiratory system. Typically, individuals aged 40 and above face a heightened risk of developing lung cancer. The primary cause of lung cancer is cigarette smoking, with exposure to other carcinogens like asbestos and radon also contributing to its occurrence. Although lung cancer is a prevalent concern, it can be significantly prevented by adopting smoking cessation strategies and through routine screening tests. While smoking cessation has shown promising results, it is essential to continue regular screening, particularly because this method is not applicable to all age groups. Screening tests typically involve imaging techniques such as low-dose computed tomography (CT) scans and chest X-rays .Early detection of lung cancer through screening can lead to highly effective treatment outcomes. However, from the patient's perspective, these screenings entail a commitment of time and financial resources (including the cost of the test and the medical visit). Additionally, screening programs face challenges related to limited healthcare resources and the need to accommodate a vast population of individuals who require screenings. Traditional screening methods are often insufficient to handle the influx of patients efficiently. Furthermore, the interpretation of screening results, particularly in radiological imaging, can be influenced by the experience and subjectivity of the medical professionals involved, introducing potential inaccuracies in decision-making . Human judgment, after all, may not always be entirely objective.The datasets used in this study are the Breast Cancer Wisconsin (Diagnostic) Data Sets from the UCI Repository.

Table 3. Wisconsin (Diagnostic) Data Set for Lung Cancer Performance Comparison

Author	**Method used**	**Results**
Singh, et al. (2019). [28]	Multilayer perceptron	Accuracy (%) : 88.55 F1 score : 0.8681 Precision : 0.8695 Recall : 0.8916
Faisal, et al. (2018, December). [29]	MLP+GBT+SVM	Accuracy (%) : 88.57 F1 score : 80.31 % Precision : 84.44 %

		Recall : 76.57%
Vieira, et al. (2021, March). [30]	ANN	Accuracy : 93% Sensitivity : 96% Specificity : 90% Precision : 91%
Xie, et al. (2021). [31]	Naïve Bayes	Accuracy : 100% Sensitivity : 100 % Specificity :100% AUC: 100 % Precision : 100%

4. Methodology

RelieF(RLF) algorithm

The Relief algorithm [32] ranks the features using a feature relevance criterion. In contrast to statistical measures that rate the quality of traits, the Relief technique considers context.

As a consequence, it can manage the properties effectively even when there is a considerable dependency between them [33]. However, the Relief algorithm is confined to two-class problems. As a result, the RLF algorithm was introduced. It is a modification of the Relief algorithm for dealing with multi-class problems with noisy and missing data. The RLF algorithm decides if a property is desirable by repeatedly selecting one instance. It is capable of handling both continuous and discrete data.

Information Gain(IG)

An entropy-based feature assessment method known as Information Gain (IG) is frequently used in machine learning. IG evaluates how much knowledge a feature imparts to the target class. In target class, IG can identify the features with the highest information. The features with a high IG have typically chosen to produce the best classification results because they are highly relevant to the target class. On the other hand, IG is unable to eliminate pointless features. As a result, we must keep getting rid of unused features. IG is derived from entropy, as shown in the equations. By calculating the probability of a particular occurrence or feature, entropy is used to quantify the uncertainty of a class. Entropy is inversely proportional to IG. The quantity of information obtained is frequently determined by two factors: the amount of information available before learning the attribute value and the amount of information available after learning the attribute value. The maximum value of IG for several classes is 1. The formula for utilizing entropy to investigate more than two classes given in these references [34,35,36].

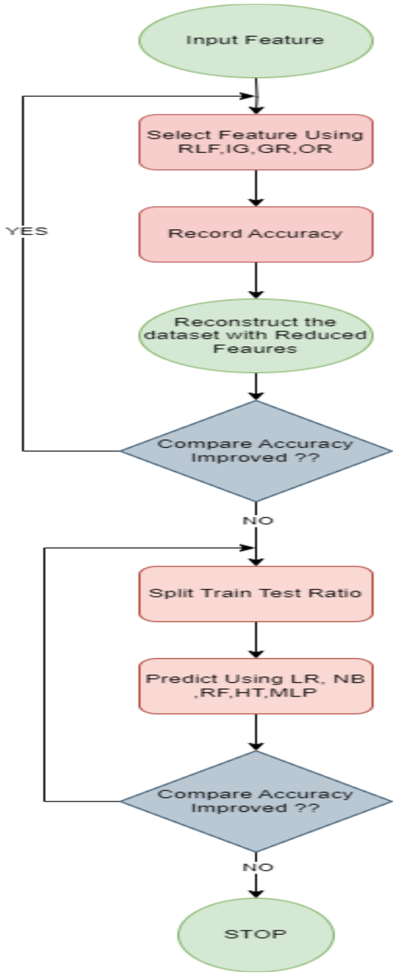
Gain ratio(GR)

It is a type of IG that reduces its bias toward high-branch qualities by selecting an attribute or feature and accounting for branch number and size. In order to repair unstable data, the inherent information of a split is taken into account. It is biased because it favors characteristics with high values. To remove this bias, the GR divides the predicted attribute's information gain by the observed attribute's entropy[37,38].

ONE-R (OR)

The OR attribute evaluates an attribute's worth. OR is a simple classifier introduced by Holte [5]. The characteristics with the lowest error rate are chosen as the single rule for this feature selection approach, which then ranks the other characteristics properly [39], [40]. It branches for each value of each characteristic as it constructs rules and tests them [41]. One-R is an efficient and straightforward machine-learning classification technique. A single-level decision tree is produced using OR. The OR technique produces one rule for each attribute in the training data. It selects the rule with the least amount of mistakes.

5. Flowchart


Input Feature
Select Feature Using RLF,IG,GR,OR
Record Accuracy
Reconstruct the dataset with Reduced Feaures
Compare Accuracy Improved ??
YES
NO
Split Train Test Ratio
Predict Using LR, NB ,RF,HT,MLP
Compare Accuracy Improved ??
NO
STOP

6. Results and Discussion

The analysis used in our study shows better accuracy for a subset of the complete feature set. This leads us to the conclusion that not all characteristics are required to accurately predict breast, lung, and cervical cancers, and that feature selection is useful when developing an effective model in these situations. The feature selection procedures IG, GR, RLF, and OR have demonstrated excellent performance in predicting various cancer risk factors. Our examination of the LR, NB, RF, HT, and MLP algorithms revealed that most of these algorithms were highly accurate in diagnosing cancers with selected features. RF and MLP classifiers have generally performed very well in detecting people exhibiting clinical signs of breast, lung, and cervical cancers, in addition to being extremely accurate through dependable outcomes with maximum accuracy. This study shows that improved feature selection approaches and repeated 10-fold cross-validation techniques can be used to develop classifier models with ML algorithms to improve prediction accuracy for various cancer detections. This study could be expanded to include forecasts for other types of cancers and additional diseases. Together, the imparted classifiers demonstrated improved performance accuracy with the optimal features' dataset.Table 3 provides a comprehensive overview of the various feature selection approaches employed to enhance the detection of lung cancer, cervical cancer, and breast cancer. Notably, it reveals distinct trends in accuracy, Kappa scores, AUC scores, and other performance metrics across different feature selection methods and the application of the random forest algorithm.In the case of lung cancer detection, the analysis demonstrates that using the Gain Ratio with a subset of nine carefully selected features leads to superior accuracy. In contrast, employing all available features with the random forest algorithm results in comparatively lower accuracy. This discrepancy underscores the significance of feature selection in optimizing the model's predictive capabilities for lung cancer.Turning our attention to cervical cancer detection, it becomes evident that other feature selection methods consistently yield 100% accuracy. However, when employing the random forest approach with all features, the accuracy drops. This highlights the importance of feature selection for achieving optimal results in cervical cancer diagnosis.Similarly, for breast cancer detection, the results reveal that accuracy is suboptimal when all features are utilized. Conversely, employing alternative feature selection techniques consistently yields 100% accuracy when coupled with the random forest algorithm.Digging deeper into the performance metrics, it is notable that Kappa scores consistently exceed 0.8 for lung cancer and reach a perfect score of 1 for breast cancer and cervical cancer when using feature selection methods. This underscores the robustness of these methods in improving model reliability.The AUC scores, which gauge the model's ability to distinguish between positive and negative cases, further substantiate the effectiveness of feature selection. In lung cancer detection, all feature selection methods except for the use of all features with random forest yield an AUC of 0.96, indicating strong discrimination ability. In breast cancer and cervical cancer detection, the AUC score attains a perfect 1, signifying an ideal classification algorithm.Additionally, sensitivity and specificity, crucial indicators of a model's ability to minimize false positives and false negatives, consistently surpass 0.9 in every feature selection method except when employing all features for lung cancer detection. In both breast cancer and cervical cancer detection, sensitivity and specificity reach the optimal value of 1, underscoring the reliability of the model.

Table 4. Features selected using various Feature Selection approaches for Various Cancer Disease

Disease	Approach	Selected features
Lung Cancer	All features(15+1)	Gender, Age, Smoking, Yellow_Fingers, Anxiety, Peer_Pressure, Chronic Disease, Fatigue , Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness Of Breath, Swallowing Difficulty, Chest Pain, Lung_Cancer
	IG with 9 features(8+1)	Gender, yellow_fingers, anxiety, peer_pressure, chronic disease, wheezing, shortness of breath, swallowing difficulty, lung_cancer
	GR with 9 features(8+1)	Gender, yellow_fingers, anxiety, peer_pressure, chronic disease, allergy , wheezing, shortness of breath, swallowing difficulty, lung_cancer
	O R With 9 features(8+1)	Gender, yellow_fingers, anxiety, peer_pressure, chronic disease, allergy , wheezing, shortness of breath, swallowing difficulty, lung_cancer
	RL F with 9 features(8+1)	Gender, anxiety, peer_pressure, chronic disease, allergy , shortness of breath, swallowing difficulty, chest pain, lung_cancer
Cervical Cancer	All features(35+1)	Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD, IUD (years), stds, stds (number), stds:condylomatosis, stds:cervicalcondylomatosis, stds:vaginalcondylomatosis, stds:vulvo-perinealcondylomatosis, stds:syphilis, stds:pelvic inflammatory disease, stds:genital herpes, stds:molluscumcontagiosum, stds:AIDS, stds:HIV, stds:Hepatitis B, stds:HPV, stds: Number of diagnosis, stds: Time since first diagnosis, stds: Time since last diagnosis, Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hinselmann, Schiller, Citology, Biopsy
	IG with 7	Stds:condylomatosis, stds:vulvo-perinealcondylomatosis, stds:genital herpes, stds:molluscumcontagiosum, stds:HPV,

	features(6+1)	Dx, Biopsy
	GR with 7 features(6+1)	Stds:condylomatosis, stds:vulvo-perinealcondylomatosis, stds:genital herpes, stds:molluscumcontagiosum, stds:HPV, Dx, Biopsy
	OR With 7 features(6+1)	Hormonal Contraceptives (years), IUD (years), stds, stds (number), stds:condylomatosis, Citology, Biopsy
	RLF with 7 features(6+1)	Smokes, Hormonal Contraceptives, stds:pelvic inflammatory disease, stds:molluscumcontagiosum, stds:HPV, Dx, Biopsy
Breast Cancer	All features(30+1)	Radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, Outcome
	IG with 7 features(6+1)	Texture_mean, compactness_mean, fractal_dimension_mean, compactness_se, concavity_se, radius_worst, Outcome
	GR with 7 features(6+1)	Texture_mean, compactness_mean, fractal_dimension_mean, compactness_se, concavity_se, radius_worst, Outcome
	OR With 7 features(6+1)	Symmetry_mean, fractal_dimension_mean, radius_se, texture_se, concavity_se, fractal_dimension_worst, Outcome
	RLF with 7 features(6+1)	Compactness_mean, fractal_dimension_mean, compactness_se, concavity_se, concave points_se, fractal_dimension_worst, Outcome

Table 5. Comparison of Accuracies with all features and selected features in 10 fold cross validation

Disease	Approach	LR	NB	RF	HT	MLP
Lung Cancer	All features(15+1)	73.14	68.9	81.88	68.94	74.43
	Information Gain with 9 features(8+1)	75.08	76.05	91.26	76.05	86.08
	Gain Ratio with 9 features(8+1)	75.08	76.10	91.30	76.10	86.08
	One R With 9 features(8+1)	75.08	76.05	91.26	76.06	86.10
	Relief F with 9 features(8+1)	75.1	76.1	91.26	76.05	86.1
Breast Cancer	All features(30+1)	83.3	80.5	89.6	80.5	87.5
	Information Gain with 7 features(6+1)	88.75	91.39	100	91	98
	Gain Ratio with 7 features(6+1)	88.75	91.39	100	91	98
	One R With 7 features(6+1)	95.6	95.43	100	94.9	97.54
	Relief F with 7 features(6+1)					

Cervical Cancer	All features(35+1)	85.31	85.9	95.8	71.9	96.97
	Information Gain with 7 features(6+1)	96.1	93.02	100	93.02	97.8
	Gain Ratio with 7 features(6+1)	96.1	93.01	100	93.02	97.8
	One R With 7 features(6+1)	96.1	93.01	100	93.02	97.8
	Relief F with 7 features(6+1)	95.22	90.2	100	94.9	100

Table 6. Comparison of Kappa Statistics with all features and selected features in 10 fold cross validation

Disease	Approach	LR	NB	RF	HT	MLP
Lung Cancer	All features(15+1)	0.46	0.38	0.64	0.38	0.49
	Information Gain with 9 features(8+1)	0.50	0.52	0.83	0.52	0.72
	Gain Ratio with 9 features(8+1)	0.50	0.52	0.83	0.52	0.72
	One R With 9 features(8+1)	0.50	0.52	0.83	0.52	0.72
	Relief F with 9 features(8+1)	0.50	0.52	0.83	0.52	0.72

Breast Cancer	All features(30+1)	0.62	0.52	0.75	0.52	0.72
	Information Gain with 7 features(6+1)	0.75	0.81	1	0.62	0.94
	Gain Ratio with 7 features(6+1)	0.75	0.81	1	0.62	0.94
	One R With 7 features(6+1)	0.83	0.81	1	0.79	0.91
	Relief F with 7 features(6+1)	0.83	0.81			
Cervical Cancer	All features(35+1)	0.60	0.60	0.88	0.1	0.92
	Information Gain with 7 features(6+1)	0.84	0.67	1	0.67	0.91
	Gain Ratio with 7 features(6+1)	0.84	0.67	1	0.67	0.91
	One R With 7 features(6+1)	0.84	0.67	1	0.67	0.91
	Relief F with 7 features(6+1)	0.90	0.80	1	0.9	1

Table 7. Comparison of AUC Score with all features and selected features in 10 fold cross validation

Disease	Approach	LR	NB	RF	HT	MLP
Lung Cancer	All features(15+1)	0.78	0.78	0.90	0.78	0.82
	Information Gain with 9 features(8+1)	0.84	.83	0.96	0.83	0.88
	Gain Ratio with 9 features(8+1)	0.84	0.83	0.96	0.83	0.88
	One R With 9 features(8+1)	0.84	0.83	0.96	0.83	0.88
	Relief F with 9 features(8+1)	0.84	0.83	0.96	0.83	0.88
Breast Cancer	All features(30+1)	0.89	0.88	0.97	0.88	0.93
	Information Gain with 7 features(6+1)	0.94	0.94	1	0.97	1
	Gain Ratio with 7 features(6+1)	0.94	0.94	1	0.97	1
	One R With 7 features(6+1)	0.96	0.96	1	0.97	0.91
	Relief F with 7 features(6+1)	0.96	0.96			
Cervical Cancer	All features(35+1)	0.85	0.85	0.96	0.60	0.97

	Information Gain with 7 features(6+1)	0.96	0.96	1	0.96	0.90
	Gain Ratio with 7 features(6+1)	0.96	0.96	1	0.96	0.98
	One R With 7 features(6+1)	0.96	0.96	1	0.96	.90
	Relief F with 7 features(6+1)	0.96	0.96	1	0.96	1

Table 8. Comparison of Sensitivity and Specificity with all features and selected features in 10 fold cross validation

Disease	**Approach**	**LR**		**NB**		**RF**		**HT**		**MLP**	
Lung Cancer		Sen	Spe	Sen	Spe	Sen	Spe	Sen	Spe	Sen	Spe
	All features(15+1)	0.72	0.75	0.68	0.70	0.82	0.82	0.65	0.70	0.75	0.74
	IG with 9 features(8+1)	0.75	0.76	0.75	0.77	0.93	0.90	0.75	0.77	0.88	0.86
	GR with 9 features(8+1)	0.75	0.76	0.75	0.77	0.93	0.90	0.75	0.77	0.88	0.84
	OR With 9 features(8+1)	0.75	0.76	0.75	0.77	0.93	0.90	0.75	0.77	0.88	0.84
	RLF with 9 features(8+1)	0.80	0.76	0.75	0.77	0.93	0.90	0.75	0.77	0.88	0.84
Breast Cancer	All features(30+1)	0.87	0.76	0.80	0.81	0.89	0.93	0.80	0.81	0.90	0.82

	IG with 7 features(6+1)	0.92	0.82	0.99	0.80	1	1	1	0.90	0.90	1
	GR with 7 features(6+1)	0.92	0.82	0.99	0.80	1	1	1	0.90	0.90	1
	OR With 7 features(6+1)	0.93	0.96	1	0.95	1	1	0.97	0.95	0.99	0.97
	RLF with 7 features(6+1)	0.93	0.96	1	0.95	1	1	0.97	0.95	0.99	0.97
Cervical Cancer	All features(35+1)	0.90	0.76	0.88	0.81	0.95	0.93	0.90	0.81	0.90	0.82
	IG with 7 features(6+1)	0.92	0.82	1	0.80	1	1	1	0.90	0.90	1
	GR with 7 features(6+1)	0.92	0.82	1	0.80	1	1	1	0.90	0.90	1
	OR With 7 features(6+1)	0.92	0.96	1	0.95	1	1	0.97	0.95	0.99	0.97
	RLF with 7 features(6+1)	0.99	0.96	1	0.95	1	1	0.97	0.95	0.99	0.97

Table 9. Comparison of F-Measure with all features and selected features in 10 fold cross validation

Disease	Approach	LR	NB	RF	HT	MLP
Lung Cancer	All features(15+1)	0.74	0.70	0.83	0.70	0.76
	Information Gain with 9	0.75	0.76	0.91	0.76	0.87

	features(8+1)					
	Gain Ratio with 9 features(8+1)	0.75	0.76	0.91	0.76	0.87
	One R With 9 features(8+1)	0.75	0.76	0.91	0.76	0.87
	Relief F with 9 features(8+1)	0.75	0.76	0.91	0.76	0.87
Breast Cancer	All features(30+1)	0.74	0.64	0.83	0.64	0.81
	Information Gain with 7 features(6+1)	0.83	0.88	1	0.95	0.99
	Gain Ratio with 7 features(6+1)	0.83	0.88	1	0.95	0.99
	One R With 7 features(6+1)	0.97	0.97	1	0.97	0.99
	Relief F with 7 features(6+1)	0.97	0.97	1	0.97	0.99
Cervical Cancer	All features(35+1)	0.70	0.69	0.91	0.83	0.94
	Information Gain with 7 features(6+1)	0.98	0.96	1	0.96	1
	Gain Ratio with 7 features(6+1)	0.98	0.96	1	0.96	1
	One R With 7 features(6+1)	0.98	0.96	1	0.96	1
	Relief F with 7 features(6+1)	0.93	0.88	1	0.93	1

7. Conclusion

In conclusion, this research study underscores the critical role of feature selection methods in the development of an effective machine learning-based cancer diagnosis technique. Cancer, as a non-communicable disease, poses a significant threat to human health, and early detection is paramount in saving lives. The study leveraged artificial intelligence-based feature selection and classification techniques to enhance the accuracy of cancer detection, with a focus on breast, cervical, and lung cancer. The research highlights that not all features in cancer datasets are equally relevant for accurate prognosis. Feature selection approaches, including IG, GR, RLF, and OR, were rigorously evaluated alongside classification algorithms such as LR, NB, RF, HT, and MLP. Through extensive experimentation and validation using various accuracy metrics, including sensitivity, specificity, f-measure, kappa score, and area under the ROC curve, the proposed framework demonstrated outstanding performance. Notably, the accuracy achieved by the proposed system reached 100% for breast and cervical cancer datasets and an impressive 91.30% for lung cancer. These results emphasize the potential of machine learning in cancer diagnosis and prognosis, particularly when combined with effective feature selection techniques. By eliminating irrelevant features and focusing on those with the most significant impact, this approach not only enhances prediction accuracy but also offers a versatile tool for extracting patterns from diverse clinical trials across various cancer conditions. In essence, this research contributes significantly to the field of cancer detection and emphasizes the pivotal role of feature selection methods in improving the effectiveness of machine learning models for early cancer diagnosis. It holds promise for the advancement of healthcare by providing a robust framework for accurate cancer detection and, ultimately, saving lives through early intervention.

Reference

- D. Hanahan, R.A. Weinberg Hallmarks of cancer: the next generation Cell, 144 (2011), pp. 646-674
- M.-Y.C. Polley, B. Freidlin, E.L. Korn, B.A. Conley, J.S. Abrams, L.M. McShane Statistical and practical considerations for clinical evaluation of predictive biomarkers J Natl Cancer Inst, 105 (2013), pp. 1677-1683.
- J.A. Cruz, D.S. Wishart Applications of machine learning in cancer prediction and prognosis Cancer Informat, 2 (2006), p. 59

- Coccia, M. (2020). Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. *Technology in Society*, *60*, 101198.
- Bhinder, B., Gilvary, C., Madhukar, N. S., &Elemento, O. (2021). Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, *11*(4), 900-915.
- Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, *153*, 1-9.
- Cruz, J. A., &Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030
- Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., & Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, *49*(3), 1-40.
- Ahmad, L., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., &Razavi, A. (2013). Using three machine learning techniques for predicting breast cancer prediction. *Journal of Health and medical informatics*, *4*(2), 1-3.
- Phan, A. C., Cao, H. P., Trieu, T. N., & Phan, T. C. (2023). Improving liver lesions classification on CT/MRI images based on Hounsfield Units attenuation and deep learning. *Gene Expression Patterns*, *47*, 119289.
- Sridevi, T., &Murugan, A. (2014). A novel feature selection method for effective breast cancer diagnosis and prognosis. International Journal of Computer Applications, 88(11).
- Aličković, E., &Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and applications, 28, 753-763.
- Hamsagayathri, P., &Sampath, P. (2017). Performance analysis of breast cancer classification using decision tree classifiers. Int J Curr Pharm Res, 9(2), 19-25.
- Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.
- Sewak, M., Vaidya, P., Chan, C. C., &Duan, Z. H. (2007, August). SVM approach to breast cancer classification. In Second international multi-symposiums on computer and computational sciences (IMSCCS 2007) (pp. 32-37). IEEE.
- Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., &Taha, F. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), 160-166.

- Kumari, S., &Arumugam, M. (2015). Application of bio-inspired krill herd algorithm for breast cancer classification and diagnosis. Indian J. Sci. Technol, 8, 30.
- Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. International Journal of Intelligent Systems and Applications (IJISA), 13(4), 24-37.
- E. Ahishakiye, R. Wario, W. Mwangi, and D. Taremwa, "Prediction of Cervical Cancer Basing on Risk Factors using Ensemble Learning," in 2020 IST-Africa Conference (IST-Africa), IEEE, May 2020, pp. 1-12.
- Y. M. S. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in Proceedings of the 2018 IISE Annual Conference, IISE, December 2018, pp.1456-1461
- J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," Future Generat. Comput. Syst., vol. 106, pp. 199-205, 2020.
- M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification," J. Phys. Conf., vol. 978, pp. 012058, 2018.
- B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," SN Applied Sciences, vol. 1, pp. 1-6, 2019.
- S. Priya and N. K. Karthikeyan, "A Heuristic and ANN based Classification Model for Early Screening of Cervical Cancer," Int. J. Comput. Intell. Syst., vol. 13, pp. 1092-1100, 2020.
- R. Sawhney, P. Mathur, and R. Shankar, "A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis," in International Conference on Computational Science and Its Applications, Cham: Springer, July 2018, pp. 438-449
- Singh, H. D. (2018). Diagnosis of Cervical Cancer using Hybrid Machine Learning Models (Doctoral dissertation, Dublin, National College of Ireland).
- A. K. Tripathi, P. Garg, A. Tripathy, N. Vats, D. Gupta, and A. Khanna, "Prediction of Cervical Cancer Using Chicken Swarm Optimization," in International Conference on Innovative Computing and Communications, Singapore: Springer, 2020, pp.591-604.
- Singh, G.A.P.; Gupta, P. Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Comput. Appl. 2019, 31, 6863–6877
- Faisal, M.I.; Bashir, S.; Khan, Z.S.; Khan, F.H. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In Proceedings

of the 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), Thrissur, Kerala, India, 18–20 January 2018; pp. 1–4.

- Vieira, E.; Ferreira, D.; Neto, C.; Abelha, A.; Machado, J. Data Mining Approach to Classify Cases of Lung Cancer. In World Conference on Information Systems and Technologies; Springer: Berlin/Heidelberg, Germany, 2021; pp. 511–521.
- Xie, Y.; Meng, W.Y.; Li, R.Z.; Wang, Y.W.; Qian, X.; Chan, C.; Yu, Z.F.; Fan, X.X.; Pan, H.D.; Xie, C.; et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl. Oncol. 2021, 14, 100907.
- Stiawan, D., Heryanto, A., Bardadi, A., Rini, D. P., Subroto, I. M. I., Idris, M. Y. B., ...&Budiarto, R. (2020). An approach for optimizing ensemble intrusion detection systems. *Ieee Access*, *9*, 6930-6947.
- Robnik-Sikonja M and Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 2003; 53: 23–69.
- Onan, A., &Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.
- Hall, M. A., & Smith, L. A. (1999, May). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference* (Vol. 1999, pp. 235-239).
- Win, T. Z., & Kham, N. S. M. (2019). *Information gain measured feature selection to reduce high dimensional data* (Doctoral dissertation, MERAL Portal).
- Karimi, Z., Kashani, M. M. R., &Harounabadi, A. (2013). Feature ranking in intrusion detection dataset using combination of filtering methods. *International Journal of Computer Applications*, *78*(4).
- Bhattacharya, S., &Selvakumar, S. (2016). Multi-measure multi-weight ranking approach for the identification of the network features for the detection of DoS and Probe attacks. *The Computer Journal*, *59*(6), 923-943.
- T. Garg and Y. Kumar, ``Combinational feature selection approach for network intrusion detection system,'' in Proc. 3rd Int. Conf. Parallel, Distrib Grid Comput., 2014, pp. 82_87,.
- R. A. Ghazy, E. S. M. El-Rabaie, M. I. Dessouky, N. A. El-Fishawy, and F. E. A. El-Samie, ``Feature selection ranking and subset-based techniques with different classi_ers for intrusion detection,'' Wireless Pers. Commun.vol. 111, no. 1, pp. 375_393, 2020.
- K. Shah and D. K. Singh, ``A survey on data mining approaches for dynamic analysis of malwares,'' in Proc. Int. Conf. Green Comput. Internet Things, 2015, pp. 495_499