The Scientific Temper (2024) Vol. 15 (3): 2931-2939
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.65
https://scientifictemper.com/

RESEARCH ARTICLE

Piecewise adaptive weighted smoothing-based multivariate rosenthal correlative target projection for lung and uterus cancer prediction with big data

V. Babydeepa, K. Sindhu*

Abstract

Cancer is the uncontrolled growth and spread of abnormal cells in the body. Early detection and prediction of cancer are crucial aspects of modern healthcare aimed at greatly improving the chances of survival for patients by reducing mortality rates and the number of people affected by this disease. Due to the large volume of data generated in the medical industry, accurate cancer detection is a challenging task. Many cancer classification systems using machine learning and deep learning models have been developed but accurate cancer detection with minimal time consumption remains a major challenging issue in the big data applications. To enhance the accuracy of cancer detection with minimal time, the Piecewise Adaptive Weighted Smoothing-based Multivariate Rosenthal Correlative Target Projection (PAWS-MRCTP) technique is introduced. This technique aims to detect lung and uterus cancers while leveraging big data. The proposed PAWS-MRCTP technique comprises three main processes namely data acquisition, preprocessing, and feature selection. In the data acquisition phase, a large number of cancer patient data are collected from lung cancer and uterus cancer detection datasets. Subsequently, the collected patient data undergo preprocessing. The preprocessing stage comprises three key processes namely handling missing data, noisy data, and outlier data. Firstly, the proposed PAWS-MRCTP is employed to address missing values, utilizing the Piecewise Adaptive Constant Interpolation method based on multiple available data points. Noisy data are identified using Gower's weighted smoothing technique, which detects data containing random variations or errors. Subsequently, outlier data are identified and removed by applying Peirce's statistical test. As a result, the pre-processed dataset is obtained resulting to minimize the time complexity. With the pre-processed dataset, the feature selection process is carried out to minimize the dimensionality of the large dataset. The proposed PAWS-MRCTP technique utilizes the Multivariate Rosenthal correlative target feature projection technique to identify the most relevant features. By selecting significant features, this approach enhances the accuracy of lung cancer and uterus cancer detection with minimal time consumption. Experimental assessment is conducted with different evaluation metrics such as cancer detection accuracy, precision, and cancer detection time and space complexity. The observed result shows the effectiveness of the proposed PAWS-MRCTP technique with higher accuracy with minimum time than the existing methods.

Keywords: Lung and uterus cancer detection, big data, preprocessing, Piecewise Adaptive Constant Interpolation method, Gower's weighted smoothing technique, Peirce's statistical test, feature selection, Multivariate Rosenthal correlative target feature projection technique.

PG & Research, Department of Computer Science, Government Arts College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanthonimalai, Karur, Tamilnadu, India.

***Corresponding Author:** K. Sindhu, PG & Research, Department of Computer Science, Government Arts College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanthonimalai, Karur, Tamilnadu, India., E-Mail: sindhupraneetha@gmail.com

How to cite this article: Babydeepa, V., Sindhu, K. (2024). Piecewise adaptive weighted smoothing-based multivariate rosenthal correlative target projection for lung and uterus cancer prediction with big data. The Scientific Temper, **15**(3):2931-2939.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.65

Source of support: Nil

Conflict of interest: None.

Introduction

Cancer is a disease characterized by the uncontrolled growth and spread of abnormal cells in the body. These abnormal cells form tumors and disrupting normal bodily functions. There are numerous different types of cancer, each with its own unique characteristics and treatment options. Early detection through screenings and medical intervention significantly improves outcomes for patients with cancer. Due to the large volume of data generated in the medical industry, accurate detection of cancer poses a significant challenge. The dimensionality of data is reduced with the aim of minimizing computational complexity in cancer detection. The Deep Neural Learning Cancer Prediction (DNLC) model was developed [1] to predict cancer in its

Received: 27/07/2024 **Accepted:** 18/08/2024 **Published** : 20/09/2024

early stages by selecting the best collection of features from datasets. However, the model failed to effectively focus on minimizing time and space complexity during cancer prediction. Explainable Machine Learning with Gradient Boosting Machine (XML-GBM), was introduced in [2], aimed to enhance the accuracy of lung cancer detection by leveraging relevant clinical features. However, it did not extend its capabilities to detect various types of cancers when confronted with big data. Feature transformation techniques and a regression model were developed in [3] for the early diagnosis of lung cancer using a dataset. However, deep learning algorithms were not employed to recognize cancer cells and their spread to other organs. Nonlinear Support Vector Machines (SVMs) with Radial Basis Function (RBF) kernels were developed [4] for multiclass classification in lung cancer detection. However, it failed to identify significant features from a large dataset. A personalized prognostic prediction tool was utilized in [5] for the detection of high-grade neuroendocrine cervical cancer. However, it exhibits a higher time complexity in cervical cancer detection. Three deep learning models, namely Bi-LSTM_simple, Bi-LSTM_dropout, and Pre-trained_BERT, were developed in [6] for lung cancer prediction. However, these models were not applied to complex datasets. The «Modified-DeepSurv» prognostic model was developed in [7], aimed to address the challenge of efficiently predicting lung carcinoma outcomes using a combination of deep learning and Cox proportional hazard regression. However, effectively predicting outcomes for lung cancer patients poses a significant challenge. Machine learning (ML) methods were developed in reference [8] to identify high-risk individuals for lung cancer detection and to mitigate long-term complications. However, accurate lung cancer detection remains a challenging issue. The TSVR and Dependent Nearest Neighbor algorithm were developed in [9] for predicting various types of cancer. However, it did not improving the accuracy of the prediction. Machine learning techniques and their performance assessment were developed in [10] for the classification of cervical cancer. A Boruta analysis and SVM method were developed in [11] for efficient feature selection and prediction modeling in cervical cancer. However, this approach did not address the diagnostic value of fully automated feature extraction in cervical cancer prediction. The machine learning model was designed in [12] to automate the recognition of persistent cervical cancer disease at an early stage. However, it was less effective in diagnosing cervical cancer based on specific evaluation criteria. A novel ensemble approach was developed in [13] to enhance the accuracy of predicting cervical cancer risk. A three ensemble-based classification techniques and Random Forest (RF) were developed in [14] for early identification of cervical cancer, utilizing the firefly algorithm. However, these methods faced challenges in effectively handling outlier data in cervical cancer detection. Ant Colony Optimization-based CNN methods were introduced in [15] with the aim of accurately recognizing cervical cancer Haitham Elwahsh (2023), Sarreha Tasmin Rikta (2023), Zunaira Munawar (2022), R. Sujitha (2021), Linlin Chen (2023), S. Mithun (2023), Jie Lei (2023), Elias Dritsas (2022), Ai-Min Yang (2021), Michał Kruczkowski (2022), Umesh Kumar Lilhore (2022), Sohely Jahan (2021), Jiayi Lu (2020), Irfan Ullah Khan (2021), R. Kavitha (2023).

- The PAWS-MRCTP technique has been developed to enhance the accuracy of cancer detection through the incorporation of preprocessing and feature selection methodologies.
- To minimize cancer detection time, the Piecewise Adaptive Constant Interpolation method is employed for missing data handling with multiple available data points. Additionally, Gower's weighted smoothing technique is employed to identify and address noisy data. Subsequently, outlier data points are detected and removed using Peirce's statistical test.
- To enhance the accuracy and precision of ulcer detection, we employ a Multivariate Rosenthal correlation-based target feature projection technique for selecting significant features from the dataset. These identified features are utilized for precise cancer detection with minimized space complexity.
- To evaluate the performance of our PAWS-MRCTP technique, complete experimentation is conducted with various evaluation metrics.

Related Works

Transfer learning and a modified generative adversarial network algorithm were introduced for lung cancer detection. However, the complexity analysis of lung cancer detection was not minimized. A Machine learning-based approaches were introduced for lung cancer detection, analyzing large and complex datasets. However, it failed to address lung cancer therapies with minimal space complexity. The Ensemble Support Vector Machine with Interpolation and Fuzzy Weight-based Recurrent Neural Networks was developed to identify the disease-free survival duration of cervical cancer. However, it failed with larger sample size for cervical cancer detection. The Feed-Forward Neural Network (FFNN) was introduced to maximize the accuracy of cancer prognosis prediction by employing an optimal set of feature selection. However, it faced difficulties when applied to various types of datasets. The Federated Deep Extreme Machine Learning method was developed for the prediction of lung diseases. However, it failed to achieve better classification results for various cancer diseases. A Generative-Discriminative framework was introduced for lung cancer detection. But it failed to perform accurate lung cancer prediction without

considering medical perspectives. A unique weight-based feature selection (WBFS) algorithm was designed to classify lung cancer subtypes from a database. However, it failed to automatically determine the optimal number of selected features from the high-dimensional dataset. An artificial intelligence system was developed with the aim of detecting lung cancer. A multi-stage framework was introduced to construct an AI-based decision support tool for predicting lung cancer patients. However, accurately predicting lung cancer patients with minimal time consumption remained a challenging issue. A convolutional neural network (CNN) was developed to quickly and accurately predict lung cancer patient data. However, it failed to diagnose and make decisions in intelligent medical systems for detecting other types of cancers. The artificial neural network (ANN) model was introduced with the aim of improving lung cancer prediction accuracy. However, it failed to achieve a high level of prediction accuracy. A Machine Learning Algorithms were designed for predicting cervical cancer. But it failed to analyze various deep learning algorithms to minimize the computational complexity. A new ensemble learning approach with SVM was developed for achieving cervical cancer prediction. But it failed to improve the performance cervical cancer prediction on higher-dimensional datasets. Ensemble classification method was developed based on majority voting for an accurate diagnosis of cervical cancer detection. Three machine learning models integrated with a stacked ensemble voting classifier were developed. However, the curse of dimensionality was not minimized Kwok Tai Chui (2023), Yawei Li (2022), Geeitha Senthilkumar (2021), Nadia G. Elseddeq (2021), Sagheer Abbas (2023), Jinpeng Li (2021), Yangyang Wang (2023), Hwa-Yen Chiu (2022), Marina Johnson (2022), Xiangbing Zhan (2021), Jason C. Hsu (2022), Naif Al Mudawi (2022), Raafat M. Munshi (2023), Qazi Mudassar Ilyas (2021), Turki Aljrees (2024).

Proposal Methodology

Cancer is a complex group of diseases distinguished by uncontrolled growth and spread of abnormal cells. These abnormal cells form tumors in nearby tissues and organs, disturbing normal physiological functions and visible in various forms. Cancer detection in medical industry aimed at identifying the presence of cancerous cells or tumors in the body at the earliest possible stage. It plays a pivotal role in improving treatment outcomes, reducing mortality rates, and enhancing the quality of life for individuals diagnosed with cancer. In this paper, a novel PAWS-MRCTP technique is developed for cancer detection at an earlier stage.

Figure 1 above illustrates the architecture diagram of the proposed PAWS-MRCTP technique for accurate cancer detection with big data. The accurate detection method involves three fundamental steps namely data acquisition, preprocessing, and feature selection. These fundamental processes of the proposed PAWS-MRCTP technique are explained briefly in the following subsections.

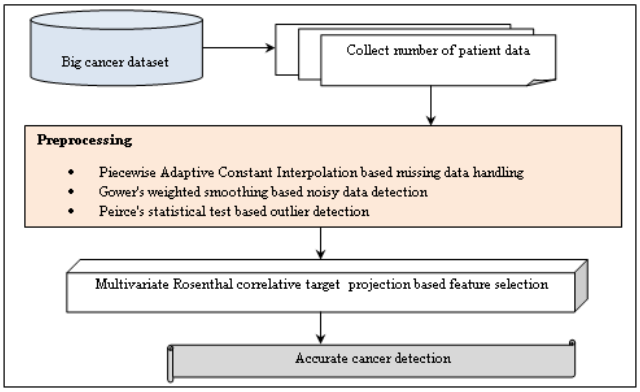

Figure 1: Architecture of proposed PAWS-MRCTP technique

Data acquisition

Data acquisition refers to the process of collecting, the patient data from the dataset. The cancer prediction dataset is taken from kaggle. This dataset contains information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss ,shortness of breath ,wheezing ,swallowing difficulty ,clubbing of finger nails and snoring.

The cervical cancer detection dataset is taken from kaggle. It is the leading causes of cancer-related deaths among women worldwide. Early detection and accurate prediction of cervical cancer significantly improve the chances of successful treatment and save lives. This dataset includes 36 features and 858 data instances.

Data preprocessing

Data preprocessing is a fundamental step in data analysis involves cleaning, transforming, and organizing raw data into a suitable format. The main aim of data preprocessing is to improve the quality of the data and enhance the performance of machine learning algorithms. The PAWS-MRCTP technique includes three major processing steps namely handling missing data, noisy data, and outlier data.

Piecewise Adaptive Constant Interpolation based missing data handling

The first step of data preprocessing is to perform the missing data handling. Missing data refers to the absence of values in a dataset for certain variables or features during data collection. Dealing with missing data is an important step in data analysis, as it affects the statistical analyses of cancer detection. Therefore, missing data handling refers to the techniques an employed to address the absence of information in a dataset. Piecewise Adaptive Constant Interpolation method is employed for handling the missing

data in the dataset. It is a mathematical technique used to find new data points based on the range of a discrete set of known data points within the dataset.

Let us consider the cancer dataset 'CDS' and the data are arranged in the form of matrix. Therefore, the input matrix is formulated in the form of matrix,

$$M = \begin{bmatrix} A_1 & A_2 & \dots & A_m \\ Dp_{11} & Dp_{12} & \dots & Dp_{1n} \\ Dp_{21} & Dp_{22} & \dots & Dp_{2n} \\ \vdots & \vdots & \dots & \vdots \\ Dp_{m1} & Dp_{m2} & \dots & Dp_{mn} \end{bmatrix} \quad (1)$$

Where, M indicates an input data matrix, each column indicates a number of features$A_1, A_2, A_3, \dots A_m$, each row indicates a number of data samples or instances or records $S_1, S_2, S_3, \dots S_n$which includes a number of data points $Dp_1, Dp_2, Dp_3, \dots Dp_n$ respectively.

The Piecewise Adaptive Constant Interpolation method, also known as nearest neighbour interpolation, is a technique used to estimate missing values in a dataset by considering the values of neighbouring data points. Some other interpolation methods that utilize information from all data points in the dataset, the proposed interpolation method focuses only on the nearest neighbours to the point.

Let us consider the data points $Dp_1, Dp_2, Dp_3, \dots Dp_n$ of the particular features. First, the nearest value of unknown data points is selected as the missing value. Then the Manhattan distance is measured for between themissing value and the nearest value.

$$d = \sum_{i=1}^{k} |Dp_m - Dp_i| \quad (2)$$

$$z = \min d \quad (3)$$

Where, z indicates an output of interpolation method, $\min d$ denotes a minimum distance between the missing value 'Dp_m' and the nearest value 'Dp_i'. After that, the minimal distance between the data point is replaced with missing value. In this way, the proposed an interpolation method effectively handles all missing values in the given dataset.

Gower's weighted smoothing based noisy data detection

Noisy data refers to random variations or errors that affect multiple data points throughout the dataset. These noisy data impact the accuracy of cancer detection. Therefore, the proposed technique utilizes the Gower's weighted smoothing technique to detect the noise data points within the dataset.

The proposedsmoothing technique assigns weights to the data points based on the distance of each observation. The function assigns higher weights to nearby data pointsand lower weights to distant data points. Based on the weight distribution, the noisy data points are smoothed.

The Gower's weighted smoothing is obtained as follows,

$$S_G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} * \beta_i}{\sum_{i=1}^{n} \beta_i} \quad (4)$$

$$\delta_{ij} = |Dp_i - Dp_j| \quad (5)$$

$$Q = \begin{cases} \delta_{ij} < T, \ H(\beta_i) \\ \delta_{ij} > T, \ L(\beta_i) \end{cases} \quad (6)$$

$$\beta_i = \exp\left(-\frac{(Dp_i - Dp_j)^2}{2\sigma^2}\right) \quad (7)$$

Where, S_G indicates an output of smoothing, δ_{ij} denotes a absolute difference between the two data points Dp_i and Dp_j respectively, β_i denotes a Gaussian weight assigned to the data points, T denotes a threshold, $H(\beta_i)$ and $L(\beta_i)$ represents the high or low weights respectively. It means that if the absolute difference between data points is less than the threshold, a higher weight is assigned. On the other hand, if the absolute difference between data points is greater than the threshold, a lower weight is assigned. Data points with absolute differences less than the threshold are considered normal, while those exceeding the threshold are considered as noisy data points. Therefore, the weighted average functions '$S_G S_G$', is used forsmoothing thenoisy data points within the dataset.

Peirce's statistical test based outlier detection

Outlier data refers to the data points that deviate significantly from the majority of the data points in a dataset. These data points are typically extreme values from the tendency of the other data distribution. Therefore, the proposed technique utilizes the Peirce's statistical test to detect the outlier data points within the dataset.

The formula for calculating the Peirce's statistical testis give below,

$$DF = |Dp_i - avg_{dp}| \quad (8)$$

$$avg_{dp} = \frac{\sum_{i=1}^{n} Dp_i}{n} \quad (9)$$

Where, Dp_i denotes a data points, avg_{dp} denotes average of the data points, DF indicates adeviation function. The following criterion is used to detect the outlier in the distribution of the data samples.

$$y = \begin{cases} \max DF, \ OD \\ otherwise, \ ND \end{cases} \quad (10)$$

Where, y denotes an outcome, $\max DF$ indicates a maximum deviation. If the deviation for a data point is maximum, then it is considered outlier data 'OD'. Otherwise, the data point is said to be anormal 'ND'. Based on this analysis, outlierdata points are removed for further processing.

The algorithm of different preprocessing steps are given below:

Algorithm 1 describes a step-by-step process of data pre-processing to minimize both time and space consumption in cancer detection. Initially, a number of patient data and features are collected from the cancer dataset. Next, missing values are identified by applying a distance measure. Based on the results, the missing values are handled. Following an interpolation method, Gower's weighted smoothing is applied to identify noisy data and smooth the dataset.

// **Algorithm 1:** Data pre-processing

Input: Dataset 'DS', features$A_1, A_2, A_3, \ldots A_m$, data samples or instances $Dp_1, Dp_2, Dp_3, \ldots Dp_n$

Output: Pre-processed dataset

Begin
Step 1: Collect number of features and data from dataset
Step 2: Formulate input vector matrix 'M'using (1)
Step 3: If missing value in dataset then
Step 4: Select the nearest value as missing value
Step 5: Compute Manhattan distance using (2)
Step 6: Find minimal distance using (3)
Step 7: Replace missing value with minimal distance value
Step 8:End if
Step 9: For each data point Dp_i and Dp_j
Step 10: Measure the absolute difference using (5)
Step 11: if ($\delta_{ij} < T$) then
Step 12: Data points are called as normal having higher weight
Step 13: else
Step 14: Data points are called as noisy having higher weight
Step 15: end if
Step 16: Obtain the Smoothing results using (4)
Step 17: end for
Step 18: Measures Peirce's statistical test using (8)
Step 19: if ($\max DF$) then
Step 20: Outlier data points
Step 21: else
Step 22: Normal data points
Step 23: end if
Step 24: Return (pre-processed dataset)
End

Finally, Peirce's statistical test is employed to detect outlier data points. This step aids in identifying data points that deviate considerably from the other maximum data points within the dataset. This pre-processing step of the PAWS-MRCTP technique minimizes the time and space complexity of cancer prediction.

Multivariate Rosenthal correlative target feature projection

Feature selection is a fundamental step to improve the performance of cancer detection. The proposed PAWS-MRCTP technique utilizes the Multivariate Rosenthal correlative target projection to projects the significant features from the dataset. Big datasets includes a more number of features causes increased computational complexity and challenges in achieving accurate cancer detection.

Figure 2 illustrates flow process of the Multivariate Rosenthal correlative target feature projection for accurate cancer detection. Let us consider the number of features $A_1, A_2, A_3, \ldots A_m$ in the given dataset. The Multivariate Rosenthal correlative target projection method refers to the degree of relationship between the feature matrix and target matrix in a dataset. Multivariate indicates the involvement of a number of features in the correlation measure.

$$H = RC\left(A_i, T_j\right) \tag{11}$$

Algorithm 2: Multivariate Rosenthal correlative target feature projection

Input: pre-processed Datasets 'DS', features$A_1, A_2, A_3, \ldots A_m$, data samples or instances $Dp_1, Dp_2, Dp_3, \ldots Dp_n$

Output: Select relevant features for cancer detection

Begin
Step 1: Collect the pre-processed dataset as input
Step 2:For each feature'A_i'
Step 3:Fortarget feature'T_i
Step 4:Measure the multivariate Rosenthal correlation using (9)
Step 5:***if*** $(RC\left(A_i, T_j\right) > TH)$then
Step 6: Features are said to be relevant
Step 7:else
Step 8: Features are said to be irrelevant
Step 9:End if
Step 10:Select the relevant features and remove irrelevant features
Step 11: End for
Step 12: End for
End

Where, H denotes an output of feature selection, $RC\left(A_i, T_j\right)$ indicates a Multivariate Rosenthal correlation between the features and target matrix. It is measured as follows,

Rosenthal correlation is a statistical method used to measure the correlation between features based on Standardized test statistics. It is formulated as follows,

$$RC\left(A_i, T_j\right) = \left[\frac{ST}{\sqrt{n}}\right] \tag{12}$$

Where,

$$ST = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(A_i - T_j\right)}{\sigma} \tag{13}$$

Where, $RC\left(A_i, T_j\right)$ indicates a Rosenthalcorrelation between the feature 'A_i' and target 'T_j', ST denotes Standardized test statistics, 'n' indicatesthe number of features, σindicates a deviation. The correlation $RC\left(A_i, T_j\right)$ provides the output results from '0 to '+1'. The threshold is set to the correlation for selecting the significant features from the dataset for accurate cancer detection.

$$X = \begin{cases} RC\left(A_i, T_j\right) > TH; & Selected \\ otherwise; & Removed \end{cases} \tag{14}$$

Where X denotes an output of the feature, TH denotes a threshold, $RC\left(A_i, T_j\right)$ indicates a correlation output. As a result, the features with high correlation than the threshold are selected for the accurate cancer detection. The features

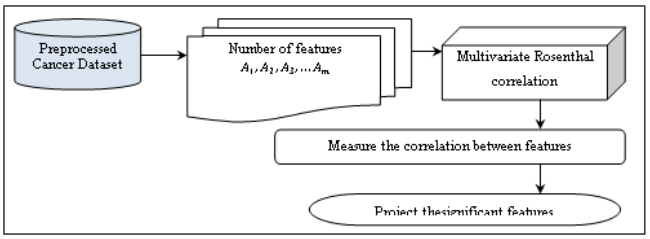

Figure 2: Flow process of Multivariate Rosenthal correlative target feature projection

with less correlation than the threshold are removed. The algorithm for Multivariate Rosenthal correlative target feature projection is given below,

Algorithm 2 described above outlines the process of selecting relevant features using Multivariate Rosenthal correlative target feature projection to enhance accurate cancer detection. Initially, the pre-processed dataset serves as input to identify relevant features. Next, Multivariate Rosenthal correlation is computed between the features and the target. This correlation measure distinguishes between relevant and irrelevant features by setting a threshold. Finally, the relevant features are selected from the dataset to improve the accuracy of cancer detection.

Experimental Scenario

In this section, the experimental evaluation of the proposed PAWS-MRCTP technique and existing DNLC [1] and XML-GBM [2] are implemented using Python coding. To conduct the experiment, two cancer dataset namely lung and cervical dataset are used. The lung cancer prediction dataset is taken from the https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link. This dataset includes 26 features and 1000 instances. This dataset contains information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss ,shortness of breath ,wheezing ,swallowing difficulty ,clubbing of finger nails and snoring.

The other dataset is a Cervical Cancer Risk Classification taken from https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classificationfor uterus cancer prediction. It is the leading causes of cancer-related deaths among women worldwide. This file contains a List of Risk Factors for Cervical Cancer leading to a Biopsy Examination. Cervical cancer is uncommon in women under the age of 20. However, numerous young women contract multiple types of human papillomavirus (HPV), heightening their likelihood of developing cervical cancer later in life. Young women exhibiting early abnormal changes who fail to undergo regular examinations face a heightened risk of localized cancer by the time they reach 40, and invasive cancer by the age of 50. This dataset includes 36 features and 858 instances.

Performance Comparision Analysis

In this section, performance of the proposed PAWS-MRCTP technique and existing DNLC and XML-GBM are assessed with various metrics, including cancer detection accuracy, precision, cancer detection time and space complexity with different number of data samples.

Cancer Detection Accuracy

It is referred to the ratio of number of patient data samples are classified as normal or cancerous defrom the total number of data samples. Therefore, cancer detection accuracy is mathematically computed as follows,

$$\boldsymbol{CDA} = \left(\frac{\boldsymbol{tps+tng}}{\boldsymbol{tps+tng+fps+fng}}\right) * \boldsymbol{100} \quad (15)$$

Where, $\boldsymbol{CDA}$ indicates an detection accuracy, $\boldsymbol{tps}$ denotes the true positive, $\boldsymbol{tng}$ denotes the true negative, $\boldsymbol{fps}$ represents the false positive, $\boldsymbol{fng}$ represents the false negative. It is measured in percentage (%). Higher the accuracy, the method is said to be more efficient.

Precision

It is the measures of ratio of true positives and false positives of cancer disease from the total number of data samples. Precision is calculated as follows,

$$PRC = \frac{\boldsymbol{tps}}{\boldsymbol{tps+fps}} \quad (16)$$

Where, PRC denotes a precision, $\boldsymbol{tps}$denotes a true positive that the data samples are correctly detected as cancer or normal, $\boldsymbol{fps}$ indicates a false positive refer to data samples incorrectly detected as cancer. Higher the precision, the method is said to be more efficient.

Cancer Detection Time

It is measured as the amount of time taken by algorithm for cancer detection. It is mathematically calculated as follows,

$$CDT = \sum_{i=1}^{n} S_i * tm\ [CD] \quad (17)$$

Where, CDT indicates the cancer detection time,$tm\ [CD]$ indicates a time for cancer detection of single data samples 'S_i'. The overall time is measured in terms of milliseconds (ms). Lesser the time consumption, the method is said to be more efficient.

Space complexity

It is measured as the amount of memory space consumed by algorithm for cancer detection. It is mathematically calculated as follows,

$$SCOM = \sum_{i=1}^{n} S_i * Mem\ [CD] \quad (18)$$

Where, $SCOM$ indicates the space complexity, $Mem\ [CD]$ indicates a memory space for cancer detection of single data samples 'S_i'. The space complexity is measured in terms of kilobytes (KB). Lesser the space complexity, the method is said to be more efficient.

Table 1 (a) and (b) illustrate a performance analysis of cancer detection accuracy versus the number of data samples collected from the lung cancer dataset and cervical cancer dataset. The overall comparison results illustrate that the PAWS-MRCTP technique increased accuracy by 5% compared to DNLC and 7% compared to XML-GBM when using the lung cancer dataset. Similarly, the overall comparison results illustrate that the PAWS-MRCTP technique increased accuracy by 4% compared to DNLC and 6% compared to XML-GBM when using the cervical cancer dataset.

Figure 3 (a) and (b) depict the performance outcomes of precision in cancer detection versus the number of data

Table 1 (a): comparison of cancer detection accuracy using lung cancer dataset

Number of data samples	*Cancer detection accuracy (%)*		
	PAWS-MRCTP	*DNLC*	*XML-GBM*
100	89	86	84
200	90.36	84.65	82.36
300	91.5	85.65	83.65
400	92.12	86.85	84.2
500	91.55	88.21	86.21
600	91.2	87.2	85.23
700	90.58	86.23	84.2
800	91.56	88.36	85.62
900	92.11	89.12	87.1
1000	90.2	87.11	85.2

Table 1 (b): Comparison of cancer detection accuracy using cervical cancer dataset

Number of data samples	*Cancer detection accuracy (%)*		
	PAWS-MRCTP	*DNLC*	*XML-GBM*
85	89.41	87.05	84.70
170	88.65	86.56	84.23
255	90.2	87.2	85.56
340	91.85	88.2	86.23
425	92.3	89.12	87.52
510	92.05	88.05	86.05
595	91.22	87.1	85.12
680	92.36	89.65	87.05
765	92.85	88.78	87.89
850	91.52	88.1	86.02

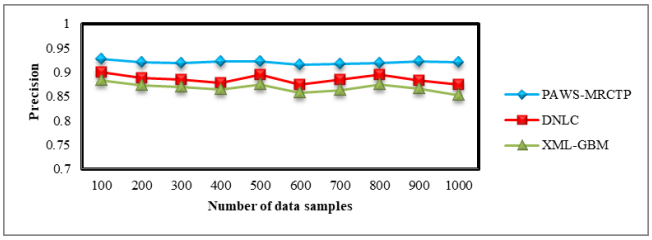

Figure 3 (a): performance results of precision using lung cancer dataset

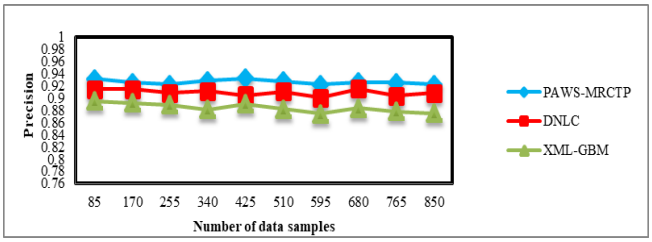

Figure 3 (b): performance results of precision using cervical cancer dataset

Table 2 (a): comparison of cancer detection time using lung cancer dataset

Number of data samples	*cancer detection time (ms)*		
	PAWS-MRCTP	*DNLC*	*XML-GBM*
100	22	25	27
200	24.2	27.5	30.8
300	26.3	28.2	31.4
400	30.5	32.6	34.5
500	32.4	34.5	36.8
600	33.8	35.1	38.9
700	36.5	38.9	40.5
800	38.4	40.5	42.4
900	40.1	42.3	44.5
1000	42.6	44.5	46.8

Table 2 (b): comparison of cancer detection time using cervical cancer dataset

Number of data samples	*Cancer prediction time (ms)*		
	PAWS-MRCTP	*DNLC*	*XML-GBM*
85	18.7	20.4	22.1
170	20.2	23.2	25.5
255	22.5	24.5	27.6
340	24.6	26.8	28.3
425	25.8	28.2	30.2
510	28.2	30.1	32.5
595	30.4	32.5	34.7
680	32.5	34.7	36.8
765	33.8	35.6	38.4
850	34.5	36.2	40.5

samples taken from the lung and cervical cancer datasets, respectively. A comparison of ten results are revealed that the precision performance during cancer detection increased by 4% and 6% than the DNLC and XML-GBM, when applying the lung cancer dataset. The overall performance result shows the precision performance during cancer detection increased by 2% and 5% compared to DNLC and XML-GBM, respectively, when applying the cervical cancer dataset.

Table 2 (a) and (b) illustrate the performance analysis of cancer detection time using three methods namely PAWS-MRCTP technique, and existing methods DNLC and XML-GBM. Through this analysis, it was found that the cancer detection time using the PAWS-MRCTP technique was minimized by 7% and 13% compared to DNLC and XML-GBM, respectively, when applying the lung cancer dataset. Similarly, when applying the cervical cancer dataset, the cancer detection time using the PAWS-MRCTP technique was minimized by 8% and 15% compared to DNLC and XML-GBM, respectively.

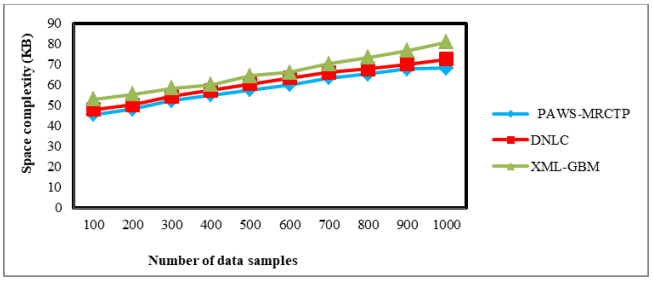

Figure 4 (a): performance results of space complexity using lung cancer dataset

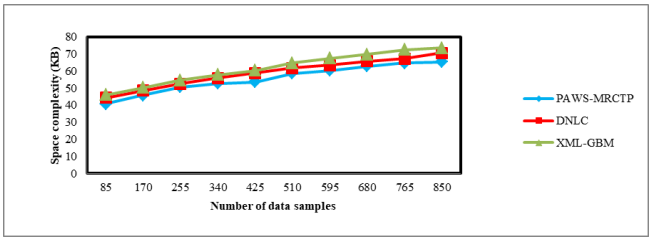

Figure 4 (b): performance results of space complexity using cervical cancer dataset

Figure 4 (a) and (b) illustrate the performance analysis of space complexity using three methods namely the PAWS-MRCTP technique, and existing methods DNLC and XML-GBM. Through this analysis, it was found that the space complexity of cancer detection using the PAWS-MRCTP technique was minimized by 4% and 11% compared to DNLC and XML-GBM, respectively. Similarly, when applying the cervical cancer dataset, the space complexity using the PAWS-MRCTP technique was reduced by 6% and 10% compared to DNLC and XML-GBM, respectively.

Conclusion

In this paper, a novel technique called PAWS-MRCTP is introduced for accurate lug and cervical cancer detection by combining two data preprocessing and feature selection to diagnose the disease at its early stage. Initially, data preprocessing is employed in the PAWS-MRCTP technique to prepare raw data into a suitable format. This involves handling missing data, detecting noisy data, and identifying outliers, thereby minimizing the time consumption of cancer detection. Following data preprocessing, the Multivariate Rosenthal Correlative Target Feature Projection technique is employed to select significant features from the cancer dataset. With these selected features, cancer disease is detected with higher accuracy. A comprehensive experiment is conducted using two different datasets namely lung cancer dataset and a cervical cancer dataset. Various metrics such as cancer detection accuracy, precision, and cancer detection time and space complexity are evaluated. The comparison results demonstrate that the presented PAWS-MRCTP technique achieves higher accuracy in cancer detection and precision with minimal cancer detection time as well as space complexity compared to existing methods.

References

Abbas, S., Issa, G. F., Fatima, A., Abbas, T., Ghazal, T. M., Ahmad, M., Yeob Yeun, C., & Khan, M. A. (2023). Fused weighted federated deep extreme machine learning based on intelligent lung cancer disease prediction model for healthcare 5.0. International Journal of Intelligent Systems, 2023, 1-14. https://doi.org/10.1155/2023/2599161

Al Mudawi, N., & Alazeb, A. (2022). A model for predicting cervical cancer using machine learning algorithms. Sensors, 22(11), 1-19. https://doi.org/10.3390/s22114132

Aljrees, T. (2024). Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. PLoS ONE, 19(1), 1-24. https://doi.org/10.1371/journal.pone.0295632

Chen, L., Chen, Y., Shi, H., & Cai, R. (2023). Enhancing prognostic accuracy: A SEER based analysis for overall and cancer specific survival prediction in cervical adenocarcinoma patients. Journal of Cancer Research and Clinical Oncology, 149, 17027–17037. https://doi.org/10.1007/s00432-023-05399-2

Chiu, H.-Y., Chao, H.-S., & Chen, Y.-M. (2022). Application of artificial intelligence in lung cancer. Cancers, 14(6), 1-17. https://doi.org/10.3390/cancers14061370

Chui, K. T., Gupta, B. B., Jhaveri, R. H., Chi, H. R., Arya, V., Almomani, A., & Nauman, A. (2023). Multiround transfer learning and modified generative adversarial network for lung cancer detection. International Journal of Intelligent Systems, 2023, 1-14. https://doi.org/10.1155/2023/6376275

Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. BDCC, 6(4), 1-14. https://doi.org/10.3390/bdcc6040139

Elseddeq, N. G., Elghamrawy, S. M., Salem, M. M., & Eldesouky, A. I. (2021). A selected deep learning cancer prediction framework. IEEE Access, 9, 151476–151492. https://doi.org/10.1109/ACCESS.2021.3124889

Elwahsh, H., Tawfeek, M. A., Abd El-Aziz, A. A., Mahmood, M. A., Alsabaan, M., & El-shafeiy, E. (2023). A new approach for cancer prediction based on deep neural learning. Journal of King Saud University - Computer and Information Sciences, 35(6), 1-12. https://doi.org/10.1016/j.jksuci.2023.101565

Hsu, J. C., Nguyen, P.-A., Phuc, P. T., Lo, T.-C., Hsu, M.-H., Hsieh, M.-S., Le, N. Q. K., Cheng, C.-T., Chang, T.-H., & Chen, C.-Y. (2022). Development and validation of novel deep-learning models using multiple data types for lung cancer survival. Cancers, 14(22), 1-14. https://doi.org/10.3390/cancers14225562

Ilyas, Q. M., & Ahmad, M. (2021). An enhanced ensemble diagnosis of cervical cancer: A pursuit of machine intelligence towards sustainable health. IEEE Access, 9, 12374–12388. https://doi.org/10.1109/ACCESS.2021.3049165

Jahan, S., Islam, M. D. S., Islam, L., Rashme, T. Y., Prova, A. A., Paul, B. K., Islam, M. D. M., & Mosharof, M. K. (2021). Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. SN Applied Sciences, 3, 1-17. https://doi.org/10.1007/s42452-021-04786-z

Johnson, M., Albizri, A., & Simsek, S. (2022). Artificial intelligence in healthcare operations to enhance treatment outcomes: A framework to predict lung cancer prognosis. Annals of Operations Research, 308, 275–305. https://doi.org/10.1007/s10479-020-03872-6

Kavitha, R., Jothi, D. K., Saravanan, K., Swain, M. P., Gonzáles,

J. L. A., Bhardwaj, R. J., & Adomako, E. (2023). Ant colony optimization-enabled CNN deep learning technique for accurate detection of cervical cancer. BioMed Research International, 2023, 1-9. https://doi.org/10.1155/2023/1742891

Khan, I. U., Aslam, N., Alshehri, R., Alzahrani, S., Alghamdi, M., Almalki, A., & Balabeed, M. (2021). Cervical cancer diagnosis model using extreme gradient boosting and bioinspired firefly optimization. Scientific Programming, 2021, 1-10. https://doi.org/10.1155/2021/5540024

Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A., Tarczewska, M., Kosowska, M., & Szczerska, M. (2022). Predictions of cervical cancer identification by photonic method combined with machine learning. Scientific Reports, 12, 1-11. https://doi.org/10.1038/s41598-022-07723-1

Lei, J., Xu, X., Xu, J., Liu, J., Wang, Y., Wu, C., Zhang, R., Zhang, Z., & Jiang, T. (2023). The predictive value of modified-DeepSurv in overall survivals of patients with lung cancer. iScience, 26(11), 1-11. https://doi.org/10.1016/j.isci.2023.108200

Li, J., Tao, Y., & Cai, T. (2021). Predicting lung cancers using epidemiological data: A generative-discriminative framework. IEEE/CAA Journal of Automatica Sinica, 8(5), 1067–1078. https://doi.org/10.1109/JAS.2021.1003910

Li, Y., Wu, X., Yang, P., Jiang, G., & Luo, Y. (2022). Machine learning for lung cancer diagnosis, treatment, and prognosis. Genomics, Proteomics & Bioinformatics, 20(5), 850-866. https://doi.org/10.1016/j.gpb.2022.11.003

Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., Vijayakumar, V., Tunze, G. B., & Hamdi, M. (2022). Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. Computational and Mathematical Methods in Medicine, 2022, 1-17. https://doi.org/10.1155/2022/4688327

Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems, 106, 199-205. https://doi.org/10.1016/j.future.2019.12.033

Mithun, S., Jha, A. K., Sherkhane, U. B., Jaiswar, V., Purandare, N. C., Rangarajan, V., Dekker, A., Puts, S., Bermejo, I., & Wee, L. (2023). Development and validation of deep learning and BERT models for classification of lung cancer radiology reports. Informatics in Medicine Unlocked, 40, 1-10. https://doi.org/10.1016/j.imu.2023.101294

Munawar, Z., Ahmad, F., Alanazi, S. A., Nisar, K. S., Khalid, M., Anwar, M., & Murtaza, K. (2022). Predicting the prevalence of lung cancer using feature transformation techniques. Egyptian Informatics Journal, 23(4), 109-120. https://doi.org/10.1016/j.eij.2022.08.002

Munshi, R. M. (2023). Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLoS ONE, 19(1), 1-20. https://doi.org/10.1371/journal.pone.0296107

Rikta, S. T., Uddin, K. M. M., Biswas, N., Mostafiz, R., Sharmin, F., & Dey, S. K. (2023). XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer. Journal of Pathology Informatics, 14, 1-16. https://doi.org/10.1016/j.jpi.2023.100307

Senthilkumar, G., Ramakrishnan, J., Frnda, J., Ramachandran, M., Gupta, D., Tiwari, P., Shorfuzzaman, M., & Mohammed, M. A. (2021). Incorporating artificial fish swarm in ensemble classification framework for recurrence prediction of cervical cancer. IEEE Access, 9, 83876–83886. https://doi.org/10.1109/ACCESS.2021.3087022

Sujitha, R., & Seenivasagam, V. (2021). Classification of lung cancer stages with machine learning over big data healthcare framework. Journal of Ambient Intelligence and Humanized Computing, 12, 5639–5649. https://doi.org/10.1007/s12652-020-02071-2

Wang, Y., Gao, X., Ru, X., Sun, P., & Wang, J. (2023). The weight-based feature selection (WBFS) algorithm classifies lung cancer subtypes using proteomic data. Entropy, 25(7), 1-16. https://doi.org/10.3390/e25071003

Yang, A.-M., Han, Y., Liu, C.-S., Wu, J.-H., & Hua, D.-B. (2021). D-TSVR recurrence prediction driven by medical big data in cancer. IEEE Transactions on Industrial Informatics, 17(5), 3508–3517. https://doi.org/10.1109/TII.2020.3011675

Zhan, X., Long, H., Gou, F., Duan, X., Kong, G., & Wu, J. (2021). A convolutional neural network-based intelligent medical system with sensors for assistive diagnosis and decision-making in non-small cell lung cancer. Sensors, 21(23), 1-24. https://doi.org/10.3390/s21237996