The Scientific Temper (2024) Vol. 15 (3): 2940-2948
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.66
https://scientifictemper.com/

RESEARCH ARTICLE

A hybrid feature selection and generative adversarial network for lung and uterus cancer prediction with big data

V. Babydeepa, K. Sindhu*

Abstract

Among all diseases affecting humanity, lung cancer has consistently stood out as one of the deadliest. It ranks among the most prevalent cancers and is a significant contributor to cancer-related deaths. The disease is often asymptomatic in its early stages, making early detection extremely challenging. To enhance the accuracy of cancer detection with minimal time, an effective hybrid feature selection and classification model is developed in this research for the efficient detection of detect lung and uterus cancers while leveraging big data. The Piecewise Adaptive Weighted Smoothing-based Multivariate Rosenthal Correlative Target Projection (PAWS-MRCTP) comprises three main processes namely data acquisition, preprocessing, and feature extraction. 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. Then the Improved Particle Swarm Optimization (IPSO) with fuzzy possibility C-Means clustering (FPCM) is introduced for the data clustering. And then the hybrid feature selection is performed using the ANFIS and Modified Chicken Swarm Optimization (MCSO). Finally, the classification of uterine and lung tumors is done using the Generative Adversarial Network (GAN). Consequently, in the experiments, the proposed model beats existing classifiers in detection accuracy while consuming the least time.

Keywords: Lung and Uterus cancer, Improved Particle Swarm Optimization (IPSO) with fuzzy possibilitic C-Means clustering (FPCM), ANFIS and Modified Chicken Swarm Optimization (MCSO) and Generative Adversarial Network (GAN).

Introduction

The lungs are the primary organs involved in respiration. Humans have a pair of lungs, situated on either side of the chest. The left lung is smaller than the right due to the presence of the heart. The lungs expand during consumption and deflate during expiration, causing the chest to expand and contract during breathing. The lungs have a vital function in the process of oxygenating the blood. While the lungs have mechanisms such as phlegm production to clear themselves, these are insufficient for smokers. Lung health can be influenced by environmental factors, genetics, heredity, or a combination of these, contributing to the development and progression of various respiratory diseases. Fertility is a factor which can influence a household's dignity. While infertility is an issue that exists in women and men's reproductive systems. Uterine fibroids are noncancerous growths of the uterus and often occur during childbirth. Since these tests are very costly, many people use ultrasound (USG) tests to produce an ultrasound. Patients with PCOS can be said to have PCOS according to the Conference of Rotterdam if they have two or more symptoms: (1) ovulation failure, (2) elevated androgens hormones, or (3) polycystic ovary conditions. Morphologically, it may be said if there are twelve or more polycystic ovaries of 2-9 mm diameter or ovarian volumes of more than 10 cm3 Malvezzi, M., Santucci, C., Boffetta, P., Collatuzzo, G., Levi, F., La Vecchia, C., & Negri, E. (2023), Carioli, G., Malvezzi, M., Bertuccio, P., Boffetta, P., Levi, F., La Vecchia,

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). A hybrid feature selection and generative adversarial network for lung and uterus cancer prediction with big data. The Scientific Temper, **15**(3):2940-2948.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.66

Source of support: Nil

Conflict of interest: None.

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

C., & Negri, E. (2021), Hippisley-Cox, J., & Coupland, C. (2015), Bhardwaj, V., Sharma, A., Parambath, S. V., Gul, I., Zhang, X., Lobie, P. E., ... & Pandey, V. (2022), Baker, W., Pelkofski, E., Te Paske, J., Erickson, S., & Duska, L. (2015), ÇElik, Ç., ÖZdemir, S., Esen, H., Balcı, O., & Yılmaz, O. (2010) .

The ultrasound image is manually examined by a doctor when the follicles are counted in the ova. The examination, however, takes a long time and takes a great deal of precision whether the patient has or has not polycystic ovary syndrome. An automated system capable of detecting PCO using ultrasound images may address this issue. Thus, this study could allow physicians to detect PCO with ultrasonic images more easily. Ovarian cancer is an ovarian cancer type. In healthcare, big data processing involves the generation, collection, analysis, and storage of clinical data that is too large or complex to be interpreted using traditional data processing methods. To enhance the accuracy of cancer detection with minimal time, this research presents a resilient hybrid model that combines feature selection and classification techniques. for the efficient detection of detect lung and uterus cancers while leveraging big data Gien, L. T., Barbera, L., Kupets, R., Saskin, R., & Paszat, L. (2009), Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019).

Literature Review

This section provides an overview of the latest approaches in deep learning and machine learning for identifying cancer in the uterus and lungs.

Kaggle Data Science Bowl 2017, and LIDC-IDRI datasets' 3D cubes were used to evaluate the proposed 03D multipath VGG-like network.0.387732 log loss and 95.60% accuracy, this architecture identifies lung nodules and diagnoses malignancy. Tumor classification models were created using decision trees, support vector machines, and feature selection methods (Wrappers and Relief-F). The findings imply that adding additional characteristics might be advantageous since it is feasible to identify tumor classes with about 68% accuracy over a wide feature space that includes both 2D and 3D features. The 2D and 3D features provide similar levels of accuracy; but, 3D features are more user-friendly. Lung-EffNet is a brand-new transfer learning-based predictor that was presented for the categorization of lung cancer. Adding top layers to the model's classification head improves Lung-EffNet, which is based on the Efficient Net architecture. The study introduced an image processing workflow for lung cancer classification, including feature selection, categorization, data pre-processing, and data collection. A median filter was used during pre-processing to eliminate noise from the images. The parameters that were returned were the area, perimeter, and centroid. The categorization of lung cancer was then performed using these characteristics as inputs. Analysis revealed that 98.15% accuracy was attained using the kNN approach. A novel approach that uses a contrast stretching-based classical features fusion methodology was put forward for the categorization of lung cancer. Three primary stages comprise the method: First, the contrast of the original CT images is enhanced using a gamma correction max intensity weights approach. Second, a fusion method based on serial canonical correlation is used to merge different textures, points, and geometric characteristics that have been derived from the improved images. Pre-trained deep-learning models for UF diagnosis from medical images may be considerably improved by fine-tuning. More specifically, InceptionV3 has 90% accuracy and ResNet50 89%. The VGG16 model did, however, show a lesser accuracy of 85%, which is remarkable Tekade, R., & Rajeswari, K. (2018, August), Basu, S., Hall, L. O., Goldgof, D. B., Gu, Y., Kumar, V., Choi, J., ... & Gatenby, R. A. (2011), Raza, R., Zulfiqar, F., Khan, M. O., Arif, M., Alvi, A., Iftikhar, M. A., & Alam, T. (2023), Abdullah, M. F., Sulaiman, S. N., Osman, M. K., Karim, N. K. A., Shuaib, I. L., & Alhamdu, M. D. I. (2020), Khan, M. A., Rubab, S., Kashif, A., Sharif, M. I., Muhammad, N., Shah, J. H., ... & Satapathy, S. C. (2020).

Proposed Methodology

A robust hybrid feature selection and classification model is developed in this research study for the efficient detection of detect lung and uterus cancers while leveraging big data. The Piecewise Adaptive Weighted Smoothing-based Multivariate Rosenthal Correlative Target Projection (PAWS-MRCTP) comprises three main processes namely data acquisition, preprocessing, and feature extraction. In the

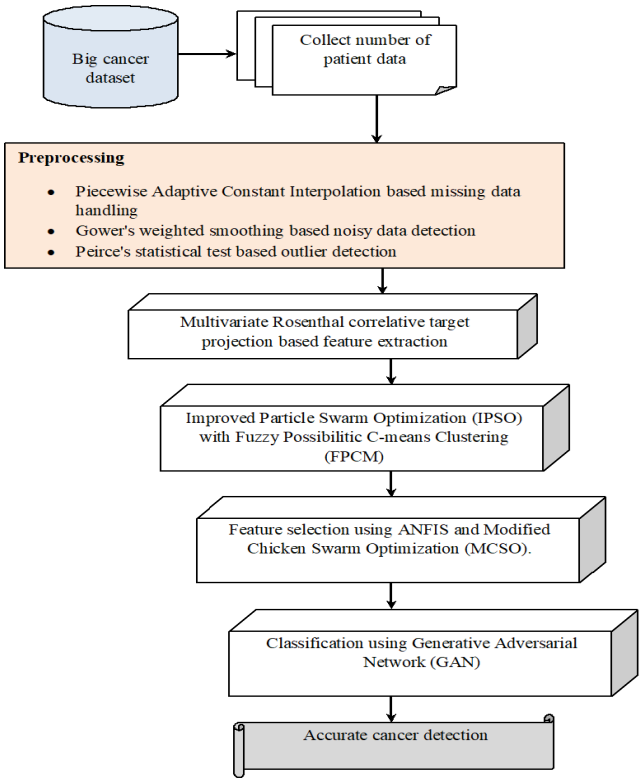

Figure 1: The proposed technique's architecture

data acquisition phase, a large number of cancer patient data are collected from lung cancer and uterus cancer detection datasets (Figure 1). 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. Then the Improved Particle Swarm Optimization with Fuzzy Possibilitic C-means (IPSO-FPCM) is introduced for the data clustering. And then the hybrid feature selection is performed using the ANFIS and Modified Chicken Swarm Optimization (MCSO). Finally, the Generative Adversarial Network (GAN) is introduced of the classifying the lung and uterus cancers.

Data acquisition

Data acquisition involves gathering patient information from a dataset, with the cancer prediction data sourced from Kaggle. This dataset encompasses various details about lung cancer patients, such as age, gender, environmental exposures, lifestyle factors, genetic predispositions, and symptoms like chest pain and coughing. Similarly, the dataset for cervical cancer detection, also from Kaggle, addresses the pressing issue of cervical cancer, a leading cause of death among women globally. Early identification and precise forecasting of cervical cancer greatly enhance treatment efficacy and ultimately save lives. This dataset includes 36 features and 858 data instances.

Data preprocessing

Cleaning, converting, and arranging raw data into an appropriate format constitute data preparation, a crucial stage in data analysis. Improving data quality and optimizing machine learning algorithm performance are the main objectives of data preparation. The ANFISMCSA-GAN 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. As missing data has an impact on statistical research of cancer detection, handling missing data is a crucial stage in data analysis. 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. Utilizing a discrete collection of known data points within the dataset as a basis, this mathematical approach finds new data points.

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.

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 the missing 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.

According to the separation between each observation, the suggested smoothing method gives weights to the data points. The function gives closer data points a greater weight and further data points a lower weight. 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}means the two data points differ completelyDp_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. This indicates that a larger weight is applied if the data point absolute difference is below

the threshold. Conversely, a lower weight is given if The data point absolute difference is below the threshold. 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', is used for smoothing the noisy data points within the dataset.

Peirce's statistical test-based outlier detection

Outlier data are data points that deviate significantly from the remainder of 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 test is give below,

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

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

Where, Dp_i denotes a data points, avg_{dp} denotes average of the data points, DF indicates a deviation 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 a normal 'ND'. Based on this analysis, outlier data points are removed for further processing.

Data Clustering

The Improved Particle Swarm Optimization (IPSO) with fuzzy possibilistic C-Means clustering (FPCM) is introduced for the data clustering.

Particle Swarm Optimization (PSO)

Multiple particles working together as a swarm to explore the search space for the best answer are used in PSO. Based on its own experiences as well as those of other particles, each particle in a multidimensional space represents a point and modifies its movement accordingly. In search of the ideal answer, these particles travel around the multidimensional space at a set velocity.

The particle's velocities are expressed as $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, the particle I indicates its position as $(x, x_{i2}, \ldots, x_{iD})$, particle I's ideal location is expressed as $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$, it is also called p_{best}.

All particle positions at their global optimum is expressed as $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$, it is also called g_{best}. To determine the fitness value, fitness functions exist for each group particle. The velocity update formula for dimension d in conventional PSO is represented in formulas (11) and (12):

$$v_{id} = w \times v_{id} + c_1 \times rand() \times (p_{id} - x_{id}) + c_2 \times Rand() \times (p_{gd} - x_{id}) \quad (11)$$

$$(X_{id} = x_{id} + v_{id}) \quad (12)$$

The population quantity (Q), inertia weight (w), acceleration constants C1 and C2, maximum velocity (v_{max}), maximum number of iterations (G_{max}), and random functions rand() and rand() with values in [0,1] are among the PSO parameters. Typically, the values of C1 and C2 take constant 2.

During optimization processes, IPSO (Incremental Particle Swarm Optimization) improves the exchange of information among populations and preserves population diversity, addressing challenges inherent in classical optimization algorithms when tackling complex engineering optimization problems characterized by multiple parameters, strong coupling, and nonlinearity. These challenges include premature convergence and susceptibility to getting trapped in local optima. The «local-global information sharing» mechanism, a key aspect of IPSO, is thoroughly examined to determine optimal parameter settings for improved performance. Subsequently, IPSO's effectiveness is evaluated alongside classical optimization algorithms using various standard functions to assess its global search capabilities.

To improve the PSO algorithm's effectiveness in finding better solutions while preserving its simplicity and rapid convergence, this study will provide an IPSO variation. This improvement is achieved by incorporating a straightforward but powerful new operation into the iterative search process. The objective of this modification is to improve the algorithm's capacity to successfully use intermediate answers and to seek additional areas of the search space for maybe better results. The modified form of PSO that depends on certain parameter combinations serves as the basis for the suggested variation.

- *Distraction factor*

Due to the typically high dimensionality of a feature vector, particles in PSO tend to converge prematurely before finding the global optimum. To address this, a distraction factor \ (K \) is introduced into PSO to ensure optimal convergence. The velocity formula is shown in Equation (11):

$$v_{id} = K[v_{id} + c_1 \times rand() \times (p_{id} - x_{id}) + c_2 \times Rand() \times (p_{gd} - x_{id})] \quad (13)$$

Algorithm 1 in this study employs the proposed formula to calculate the distraction factor \ (K \). The values for \(c1 \) and \(c2 \) are set to 2.05, consistent with Clerc's experiment. For the purposes of the experiment, \(K \) is rounded to four decimal places. The detailed velocity formula is given in Equation (13):

$$v_{id} = 0.7298 \times [v_{id} + 2.05 \times rand() \times (p_{id} - x_{id}) + 2.05 \times Rand() \times (p_{gd} - x_{id}) \quad (14)$$

Based on this principle, the functional form of the distraction factor, structured using the cosine function, is presented in Equation (15):

$$K = \frac{\cos((\pi/G_{max}) \times T) + 2.5}{4} \quad (15)$$

T iterations. The K-changing curve appeared when G_{max} G_{max} was 40. K curve in the function is first convex and then

becomes concave. Formula (6) is changed to formula (16) once the value K is inserted. The following is an explanation of formula (16):

$$v_{id} = \left(\frac{\cos((\pi \times T/G_{max}))\times 2.5}{4}\right) \times [v_{id} + 2 \times rand() \times (p_{id} - x_{id}) + 2 \times Rand() \times (p_{gd} - x_{id})] \quad (16)$$

However, the process is slowdown and also iterations are also take more time to convergence. The fuzzy Possibilistic c-means algorithm clustering method is suggested as a solution to this issue since it speeds up the whole process more quickly than a single algorithm.

Fuzzy Possibilistic C Means (FPCM) clustering algorithm

A possibilistic membership function is used by the PCM method to represent the degree of belonging. An advantage of this approach is that representative feature points have high membership values, while non-representative points have low membership values. Each cluster in the PCM method corresponds represents a concentrated area within the dataset, and all clusters are separate from each other. The objective is to minimize the objective function Kannan, S. R., Devi, R., Ramathilagam, S., & Hong, T. P. (2017).

$$J_{PCM} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{ij})^m \quad (17)$$

where η_i , appropriate positive values represent the i-th cluster's scale parameter. The first condition requires the u_{ij} to be as big as feasible, while the second phase mandates the minimum distances between data points and the prototypes. The components of U meet the requirements listed below:

$$u_{ij} \epsilon [0,1], \qquad \forall_i \ \ and \ j \quad (18)$$

$$0 < \sum_{j=1}^{n} u_{ij} < n, \qquad \forall_i \quad (19)$$

It is recommended to select ηi as

$$\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{n} u_{ij}^m} \qquad for\ 1 \le i \le c \quad (20)$$

d_{ij}^2 is the value of sample xj possibilistic typicality, and it belongs to cluster i. m Є [1, ∞] the possibilistic parameter, a weighting factor. The FPCM clustering algorithm merges the features combining possibilistic and fuzzy c-means techniques. In this approach to effectively reflect the data substructure within a clustering issue, memberships and typicalities are essential. The following is the expression for the FPCM goal function, which takes memberships and typicalities into account:

$$Minimize\ J_{FPCM} - \sum_{i=1}^{c}\sum_{j=1}^{n} (u_{ij}^m + t_{ij}^\eta)\, d_{ij}^2 \quad (21)$$

with the following constraints

$$\sum_{i=1}^{c} u_{ij} = 1, \qquad for\ 1 \le j \le n$$

$$\sum_{i=1}^{c} t_{ij} = -1, \qquad for\ 1 \le j \le c$$

Feature Selection

Then the hybrid feature selection is performed using the ANFIS and Modified Chicken Swarm Optimization (MCSO).

ANFIS

Neuro-fuzzy systems are neural networks that utilize fuzzy logic to identify properties (such as sets and rules) by processing datasets. When modelling non-linear functions, ANFIS has shown impressive results. A dataset that depicts the behavior of the system is used to generate membership functions (MFs) in this situation. According to specified error criteria, ANFIS modifies system settings by using characteristics extracted from the Pima Indian diabetes dataset. Neural network techniques in this design are used to forecast the parameters of the fuzzy inference system and to assist learning, Squares denote adaptive nodes in this figure, whereas circles represent fixed nodes. Let's examine two inputs, «S» and «MS,» which stand for missing and non-missing samples, respectively, to make the explanation simpler to understand. The output, represented by \(z \), is part of the Fuzzy Inference System (FIS). This research proposes an ANFIS to implement a first-order Sugeno fuzzy model, which is defined as follows Al-Hmouz, A., Shen, J., Al-Hmouz, R., & Yan, J. (2011):

The hybrid learning algorithm addresses this issue by integrating the least squares method with backpropagation (BP). As the learning process progresses, the layers 1 and 4 parameters, both premise and consequent, are continually adjusted until the FIS produces the desired response. This procedure involves two distinct steps.

- *Step 1*

The functional values are the only ones that advance to the fourth layer by maintaining the premise parameters as constraints, and the resulting parameters are found by the least squares technique.

- *Step 2*

The subsequent parameters in this method are maintained constant while the error values propagate from the output to the input, represented by the derivatives of the error measure for each node output. After that, The BP method is used to update the premise parameters. For the differential equation under consideration, a trial solution would be as follows: (22).

$$y_p(t) = f_i\left(t, y_0^{(0)}, y_0^{(1)}, \ldots, y_0^{(n-1)}\right) + g(h_i, u) \quad (22)$$

$$= f_i(t, C) + g(h_i, u) \quad (23)$$

where f_i(t, C) satisfies the initial requirements as a function and g(h_i, u) A function is defined to have a value of 0 at the starting points and a value of 'u' at all other points. In the absence of data points at the first stage, the unsupervised neuro-fuzzy inference system 'u' uses the ANFIS with the corrected samples 'S' to detect missing values. Later, u is the primary response. The value of u, which is specified in terms of a fuzzy logic system, is adjusted using a hybrid learning strategy that combines the least squares approach with the backpropagation algorithm. At the beginning and boundary

points, hi is used to minimize the g(h_i, u) function. Thus, h_iu is a flexible variable for the function g(h_i, u).

The determination of fi (t, C) and hi may vary depending on the initial conditions and the series of differential equations used to choose the most suitable ones. The formula (24) is used to normalize the unsupervised filter for data normalization, with x' representing the variable's arithmetic mean and sd representing its standard deviation. This process ensures that the data is transformed into the range [0, 1].Value is the new normalized value. This speeds up computations and reduces computation complexity.

$$Value = \frac{Value - x'}{sd} \quad (24)$$

Modified Chicken Swarm Optimization (MCSO)

Chicken Swarm Optimization (CSO) is an innovative optimization approach that draws inspiration from biological processes. The actions and hierarchical structure of a swarm of chickens are mimicked by this program. A rooster and many hens and chicks are housed in each group that the swarm is separated. Different hens have different ways of moving, and within the set order, they compete with each other. The dominant hens in a flock exercise authority over the weaker hens, and this hierarchy is important to understanding the social dynamics of hens. At the outside of the group, the more submissive hens and roosters are situated, While the most powerful hens remain near the front, the roosters also do the same. Adding or removing chickens from a group temporarily disrupts the social order until a new hierarchy is established Gandomi, A. H., Yang, X. S., & Alavi, A. H. (2013).

- *Movement of the chickens*

Roosters with higher fitness ratings get prioritized over those with lower fitness values when it comes to food access. To simplify, a broader region may be searched for food by higher-valued roosters are more fit than lower-valued ones, simulating this situation. The following is an expression for this concept:

$$x_{i,j}^{t+1} = x_{i,j}^{t} * (1 + Randn(0,\sigma^2)) \quad (25)$$

$$\sigma^2 = \begin{cases} 1, if\ f_i \le f_k \\ \exp\left(\frac{(f_k - f_i)}{|f_i| + \varepsilon}\right), otherwise, \end{cases} k \in [1,N], k \ne i \quad (26)$$

$$x_{i,j}^{t+1} = x_{i,j}^{t} + S1 * Rand * \left(x_{r1,j}^{t} - x_{i,j}^{t}\right) + S2 * Rand * \left(x_{r2,j}^{t} - x_{i,j}^{t}\right) \quad (27)$$

$$S1 = \exp((f_i - f_{r1})/(abs(f_i) + \varepsilon)) \quad (28)$$

$$S2 = \exp((f_{r2} - f_i) \quad (29)$$

In this case, the uniform random number across [0, 1] is denoted *Rand*.$r_1 \in [1, \ldots, N]$is a measure of the rooster, a group member of the ith hen, while $r_2 \in [1, \ldots, N]$is a random selection made from a flock of chickens, either a hen or a rooster. $r_1 \ne r_2$.Obviously, $f_i > f_{r1}, f_i > f_{r2}$, thus *S2* <1< *S1*. The *ith* hen would search for food first, followed by other hens, assuming S1=0. The chicks follow their mother around in search of food. Below is how this is put together.

$$x_{i,j}^{t+1} = x_{i,j}^{t} + FL * (x_{m,j}^{t} - x_{i,j}^{t}) \quad (30)$$

Where $x_{m,j}^{t}$ stands for the status of the mother of the ith chick $(m \in [1,N]).FL(FL \in (0,2))$ The parameter signifies the chick follows its mother for food. Each chick's FL would random between 0 and 2 due to individual variations.

Classification using Generative adversarial network (GAN)

Using a GAN, extract knowledge from an unknown data distribution and produce similar samples. A GAN has a generator and a discriminator. Its primary purpose is to generate samples with a distribution that fits actual data. Separating authentic from fraudulent data requires the discriminator to assess original and created samples. The generation (G) and discrimination (D) models optimize separately via alternating iterative training. The loss function for the whole network Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020):

$$\min_c \max_D L(D,G) = E_{x\sim P_r}\left[\log\left(D(x)\right)\right] + E_{x\sim P_g}\left[\log\left(D(x')\right)\right] \quad (31)$$

In this context, *x* denotes real data, while *x'* signifies fake sample data created by the generator G. The distribution of created data is represented by Pg, while the distribution of actual data is represented by Pr. When Pr and Pg share identical distributions, it is difficult for discriminator D to distinguish between authentic and fraudulent samples, resulting in a probability of 0.5. Consequently, the generator is capable of producing sufficiently realistic samples.

ACGAN merges the strengths of various generative adversarial network models, including CGAN, SGAN, and infoGAN. Each produced sample in ACGAN has a class label. To distinguish produced samples, these labels are one-hot encoded. The generator G generates Xfake = G(z, c) from noise z and class label c, while the discriminator D assesses actual and fake sample likelihood and class label probability.

$$D(X) = P(S|X).P(C|X) \quad (32)$$

$$L_s = E[logp(s = real|X_{real})] + E\left[logp\left(s = fake|X_{fake}\right)\right] \quad (33)$$

$$L_c = E[logp(C = c|X_{real})] + E\left[logp\left(C = c|X_{fake}\right)\right] \quad (34)$$

Discriminator D aims to maximize $L_s + L_c$, whereas generator G aims to maximize L_s- L_c throughout training. To restore the original data distribution after information loss, a new regularization penalty function is needed.We evaluate the informational disparity between the data generated and the data source in terms of similarity and distance to identify the information loss. There is an explanation of the distance loss function:

$$L_{dis} = \left\| E(f_x)_{x\sim p_r(x)} - E(f_{x'})_{x'\sim p_r(x')} \right\|_2 \quad (35)$$

In this context, fx and fx' denote the multi-dimensional characteristics of initial and generator-produced samples. The function 'E(·)' averages out the characteristics of all samples within a batch. Specifically, cosine similarity is employed in this research to evaluate the resemblance between the original and generated data.

Results and Discussion

In this section, the experimental evaluation of the proposed ANFISMCSA-GAN technique and existing PAWS-MRCTP,DNLC and XML-GBM are implemented using Python coding. To conduct the experiment, two cancer dataset namely lung and cervical dataset are used. The https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link is where the lung cancer prediction dataset may be found.

Figure 2 and Table 1 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 observed result shows that the accuracy of the PAWS-MRCTP technique was higher for both datasets compared to the existing methods DNLC and XML-GBM. Regarding the quantity of data samples, ten distinct results were noted for every technique. Finally, the observed results of the ANFISMCSA-GAN technique are compared to the existing methods.

Figure 3 and Table 2 depict the performance outcomes of precision in cancer detection versus the number of data samples taken from the lung and cervical cancer datasets, respectively. The graph illustrates the number of data samples on the 'x' axis and the precision performance observed on the 'y' axis. Among the three methods, ANFISMCSA-GAN demonstrates improved precision performance compared to other existing methods. This improvement is achieved by utilizing the target feature projection based on correlation functions. These selected target features are employed to classify normal and cancer data samples, minimizing false positives and increasing the true positive rate.

Figure 4 and Table 3 illustrate the performance analysis of cancer detection time using three methods namely ANFISMCSA-GAN technique, and existing methods PAWS-MRCTP,DNLC, and XML-GBM. For each method, 10 runs were performed with distinct numbers of data samples. From the graph, it is evident that increasing the number

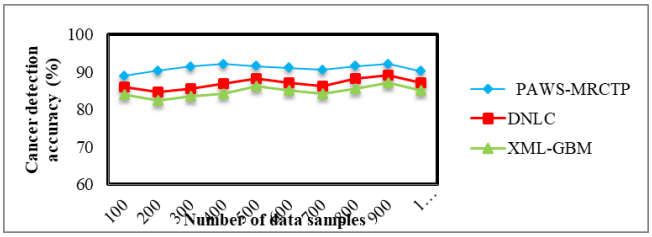

Figure 2: performance results of cancer detection accuracy using lung cancer dataset

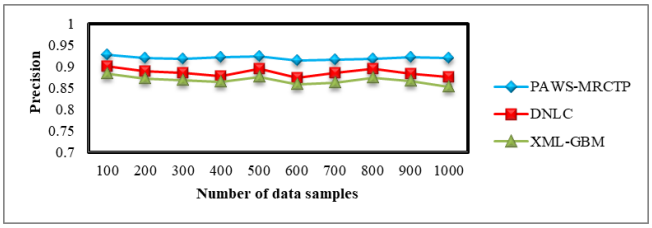

Figure 3: performance results of precision using lung cancer dataset

Table 1: Comparison of cancer detection accuracy using lung cancer dataset

Number of data samples	*Cancer detection accuracy (%)*			
	ANFISMCSA-GAN	*PAWS-MRCTP*	*DNLC*	*XML-GBM*
100	91.45	89	86	84
200	90.85	90.36	84.65	82.36
300	91.36	91.5	85.65	83.65
400	92.35	92.12	86.85	84.2
500	92.68	91.55	88.21	86.21
600	91.56	91.2	87.2	85.23
700	90.92	90.58	86.23	84.2
800	92.36	91.56	88.36	85.62
900	91.57	92.11	89.12	87.1
1000	91.67	90.2	87.11	85.2

Table 2: Comparison of precision using lung cancer dataset

Number of data samples	*Precision*			
	ANFISMCSA-GAN	*PAWS-MRCTP*	*DNLC*	*XML-GBM*
100	0.934	0.927	0.9	0.883
200	0.947	0.92	0.889	0.872
300	0.927	0.918	0.885	0.869
400	0.937	0.922	0.878	0.865
500	0.947	0.923	0.895	0.875
600	0.938	0.915	0.874	0.858
700	0.927	0.917	0.885	0.862
800	0.948	0.918	0.895	0.874
900	0.937	0.922	0.883	0.866
1000	0.948	0.921	0.875	0.852

Table 3: Comparison of cancer detection time using lung cancer dataset

Number of data samples	*cancer detection time (ms)*			
	ANFISMCSA-GAN	*PAWS-MRCTP*	*DNLC*	*XML-GBM*
100	21	22	25	27
200	22.54	24.2	27.5	30.8
300	24.51	26.3	28.2	31.4
400	26.41	30.5	32.6	34.5
500	24.57	32.4	34.5	36.8
600	30.25	33.8	35.1	38.9
700	31.54	36.5	38.9	40.5
800	32.68	38.4	40.5	42.4
900	36.74	40.1	42.3	44.5
1000	39.57	42.6	44.5	46.8

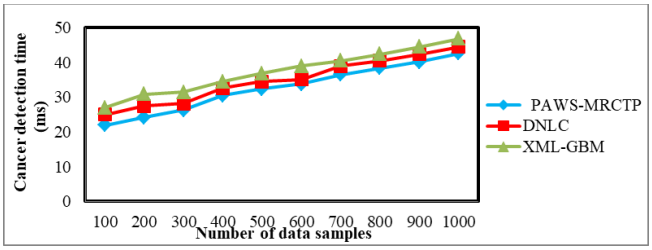

Figure 4: Performance results of cancer detection time using lung cancer dataset

of data samples also increases the time incurred for cancer detection. Through this analysis, it was found that the cancer detection time using the ANFISMCSA-GANtechnique was minimized by 7% and 13% compared to existing classifiers respectively, when applying the lung cancer dataset.

Conclusion

The lungs serve as the primary respiratory organs in the human body. Breathing continues until death as the lungs provide oxygen to the blood, crucial for sustaining life. Lung cancer ranks as the top cause of mortality among both men and women due to malignant tumors. 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. Then the Improved Particle Swarm Optimization (IPSO) with fuzzy possibility C-Means clustering (FPCM) is introduced for the data clustering. And then the hybrid feature selection is performed using the ANFIS and Modified Chicken Swarm Optimization (MCSO). Finally, the Generative Adversarial Network (GAN) is introduced of the classifying the lung and uterus cancers. The pre-operative steps are mostly determined by personal knowledge and experience, which differ from person to person. The results were reasonable. From the result it is found that the proposed ANFISMCSA-GAN model provide the best detection accuracy than the existing methods.

References

Abdullah, M. F., Sulaiman, S. N., Osman, M. K., Karim, N. K. A., Shuaib, I. L., & Alhamdu, M. D. I. (2020, August). Classification of lung cancer stages from CT scan images using image processing and k-Nearest neighbours. In 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC) (pp. 68-72). *IEEE*. https://doi.org/10.1109/ICSGRC49013.2020.9232662

Al-Hmouz, A., Shen, J., Al-Hmouz, R., & Yan, J. (2011). Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning. *IEEE Transactions on Learning Technologies*, 5(3), 226-237. https://doi.org/10.1109/TLT.2011.12

Baker, W., Pelkofski, E., Te Paske, J., Erickson, S., & Duska, L. (2015). Preoperative imaging of uterine malignancy: A low-value service. *Gynecologic Oncology*, 137(2), 285-290. https://doi.org/10.1016/j.ygyno.2015.02.021

Basu, S., Hall, L. O., Goldgof, D. B., Gu, Y., Kumar, V., Choi, J., ... & Gatenby, R. A. (2011, October). Developing a classifier model for lung tumors in CT-scan images. In 2011 IEEE International Conference on Systems, Man, and Cybernetics (pp. 1306-1312). *IEEE*. https://doi.org/10.1109/ICSMC.2011.6083862

Bhardwaj, V., Sharma, A., Parambath, S. V., Gul, I., Zhang, X., Lobie, P. E., & Pandey, V. (2022). Machine learning for endometrial cancer prediction and prognostication. *Frontiers in Oncology*, 12, 852746. https://doi.org/10.3389/fonc.2022.852746

Carioli, G., Malvezzi, M., Bertuccio, P., Boffetta, P., Levi, F., La Vecchia, C., & Negri, E. (2021). European cancer mortality predictions for the year 2021 with focus on pancreatic and female lung cancer. *Annals of Oncology*, 32(4), 478-487. https://doi.org/10.1016/j.annonc.2020.12.011

Çelik, Ç., Özdemir, S., Esen, H., Balcı, O., & Yılmaz, O. (2010). The clinical value of preoperative and intraoperative assessments in the management of endometrial cancer. *International Journal of Gynecologic Cancer*, 20(3), 358-362. https://doi.org/10.1111/IGC.0b013e3181d6de64

Gandomi, A. H., Yang, X. S., & Alavi, A. H. (2013). Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems. *Engineering with Computers*, 29, 17-35. https://doi.org/10.1007/s00366-011-0241-y

Gien, L. T., Barbera, L., Kupets, R., Saskin, R., & Paszat, L. (2009). Utilization of preoperative imaging in uterine cancer patients. Gynecologic Oncology, 115(2), 226-230. https://doi.org/10.1016/j.ygyno.2009.07.023

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144. https://doi.org/10.1145/3422622

Hippisley-Cox, J., & Coupland, C. (2015). Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: Prospective cohort study. *BMJ Open*, 5(3), e007825. https://doi.org/10.1136/bmjopen-2015-007825

Kannan, S. R., Devi, R., Ramathilagam, S., & Hong, T. P. (2017). Effective fuzzy possibilistic c-means: Analyzing cancer medical database. *Soft Computing*, 21, 2835-2845. https://doi.org/10.1007/s00500-016-2485-2

Khan, M. A., Rubab, S., Kashif, A., Sharif, M. I., Muhammad, N., Shah, J. H., & Satapathy, S. C. (2020). Lungs cancer classification from CT images: An integrated design of contrast-based classical features fusion and selection. *Pattern Recognition Letters*, 129, 77-85. https://doi.org/10.1016/j.patrec.2019.11.012

Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, 92, 374-382. https://doi.org/10.1016/j.future.2018.10.009

Malvezzi, M., Santucci, C., Boffetta, P., Collatuzzo, G., Levi, F., La Vecchia, C., & Negri, E. (2023). European cancer mortality predictions for the year 2023 with focus on lung cancer. *Annals of Oncology*, 34(4), 410-419. https://doi.org/10.1016/j.annonc.2023.01.002

Raza, R., Zulfiqar, F., Khan, M. O., Arif, M., Alvi, A., Iftikhar, M. A., &

Alam, T. (2023). Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. *Engineering Applications of Artificial Intelligence*, 126, 106902. https://doi.org/10.1016/j.engappai.2023.106902

Shahzad, A., Mushtaq, A., Sabeeh, A. Q., Ghadi, Y. Y., Mushtaq, Z., Arif, S., & Jamil, F. (2023, May). Automated uterine fibroids detection in ultrasound images using deep convolutional neural networks. In Healthcare (Vol. 11, No. 10, p. 1493). *MDPI*. https://doi.org/10.3390/healthcare11101493

Tekade, R., & Rajeswari, K. (2018, August). Lung cancer detection and classification using deep learning. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). *IEEE*. https://doi.org/10.1109/ICCUBEA.2018.8697746