Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimizing Cancer Diagnosis: IRPO-Driven Integrated Analysis of Gene Expression Microarray Data

P. Nancy Vincentina Mary[1], Dr. R. Nagarajan[2]

[1]Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, India
[2]Assistant Professor, Department of Computer and Information Science, Faculty of Science, Annamalai University, India
Email: nancy.vincentina.mary@gmail.com

Cancer diagnosis using gene expression microarray data involves analyzing gene expression patterns to classify samples as cancerous or non-cancerous, aiding in early detection and treatment planning for various types of cancer. Challenges in cancer diagnosis from gene expression microarray data include noise and variability in data, feature selection from high-dimensional datasets, overfitting, class imbalance, and the need for robust algorithms to effectively distinguish between cancerous and non-cancerous samples. This work involves a comprehensive approach to analyzing microarray data for cancer diagnosis. It begins with the selection of relevant microarray data, followed by essential data pre-processing steps such as normalization and handling missing values to ensure data quality. Dimensionality reduction techniques, particularly Linear Discriminant Analysis (LDA), are employed to reduce the complexity of the dataset. Feature selection is then performed using the Improved Red Panda Optimization (IRPO) algorithm. Subsequently, a classification model, Convolutional Artificial Neural Networks (CANN), is utilized for accurate cancer diagnosis. This integrated approach ensures that the most relevant features are extracted from the data, optimizing classification performance while mitigating the effects of noise and high dimensionality inherent in microarray datasets, ultimately enhancing the accuracy and reliability of cancer diagnosis. After classifying cancer types, pathway analysis is conducted to understand the molecular characteristics of various cancers. This informs the search for biomarkers, therapeutic targets, and subtype-specific therapies. By using this data, customized treatment plans based on the molecular features of each patient's tumor can be created, leading to better patient outcomes and enabling precision medicine techniques in cancer treatment. The proposed model achieves a high

accuracy rate of approximately 99.758%, demonstrating its effectiveness in cancer diagnosis.
Keywords: Cancer Diagnosis; Gene Analysis; Pathway Analysis; LDA; IRPO; CANN.

1. Introduction

Cancer diagnosis stands at the forefront of modern medical challenges, representing a critical juncture where early detection can significantly alter patient outcomes. This process involves a multifaceted approach that integrates advanced technologies and rigorous clinical methodologies to identify and characterize cancerous cells or tissues within the human body [1,2]. At its essence, cancer diagnosis begins with a suspicion often triggered by symptoms reported by patients or abnormalities detected through routine screening tests. These indications prompt healthcare professionals to initiate a thorough investigation, typically starting with imaging techniques such as X-rays, CT scans, MRI scans, or ultrasound [3,4]. These imaging modalities provide initial insights into the location, size, and possible spread of suspected tumors or abnormal growths, guiding further diagnostic pathways. Following initial imaging, clinicians may proceed to more targeted diagnostic procedures, depending on the suspected type of cancer and the location of abnormalities [5]. Biopsy remains a cornerstone in confirming a cancer diagnosis, involving the extraction and examination of tissue samples from suspicious areas. This procedure, often performed under local anaesthesia, allows pathologists to scrutinize cells microscopically, identifying malignant characteristics such as abnormal cell structure, rapid growth patterns, and potential invasion into surrounding tissues [6,7].

In addition to traditional pathology, molecular diagnostics have revolutionized cancer diagnosis by delving into the genetic and molecular makeup of tumors [8]. Techniques like polymerase chain reaction (PCR) and next-generation sequencing (NGS) enable clinicians to identify specific genetic mutations or biomarkers associated with different types of cancer [9]. This molecular profiling not only aids in confirming diagnoses but also informs personalized treatment strategies tailored to the genetic profile of each patient's cancer. Moreover, advancements in medical imaging and diagnostic technologies have led to the development of non-invasive or minimally invasive diagnostic tools [10,11]. Liquid biopsies, analyze blood samples for circulating tumor cells, cell-free DNA, or other biomarkers shed by tumors into the bloodstream. These biomarkers provide valuable information about the presence of cancer, its molecular characteristics, and even its response to treatment, offering a less invasive alternative to traditional tissue biopsies [12]. Furthermore, the role of artificial intelligence (AI) and machine learning algorithms continues to expand in cancer diagnosis. These technologies analyze vast amounts of medical data, including imaging scans, genetic profiles, and patient histories, to assist radiologists, pathologists, and oncologists in making more accurate and timely diagnostic decisions [13]. AI-driven tools can enhance the sensitivity and specificity of cancer detection, reducing the risk of false positives and miss diagnoses, thereby improving overall patient care and outcomes.

Cancer diagnosis represents a complex and evolving field within modern medicine, integrating a spectrum of clinical, technological, and scientific advancements [14]. From initial suspicions

based on symptoms to confirmatory tests like biopsies and molecular profiling, each step in the diagnostic journey plays a crucial role in guiding treatment decisions and improving patient prognosis. As research continues to push the boundaries of diagnostic accuracy and accessibility, the ongoing refinement of diagnostic tools promises to further enhance our ability to detect cancer earlier, ultimately leading to improved survival rates and quality of life for patients worldwide [15]. The motivation behind utilizing gene expression microarray data for cancer diagnosis lies in its potential to revolutionize early detection and treatment strategies. By analyzing gene expression patterns, this approach aims to identify subtle molecular signatures indicative of cancerous conditions, facilitating earlier interventions and personalized treatment plans tailored to each patient's unique genetic profile. This work addresses critical challenges such as data variability and complexity through advanced computational techniques, aiming to enhance diagnostic accuracy and pave the way for precision medicine applications in oncology. Ultimately, the goal is to improve patient outcomes by leveraging molecular insights to optimize therapeutic strategies and improve overall survival rates.

The contributions of this paper are manifested below,

- By integrating advanced computational techniques such as LDA for dimensionality reduction and IRPO for feature selection, the approach optimizes the identification of relevant gene expression patterns associated with cancer. This leads to improved accuracy in distinguishing between cancerous and non-cancerous samples, crucial for early detection and precise treatment planning.

- This work performed pathway analysis after classifying cancer types to understand the molecular characteristics of various cancers. The analysis of molecular characteristics through microarray data facilitates the discovery of biomarkers and therapeutic targets specific to different cancer types. This enables the development of personalized treatment plans based on the genetic profile of each patient's tumor, potentially improving treatment efficacy and patient outcomes through tailored therapies.

- This work addresses key challenges in microarray data analysis, such as noise reduction, handling high-dimensional datasets, and mitigating overfitting and class imbalance issues. The application of CANN as a classification model further enhances the robustness and reliability of cancer diagnosis, paving the way for more effective utilization of gene expression data in clinical settings.

The rest of this paper is organized as follows. The section II provides both related works and problem statement. The proposed protocol is introduced and explained in the section III. The result and discussion are then presented in the section IV, followed by the conclusion in the section V.

2. Literature Review

In 2021, Bartha and Győrffy [16] developed an integrated database and web platform to mine this data in real time, using gene array data from NCBI-GEO and RNA-seq data from TCGA, TARGET, and GTEx. The database includes 56,938 samples from various sources. Key

upregulated genes were TOP2A, SPP1, and CENPA, while ADH1B was downregulated.

In 2019, Ghosh et al. [17] used a Recursive Memetic Algorithm (RMA) for gene selection, outperforming standard Memetic Algorithms (MA) and Genetic Algorithms (GA). Applied to seven microarray datasets (AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL, and SRBCT), RMA achieved higher accuracy with fewer features. The results, validated biologically using Gene Ontology, KEGG pathways, and heat maps, demonstrate the effectiveness of our approach.

In 2020, Yuan et al. [18] proposed machine learning to analyze gene expression profiles of lung AC and SCC from the Gene Expression Omnibus. Monte Carlo feature selection ranked features by importance, and the incremental feature selection method identified optimal features for SVM classification. Key genes (e.g., CSTA, TP63, SERPINB13) were identified as differentially expressed. Additionally, rule learning provided classification rules, highlighting distinct gene expression patterns between lung AC and SCC.

In 2022, Su et al. [19] used gene expression data from The Cancer Genome Atlas (TCGA) for diagnosing and staging colon cancer. Weighted Gene Co-expression Network Analysis (WGCNA) identified key gene modules, and the Lasso algorithm extracted characteristic genes. Random Forest (RF), SVM, and decision trees were used for diagnosis, with RF achieving the best results: 99.81% accuracy for diagnosis, 91.5% for staging.

In 2019, Guan et al. [20] investigated circRNA regulatory mechanisms in GC and analyzed circRNA expression profiles from four GEO microarray datasets and miRNA/mRNA profiles from the TCGA database. Differentially expressed circRNAs (DEcircRNAs) were identified using robust rank aggregation, and a ceRNA network was constructed. Functional and pathway enrichment analyses were performed, and protein interactions predicted using Cytoscape. A subnetwork regulatory module was developed with the MCODE plugin.

In 2018, Shukla et al. [21] developed a hybrid gene selection method to enhance classification accuracy and reduce computational time. Our two-stage method first applies the EGS method with a multi-layer and f-score approach to filter noisy and redundant genes. In the second stage, an adaptive genetic algorithm (AGA) identifies significant gene subsets using SVM and Naïve Bayes classifiers.

In 2021, Liu et al. [22] conducted a comprehensive bioinformatics analysis across multiple databases to assess Keap1 mRNA's diagnostic and prognostic significance in lung cancer. ROC curve analysis indicated strong diagnostic potential for lung squamous cell carcinoma (LUSC). High Keap1 mRNA levels emerged as an independent risk factor for overall lung cancer mortality but exhibited conflicting implications for lung adenocarcinoma (LUAD). Co-expressed genes with Keap1 and Nfe2L2 were identified, highlighting their involvement in the oxidative stress-induced gene expression pathway via Nrf2, implicating these mechanisms in lung cancer pathogenesis.

In 2019, Musheer et al. [23] introduced an ABC-based feature selection method for microarray data. Our approach combines ICA for data reduction and ABC for optimizing feature vectors. Extensive experiments validate our method, showing it outperforms existing approaches in gene selection for the Naïve Bayes classifier across multiple cancer classification datasets, as confirmed by statistical hypothesis testing.

In 2020, Millstein et al. [24] aimed to establish a robust prognostic signature for overall survival (OS) in women with high-grade serous ovarian cancer (HGSOC). Expression levels of 513 genes, identified from a meta-analysis of 1455 tumors and additional candidates, were assessed using NanoString technology on formalin-fixed paraffin-embedded tumor samples from 3769 patients. Elastic net regularization was employed for survival analysis, developing a predictive model for 5-year OS. The model was trained on 2702 tumors from 15 studies and validated on an independent cohort of 1067 tumors from six studies.

In 2019, Algamal and Lee [25] proposed a two-stage approach. The first stage employs sure independence screening to identify genes highly correlated with cancer class levels. In the second stage, adaptive lasso with new weights handles correlations among these genes. Experimental results across four gene expression datasets demonstrate superior performance in classification metrics and highlight biologically relevant genes, making it a promising method for clinical cancer classification.

2.1. Problem Statement

The problem statement for cancer diagnosis using gene expression microarray data revolves around the need to effectively utilize complex biological data to improve early detection and treatment planning. Gene expression microarray technology offers a wealth of information about cellular activities and molecular profiles associated with cancer. However, challenges such as data noise, high dimensionality, and variability across samples hinder accurate classification of cancerous and non-cancerous tissues. Current methodologies must navigate these obstacles to ensure robust algorithms capable of distinguishing subtle gene expression patterns indicative of different cancer types. Moreover, the identification of relevant biomarkers and therapeutic targets from these datasets requires sophisticated data pre-processing, feature selection, and classification techniques. Addressing these challenges is crucial for enhancing diagnostic accuracy, enabling personalized treatment strategies, and advancing the field of precision medicine in oncology. Therefore, the overarching goal is to develop computational models that optimize classification performance and translate molecular insights into actionable clinical outcomes for cancer patients.

3. Proposed Methodology

Cancer diagnosis using gene expression microarray data involves examining gene activity patterns to detect cancer early and devise effective treatment plans. This process faces challenges such as data noise, high dimensionality, and variability across samples. To address these issues, several advanced techniques are employed: pre-processing to clean the data, dimensionality reduction to simplify the dataset, feature selection to identify the most relevant genes, and classification algorithms to accurately categorize cancer types. These methods enhance diagnostic accuracy and reliability by mitigating noise and preventing overfitting. Robust algorithms are essential for handling the complexity of the data and ensuring reproducible results. Ultimately, these techniques facilitate early cancer detection and enable personalized treatment strategies for patients. Figure 1 illustrates the overall proposed architecture.

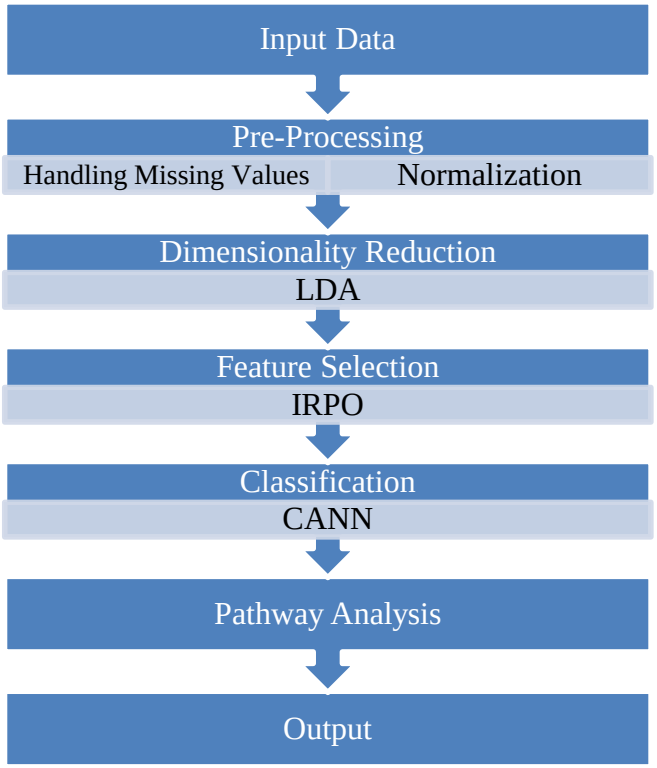


Figure 1: Overall Proposed Architecture

3.1 Pre-processing

The pre-processing phase involves normalizing data to ensure consistent scaling and employing techniques to handle missing data, thereby addressing any gaps or null values in the dataset.

3.1.1 Normalization

Normalization is an essential step in data pre-processing that standardizes the scale of features within a dataset. This process ensures that all variables contribute equally to the analysis, regardless of their original units or scales. Normalization involves transforming numerical values to a common scale, typically between 0 and 1 or -1 and 1. One common technique for normalization is Min-Max scaling, which adjusts each feature's values proportionally to fit within a specified range. This is done by subtracting the minimum value from each observation and then dividing by the range (the difference between the maximum and minimum values). Another approach is Z-score normalization (standardization), which involves subtracting the mean from each observation and dividing by the standard deviation. This centres the data around zero and adjusts it to have a standard deviation of 1. By normalizing data, features with larger scales do not overshadow those with smaller scales during analysis, thereby enhancing the performance and convergence of algorithms.

3.1.2 Handling Missing Data

Handling missing data is a crucial part of data pre-processing, aimed at managing the absence of values in a dataset. Missing data can occur due to various reasons, such as measurement errors, data corruption, or intentional non-response. Ignoring missing values can lead to biased analyses and inaccurate results. Several strategies exist for handling missing data, including deletion, imputation, and prediction. Deletion involves removing observations or variables with missing values, which can result in the loss of valuable information and a reduced sample size. Imputation methods involve replacing missing values with estimated ones based on statistical measures like mean, median, or mode, though this can introduce bias and alter the data's distribution. Prediction methods utilize machine learning algorithms to predict missing values based on other variables in the dataset, offering a more sophisticated approach to handling missing data.

3.2. Dimensionality Reduction

Dimensionality reduction streamlines modelling by reducing the number of variables in a dataset. It includes feature selection, which involves choosing the most significant variables, and feature extraction, which transforms high-dimensional data into fewer dimensions. This process accelerates model training and improves accuracy by mitigating overfitting. In this study, LDA was used for dimensionality reduction. LDA identifies the linear combinations of features that best separate different classes, enhancing the discriminatory power of the model. By focusing on the most informative features and reducing data complexity, LDA helps in building more efficient and accurate predictive models.

3.2.1. LDA

LDA is a technique used in statistics, pattern recognition, and machine learning to find a linear combination of features that best represents a dependent variable. Unlike Principal Component Analysis (PCA) and factor analysis, which focus on similarities, LDA explicitly models differences between data classes. It identifies vectors in the data space that best discriminate between classes, aiming to maximize the separation between multiple classes. LDA works by finding a linear combination of independent features to maximize the mean differences between classes. This is mathematically expressed in terms of two scatter matrices as per Eq. (1) and Eq. (2).

$$sw1 = \sum_{j=1}^{c} \sum_{i=1}^{nj} \left(x_i^j - \mu_j\right)\left(x_i^j - \mu_j\right)^w \quad (1)$$

x_i^j denotes the ith sample of class, μ_j is the mean of class j, c represents the number of classes, nj signifies the number of samples in class j, μ denotes the mean of all classes.

$$sw1 = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^v \quad (2)$$

LDA aims to maximize the ratio of the between-class scatter to the within-class scatter, effectively increasing the separation between different classes while reducing the spread within each class. By doing so, LDA enhances the discriminatory power of the resulting linear combinations, making it a powerful tool for classification tasks.

3.3 Feature Selection

In this study, the Improved Red Panda Optimization (IRPO) algorithm is employed to enhance classification accuracy by selecting relevant features and refining the collected data, ultimately improving model performance.

3.3.1 IRPO

The red panda, native to southern China and the eastern Himalayas, is a small mammal known for its reddish-brown fur and distinctive markings. Thriving in temperate forests with dense bamboo cover, it excels in climbing trees. Feeding mainly on bamboo leaves and shoots, it relies on keen senses and climbing abilities. The Red Panda Optimization (RPO) algorithm's design is inspired by these natural characteristics.

3.3.1.1 Mathematical Modelling

3.3.1.1.1 Initialization

As a population-based metaheuristic algorithm, the RPO technique uses red pandas to symbolize each individual member. Each red panda represents a candidate solution in the search space. The positions of these red pandas are initialized randomly to explore the search space effectively. The red panda's position is represented mathematically as a vector, forming a population matrix Y. This matrix is initialized using Eq. (3) and Eq. (4):

$$X = \begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_i \\ . \\ . \\ X_M \end{bmatrix}_{M\times n} = \begin{bmatrix} X_{1,1} \dots X_{1,j} \dots X_{1,n} \\ . \\ . \\ . \\ X_{i,1} \dots X_{i,j} \dots X_{i,n} \\ . \\ . \\ X_{M,1} \dots X_{M,j} \dots X_{M,n} \end{bmatrix}_{M\times n} \tag{3}$$

$$x_{i,j} = lob_j + r_{i,j} \cdot \left(upb_j - lob_j\right), i = 1,2, \dots, M, j = 1,2, \dots, n \tag{4}$$

Here, X_i stands for the i-th red panda (candidate solution) and $X_{i,j}$ for its j-th dimension problem variable). M is the total number of red pandas, and n is the number of problem variables. The j-th problem variable's lower and upper limits are denoted by lob_j and upb_j, respectively, and $r_{i,j}$ are random values in the interval [0,1]. The objective function values of the initialized solutions are evaluated and represented as per Eq. (5).

$$f = \begin{bmatrix} f_1 \\ . \\ . \\ . \\ f_i \\ . \\ . \\ f_M \end{bmatrix}_{M\times 1} = \begin{bmatrix} f(X_1) \\ . \\ . \\ . \\ f(X_i) \\ . \\ . \\ f(X_M) \end{bmatrix}_{M\times 1} \tag{5}$$

Where, f_i is the objective function value for the i-th red panda. These values help evaluate the quality of potential solutions, identifying the best and worst solutions to guide the optimization process.

3.3.1.1.2. Phase 1: Exploration Strategy - Foraging

In the first phase of RPO, red pandas' positions mimic their natural foraging behavior. They use their keen senses to move towards food sources. In the algorithm, each red panda considers the locations of others that yield superior objective function values as potential food sources. These proposed food positions are determined based on objective function value comparisons, with one position randomly chosen by each red panda using Eq. (6):

$$pfs_i = \{X_k | k \in \{1,2,..,M\} \text{and } f_k < f_i\} \cup \{X_{best}\} \quad (6)$$

Based on a comparison with the location of the best candidate solution Y_{best}, the suggested food sources for each red panda pfs_i are identified. Approaching these sources causes large positional shifts that improve ability of algorithm to globally search and explore. By determining new locations in relation to the food source (best candidate solution), red pandas' foraging behaviour can be replicated. Eq. (7) to Eq. (9) are used to update the red panda's location if the objective function value improves at the new location.

$$X_i^{p1}: x_{i,j}^{p1} = x_{i,j} + r.\left(sfs_{i,j} - Is.y_{i,j}\right) + x_i' \quad (7)$$

$$x_i' = X_i + \Delta X_i \quad (8)$$

$$X_i = \begin{cases} X_i^{p1}, f_i^{p1} < f_i \\ X_i, else \end{cases} \quad (9)$$

Gaussian mutation balances exploration and exploitation by adjusting the standard deviation. It's simple to implement and adaptable, with mutation strength decreasing over time to enhance convergence in optimization algorithms. The new location of the ith red panda as ascertained from the RPO's first phase is represented by X_i^{p1}. Objective function is denoted by f_i^{p1}, and its position in the jth dimension is indicated by $x_{i,j}^{p1}$. For the ith red panda, sfs_i denotes the preferred food source, and $sfs_{i,j}$ denotes its location in the jth dimension. Is is a randomly chosen number from the set {1, 2}, and the variable r is a random value between 0 and 1.

3.3.1.1.3 Phase 2: Proficiency in Ascending and Perching on Trees (Exploitation)

In the second phase of the RPO, red pandas' ability to climb and rest on trees is modelled. Red pandas spend much of their time on trees, climbing to obtain food after foraging on the ground. This behavior results in minor positional changes, improving the exploitation and local search capabilities of the RPO algorithm in promising areas. The tree-climbing behavior is mathematically modelled to calculate new positions for each red panda and replace previous positions if the objective function improves, as represented by Eq. (10) and Eq. (11):

$$X_{i,j}^{p2} = x_{i,j} + \frac{lob_j + r_{i,j} \cdot (upb_j - lob_j)}{t}, i = 1,2,\ldots,M, j = 1,2,\ldots,n, t = 1,2,\ldots,T \quad (10)$$

$$X_i = \begin{cases} X_i^{p2}, f_i^{p2} < f_i \\ X_i, else \end{cases}$$

(11)

The ith red panda's modified position, obtained from the second phase of RPO, is represented by X_i^{p2}. Objective function is shown by f_i^{p2}, and its position in the jth dimension is indicated by $X_{i,j}^{p2}$. A random number between 0 and 1 represents the variable r. The symbol t denotes the algorithm's iteration counter, whereas T stands for the maximum iterations. This phase refines the red pandas' positions, enhancing the algorithm's ability to exploit local optima and converge to the best solution.

3.4. Classification - CANN

One class of deep neural networks that is mainly utilized for the analysis of visual vision is called CNN. They excel at tasks like image identification and classification because of their organized ability to adaptively and automatically extract spatial hierarchies of characteristics from incoming data. Convolutional layers, pooling layers, fully linked layers, activation functions, and normalizing layers are the essential parts of a CNN. Learnable filters or kernels are used by convolutional layers to apply convolution operations, which convolve over the input image to extract features. These filters identify particular patterns, like textures, edges, or intricate structures.

Convolutional features spatial dimensions can be decreased while maintaining crucial information by using pooling layers. The pooling operations max-pooling and average-pooling are often used. The network can learn intricate correlations in the data by introducing non-linearity to its output with activation functions such as ReLU. By dividing the input into several classes according to the features that convolutional layers extracted, fully connected layers carry out high-level reasoning. Normalization layers, such as Batch Normalization, normalize the input to a layer, improving stability and speeding up training.

Convolutional Layer: This layer applies convolution operations to the input data using learnable filters. The output of each filter, known as a feature map, captures specific patterns or features from the input. The convolution operation is represented as per Eq. (12).

$$\mathrm{Conv}(i,j) = \sum_{m=0}^{m-1} \sum_{n=0}^{n-1} I(i+m, j+n) \times K(m,n) \quad (12)$$

Where I is the input matrix, K is the filter/kernel, and m and n are the dimensions of the filter.

Rectified Linear Unit (ReLU): ReLU is an activation function that introduces non-linearity to the network by replacing negative values with zero. It is defined as per Eq. (13).

$$\mathrm{relu}(x) = \max(0, x) \quad (13)$$

Pooling Layer: By lowering spatial dimensions, layer down samples the feature maps that were acquired from convolutional layers. A common pooling operation is max pooling, where the maximum value within each pooling window is retained. It helps in reducing computational complexity and controlling overfitting. Fig. 2 depicts the CANN architecture.

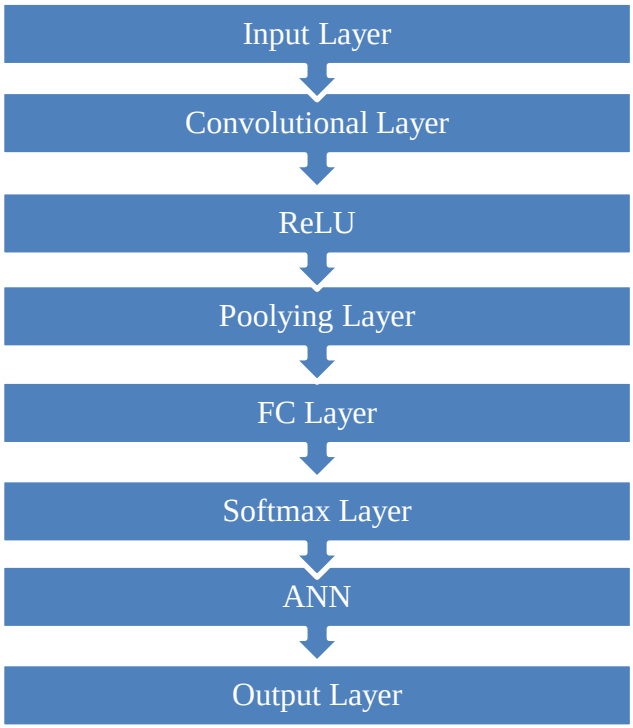


Figure 2: CANN

Fully Connected (FC) Layer: Also known as the dense layer, this layer connects every neuron from the previous layer to every neuron in the current layer. It learns complex patterns by combining features extracted from previous layers.

Softmax Layer: This layer is typically used as the output layer in classification tasks. It converts the raw scores (logits) from the previous layer into probabilities for each class using the Softmax function using Eq. (14).

$$SOp_x \frac{e^{z_x}}{\sum_{x=1}^{M} e^{z_x}} \quad (14)$$

Where, z is the vector of logits, z_x represents output-count, SOp_x denotes softmax output, and M represents the totality of output nodes.

The architecture of an Artificial Neural Network (ANN) typically consists of three types of layers: the input layer, hidden layers, and the output layer. This architecture is designed to handle complex data, improve accuracy, and make reliable predictions.

- Input Layer

The input layer receives raw data or features and forwards them to the hidden layers for processing. It contains one neuron per feature, with no computations performed within this layer.

- Hidden Layers

Hidden layers perform the primary computations in neural networks. Each neuron in a hidden layer receives inputs from the preceding layer, computes a weighted sum, and applies an activation function to generate an output, as represented by Eq. (15):

$$a_{ij} = f\left(\sum_{k=1}^{n_{i-1}} w_{ik} a_{ik} + b_{ij}\right) \quad (15)$$

Here, a_{ij} is the activation of the j-th neuron in the i-th layer, w_{ik} are the weights, a_{ik} are the activations from the previous layer, and $b_{ij}$is the bias term.

- Output Layer

The output layer generates the final output of the neural network. The number of neurons in this layer depends on the type of problem being addressed. After classifying cancer types, pathway analysis is conducted to delve into the molecular characteristics of different cancers. This detailed examination helps identify biomarkers, therapeutic targets, and subtype-specific therapies. Utilizing this data allows for the creation of customized treatment plans changed to the molecular features of each patient's tumor. These personalized plans enhance patient outcomes by targeting the specific pathways and mechanisms involved in their cancer. By adopting precision medicine techniques, this approach significantly improves the effectiveness of cancer treatments, leading to more successful management and potential cures for patients with diverse cancer types.

3.5. Pathway Analysis

In this work, pathway analysis is a critical step that follows the feature selection process. After pre-processing the microarray data and reducing its dimensionality using LDA, IRPO algorithm selects the most relevant features (genes). This step ensures that only the most significant genes, which are likely to be involved in cancer progression, are retained. Once the relevant genes are identified, pathway analysis is performed to understand how these genes interact within biological pathways. Pathway analysis involves mapping these selected genes onto known biological pathways to identify which pathways are enriched or deregulated in cancerous samples compared to non-cancerous samples. The steps in pathway analysis are,

a. Pathway Mapping

The selected genes are mapped to predefined biological pathways from databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, or BioCarta. This mapping helps in identifying the pathways in which these genes play a role.

b. Enrichment Analysis

Enrichment analysis is then conducted to determine if the identified genes are significantly overrepresented in specific pathways compared to what would be expected by chance. Statistical methods such as Fisher's exact test, hypergeometric test, or Gene Set Enrichment Analysis (GSEA) are typically used. The hypergeometric test is commonly used to determine if a set of selected genes is overrepresented in a particular pathway. The test calculates the probability of observing a certain number of selected genes in a pathway by chance. Fisher's exact test is used to calculate the exact probability of the observed association between the

selected genes and the pathway. GSEA calculates an enrichment score to determine if the members of a gene set are randomly distributed throughout the ranked list of genes or primarily found at the top or bottom.

c. Quantifying Pathway Deregulation

To measure the extent of pathway deregulation, algorithms such as Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) or Pathifier can be used. These tools score pathways based on their activity levels and deviations from normal behavior in individual samples.

Pathway analysis provides insights into the molecular mechanisms driving cancer progression. By identifying which pathways are deregulated, researchers can better understand the biological processes involved in tumor growth, metastasis, and resistance to therapy. Knowing which pathways are altered in cancer can help identify potential therapeutic targets. Drugs can be designed or repurposed to specifically inhibit or modulate these pathways, leading to more effective treatments. By integrating pathway analysis with gene expression data, it is possible to create customized treatment plans based on the specific pathways that are active in a patient's tumor. This personalized approach can improve treatment efficacy and patient outcomes. Pathways that are consistently deregulated across different patients can serve as biomarkers for diagnosis, prognosis, or treatment response. These biomarkers can be used in clinical settings to identify patients who are likely to benefit from specific therapies.

Pathway analysis in this work enhances the overall understanding of cancer biology by linking gene expression data to biological pathways. It provides a deeper insight into the molecular underpinnings of cancer, enabling the development of targeted therapies and personalized medicine approaches. By focusing on pathway-level changes, researchers can uncover the complex interactions and regulatory mechanisms that drive cancer progression, ultimately leading to improved diagnostic and therapeutic strategies. Algorithm 1 illustrates the process of cancer diagnosis using IRPO for feature selection and CANN for accurate classification.

Algorithm 1: IRPO-CANN
BEGIN
INPUT: Gene expression microarray data (raw data)
Pre-processing
Normalize the data
FOR each feature in the dataset
Apply Min-Max Scaling or Z-score normalization to standardize values
Handle missing data
IF missing data exists
Impute missing values using mean, median, or prediction methods
Dimensionality Reduction
Apply LDA

Compute within-class scatter matrix and between-class scatter matrix
Maximize the ratio of between-class scatter to within-class scatter
Reduce dimensions by projecting data onto the LDA components
Feature Selection
Initialize IRPO (Improved Red Panda Optimization) Algorithm
Initialize population of red pandas (candidate solutions)
Evaluate fitness of each candidate (based on selected features)
Exploration Phase
FOR each red panda in the population
Select potential food source (best candidate solution)
Update red panda position using foraging behavior formula
Exploitation Phase
FOR each red panda in the population
Adjust position based on tree-climbing behavior to refine local search
Update red panda position if objective function improves
Select top features based on best red panda's position
Classification
Build CANN
Train CANN
Backpropagate errors and adjust weights using optimization algorithm
Continue training until convergence or max iterations reached
Pathway Analysis
OUTPUT: Cancer classification results and pathway analysis insights
END

4. Result and Discussion

4.1. Experimental Setup

The proposed model is implemented using the Python platform and benchmarked against existing models like Improved Red Panda Optimization Recurrent Neural Networks (IRPO-RNN), Convolutional Neural Networks (IRPO-CNN), and Artificial Neural Networks (IRPO-ANN). Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate its effectiveness in cancer diagnosis. By comparing these metrics, the proposed

model's superiority over established methods can be determined. Additionally, pathway analysis is integrated into the workflow after feature selection to identify enriched biological pathways. This step involves mapping the selected genes to known pathways and performing enrichment analysis to identify overrepresented pathways. Pathway deregulation scores are computed for each sample to quantify the extent of pathway deregulation, providing deeper insights into the molecular mechanisms driving cancer progression. This comparative analysis, including pathway analysis, provides valuable insights into the model's accuracy and reliability, demonstrating its potential to improve early detection and personalized treatment strategies in cancer diagnosis.

4.2. Dataset Collection

The Ovarian Cancer dataset [26] is an extensive genomic data collection created for ovarian cancer research. It consists of 253 samples (genes/features) and 162 classes of people (162 with ovarian cancer diagnosis and 91 healthy controls). The WCX2 protein chip technology was utilized to gather continuous numeric data that represents the levels of gene expression in the dataset. Important new understandings of the molecular processes behind ovarian cancer are made possible by this comprehensive genomic data. It is particularly helpful in the development of prediction models that aid in illness diagnosis and in the identification of potential biomarkers for targeted treatments. Using this dataset, researchers may investigate the genetic makeup of ovarian cancer, leading to more accurate diagnoses and more individualized treatment plans for patients.

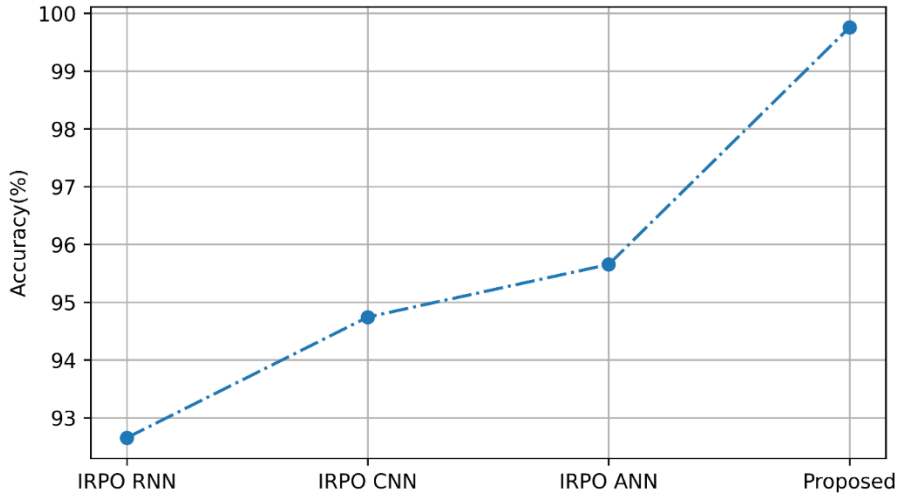
4.3. Overall Performance Analysis

Table 1 presents a comparative performance analysis of existing models and the proposed model for cancer diagnosis using gene expression data. The models compared include (IRPO-RNN), (IRPO-CNN), (IRPO-ANN), and the newly proposed model. Starting with the IRPO-RNN model, it achieved an accuracy of 92.654%, precision of 94.954%, recall of 94.765%, and an F1 score of 93.654%. This indicates that while the IRPO-RNN performs reasonably well, there is still room for improvement in accurately identifying cancerous samples. The IRPO-CNN model shows an improved performance over the IRPO-RNN, with an accuracy of 94.743%, precision of 95.654%, recall of 95.756%, and an F1 score of 95.765%. These metrics suggest that IRPO-CNN is more effective in distinguishing between cancerous and non-cancerous samples. The IRPO-ANN model further enhances performance with an accuracy of 95.654%, precision of 96.654%, recall of 97.954%, and an F1 score of 97.964%. This indicates a high level of precision and recall, making it a strong contender for cancer diagnosis. The proposed model, however, surpasses all existing models with remarkable metrics: an accuracy of 99.758%, precision of 99.999%, recall of 99.789%, and an F1 score of 99.879%. These results demonstrate the proposed model’s superior performance in accurately diagnosing cancer. The integration of pathway analysis post-feature selection significantly contributes to its efficacy, allowing for the identification of enriched biological pathways and providing deeper insights into the molecular mechanisms driving cancer progression. This comprehensive approach ensures that the proposed model not only excels in performance metrics but also enhances the understanding of cancer biology, paving the way for improved early detection and personalized treatment strategies.

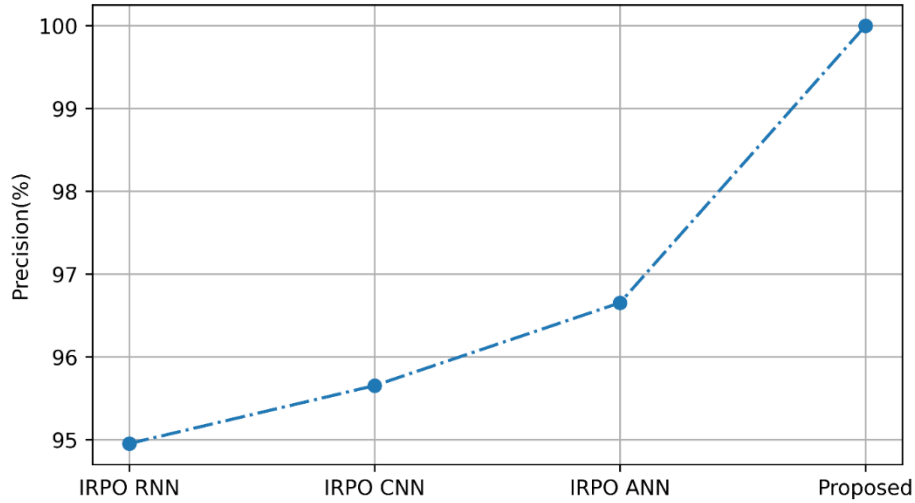
Table 1: Performance Analysis of Existing and Proposed Model

Methods	Accuracy	Precision	Recall	F1 Score
IRPO RNN	92.654	94.954	94.765	93.654
IRPO CNN	94.743	95.654	95.756	95.765
IRPO ANN	95.654	96.654	97.954	97.964
Proposed	99.758	99.999	99.789	99.879

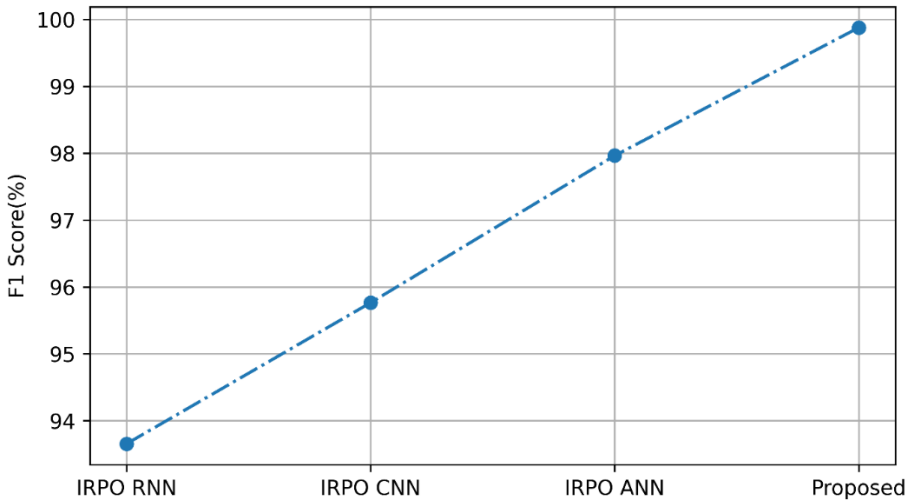
4.4. Graphical Representation



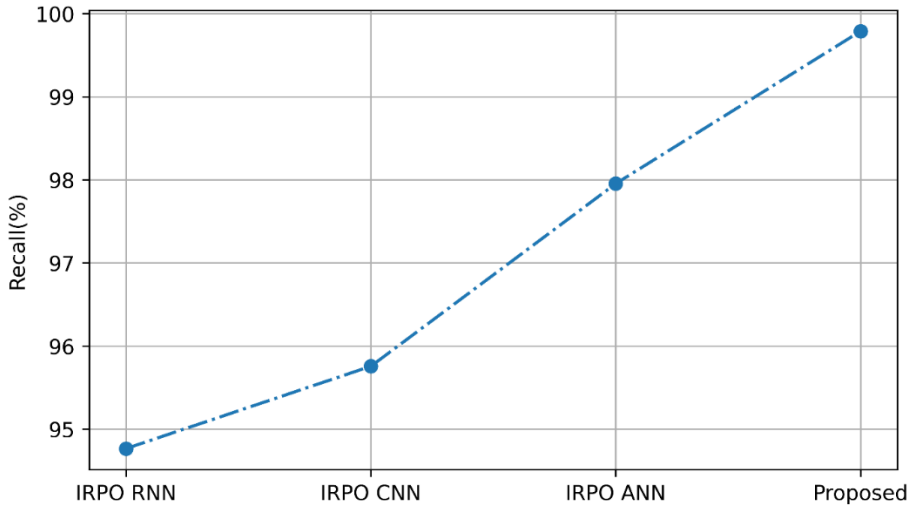
(a)



(b)



(c)



(d)

Figure 3: Graphical Representation of Existing and Proposed Model (a) Accuracy (b) Precision (c) F1-Score (d) Recall

Fig. 3 visually compares the performance metrics such as, accuracy, precision, F1 score, and recall of the proposed model and existing models (IRPO-RNN, IRPO-CNN, IRPO-ANN) for cancer diagnosis using gene expression data. Each subplot (a) to (d) shows a bar chart representing the respective metric for each model. The proposed model consistently demonstrates superior performance across all metrics compared to IRPO-RNN, IRPO-CNN, and IRPO-ANN, highlighting its effectiveness in accurately classifying cancerous and non-cancerous samples. This graphical representation underscores the significant advancement in accuracy, precision, and overall diagnostic capability achieved by the proposed model,

emphasizing its potential for enhancing early cancer detection and treatment planning.

5. Conclusion

In order to diagnose cancer, this paper used a thorough method of microarray data analysis. First, pertinent microarray data were chosen. Next, necessary data pre-processing procedures including normalization and handling missing values were carried out to guarantee the quality of the data. The dataset's complexity was decreased by using dimensionality reduction techniques, especially LDA. The IRPO algorithm was then used for feature selection. CANN were then applied as a classification model to accurately diagnose cancer. The accuracy and dependability of cancer detection were eventually improved by this integrated strategy, which made sure that the most pertinent features were extracted from the data, optimizing classification performance while reducing the impacts of noise and high dimensionality inherent in microarray datasets. Following the classification of cancer types, pathway analysis was carried out to comprehend the molecular traits of distinct cancer types. This provided guidance in the hunt for therapeutic targets, biomarkers, and subtype-specific treatments. Precision medicine techniques in cancer therapy were made possible by the creation of personalized treatment plans using this data, which were based on the molecular characteristics of each patient's tumor and improved patient outcomes. The efficacy of the suggested approach in diagnosing cancer is demonstrated by its high accuracy rate of almost 99.758%.

References

1. Adiwijaya, W.U., Lisnawati, E., Aditsania, A. and Kusumo, D.S., 2018. Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. Journal of Computer Science, 14(11), pp.1521-1530.
2. Dwivedi, A.K., 2018. Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 29, pp.1545-1554.
3. Ghosh, M., Adhikary, S., Ghosh, K.K., Sardar, A., Begum, S. and Sarkar, R., 2019. Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & biological engineering & computing, 57, pp.159-176.
4. Ayyad, S.M., Saleh, A.I. and Labib, L.M., 2019. Gene expression cancer classification using modified K-Nearest Neighbors technique. Biosystems, 176, pp.41-51.
5. Sayed, S., Nassef, M., Badr, A. and Farag, I., 2019. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications, 121, pp.233-243.
6. Zeebaree, D.Q., Haron, H. and Abdulazeez, A.M., 2018, October. Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE.
7. AbdElNabi, M.L.R., Wajeeh Jasim, M., El-Bakry, H.M., Hamed N. Taha, M. and Khalifa, N.E.M., 2020. Breast and colon cancer classification from gene expression profiles using data mining techniques. Symmetry, 12(3), p.408.
8. Alanni, R., Hou, J., Azzawi, H. and Xiang, Y., 2019. A novel gene selection algorithm for cancer classification using microarray datasets. BMC medical genomics, 12, pp.1-12.
9. Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M., Suri, H.S., Biswas, M., El-Baz, A., Bangeas, P., Tsoulfas, G. and Suri, J.S., 2019. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning

paradigms. Computer methods and programs in biomedicine, 176, pp.173-193.
10. Shukla, A.K., Singh, P. and Vardhan, M., 2018. A two-stage gene selection method for biomarker discovery from microarray data for cancer classification. Chemometrics and Intelligent Laboratory Systems, 183, pp.47-58.
11. Sampathkumar, A., Rastogi, R., Arukonda, S., Shankar, A., Kautish, S. and Sivaram, M., 2020. An efficient hybrid methodology for detection of cancer-causing gene using CSC for micro array data. Journal of Ambient Intelligence and Humanized Computing, 11, pp.4743-4751.
12. Chen, L., Lu, D., Sun, K., Xu, Y., Hu, P., Li, X. and Xu, F., 2019. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. Gene, 692, pp.119-125.
13. Turgut, S., Dağtekin, M. and Ensari, T., 2018, April. Microarray breast cancer data classification using machine learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-3). IEEE.
14. Kilicarslan, S., Adem, K. and Celik, M., 2020. Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. Medical hypotheses, 137, p.109577.
15. Zhou, X., Lu, Z., Wang, T., Huang, Z., Zhu, W. and Miao, Y.I., 2018. Plasma miRNAs in diagnosis and prognosis of pancreatic cancer: A miRNA expression analysis. Gene, 673, pp.181-193.
16. Bartha, Á. and Győrffy, B., 2021. TNMplot. com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues. International journal of molecular sciences, 22(5), p.2622.
17. Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D. and Maulik, U., 2019. Recursive memetic algorithm for gene selection in microarray data. Expert Systems with Applications, 116, pp.172-185.
18. Yuan, F., Lu, L. and Zou, Q., 2020. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1866(8), p.165822.
19. Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., Jia, D., Li, H. and Lv, X., 2022. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Computers in biology and medicine, 145, p.105409.
20. Guan, Y.J., Ma, J.Y. and Song, W., 2019. Identification of circRNA–miRNA–mRNA regulatory network in gastric cancer by analysis of microarray data. Cancer cell international, 19, pp.1-9.
21. Shukla, A.K., Singh, P. and Vardhan, M., 2018. A hybrid gene selection method for microarray recognition. Biocybernetics and Biomedical Engineering, 38(4), pp.975-991.
22. Liu, G.Y., Zhang, W., Chen, X.C., Wu, W.J. and Wan, S.Q., 2021. Diagnostic and Prognostic Significance of Keap1 mRNA Expression for Lung Cancer Based on Microarray and Clinical Information from Oncomine Database. Current Medical Science, 41, pp.597-609.
23. Musheer, R.A., Verma, C.K. and Srivastava, N., 2019. Novel machine learning approach for classification of high-dimensional microarray data. Soft Computing, 23, pp.13409-13421.
24. Millstein, J., Budden, T., Goode, E.L., Anglesio, M.S., Talhouk, A., Intermaggio, M.P., Leong, H.S., Chen, S., Elatre, W., Gilks, B. and Nazeran, T., 2020. Prognostic gene expression signature for high-grade serous ovarian cancer. Annals of Oncology, 31(9), pp.1240-1250.
25. Algamal, Z.Y. and Lee, M.H., 2019. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in data analysis and classification, 13(3), pp.753-771.
26. Dataset taken from: “https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer”, dated 1/7/2024.