Integrating HRPGW optimization with ALSTM for enhanced cancer diagnosis in gene expression microarray data analysis

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

[1]Department of MCA, Fatima College, Madurai, Tamil Nadu, India & Research Scholar, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India
[2]Assistant Professor, Department of Computer and Information Science, Faculty of Science, Annamalai University, Chidambaram, Tamil Nadu, India
*Corresponding Author

Received: 12.07.2024 Revised: 15.08.2024 Accepted: 20.09.2024

ABSTRACT
Cancer diagnosis using gene expression microarray data involves analysing 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 analysing 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 Principal Component Analysis (PCA), are employed to reduce the complexity of the dataset. Feature selection is then performed using the Hybrid Red Piranha Grey Wolf (HRPGW) algorithm, which combines the optimization power of Grey Wolf Optimization with the exploration capability of Red Piranha. Subsequently, classification models Artificial Long Short-Term Memory (ALSTM), including Artificial Neural Networks (ANN) and Long Short-Term Memory Networks (LSTM), are 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.

Keywords: Cancer Diagnosis, Gene Expression Microarray, PCA, HRPGW, ALSTM, ANN, Deep Learning.

1. INTRODUCTION

Cancer diagnosis using gene expression data obtained from microarray technology represents a significant advancement in oncology, offering insights into the molecular mechanisms underlying cancer development and progression. This approach has revolutionized the field by enabling the identification of biomarkers and molecular signatures that can aid in early detection, classification of cancer types, and personalized treatment strategies [1, 2]. Gene expression microarrays allow researchers and clinicians to simultaneously measure the expression levels of thousands of genes within a biological sample. These data provide a snapshot of the genetic activity within cells, offering clues about their physiological state and potential abnormalities such as cancerous growth [3, 4]. The technology has been particularly instrumental in understanding the heterogeneity of cancer, where each type and subtype may exhibit distinct genetic profiles that influence disease progression and response to treatment [5, 6].

Early detection of cancer is crucial for improving patient outcomes, as treatment efficacy is often higher when the disease is detected in its initial stages. Gene expression profiling facilitates the identification of biomarkers that can serve as early indicators of cancer development [7]. By comparing gene expression patterns between cancerous and normal tissues, researchers can pinpoint genes that are either overexpressed or under expressed in cancer cells. These biomarkers not only aid in diagnosis but also provide insights into the molecular pathways driving cancer, which can guide the development of targeted therapies [8, 9].

Moreover, gene expression data are invaluable for classifying different types of cancer and predicting their clinical behavior. Cancer is not a single disease but rather a collection of diseases characterized by uncontrolled cell growth and proliferation. Each type of cancer can have multiple subtypes with varying genetic profiles and clinical outcomes [10]. Gene expression profiling allows for the categorization of cancers based on their molecular signatures, enabling oncologists to tailor treatment approaches that are specific to the molecular characteristics of the tumour. In clinical practice, the integration of gene expression data into diagnostic workflows has the potential to enhance the accuracy and reliability of cancer diagnosis. Traditional diagnostic methods such as imaging and histopathology provide important information about the anatomical location and morphological features of tumors [11]. However, these methods may not always capture the molecular complexity of cancer. Gene expression profiling complements existing diagnostic tools by providing molecular insights that can aid in differential diagnosis and inform treatment decisions [12]. Nevertheless, the application of gene expression data in cancer diagnosis is not without challenges. One of the primary challenges is the high dimensionality and complexity of the data. Microarray experiments generate vast amounts of gene expression measurements, often far exceeding the number of samples available for analysis. Managing and interpreting these data require sophisticated computational and statistical methods for data pre-processing, normalization, and analysis to ensure reliable results [13].

Data quality is another critical consideration in gene expression analysis. Variability in experimental conditions, sample preparation techniques, and biological factors can introduce noise and bias into the data, potentially leading to erroneous conclusions. Robust quality control measures and rigorous validation protocols are essential to minimize these sources of variability and ensure the reproducibility of research findings. Furthermore, the development of effective algorithms and computational tools is essential for extracting meaningful information from gene expression data. Techniques such as dimensionality reduction, feature selection, and machine learning algorithms play a crucial role in identifying informative biomarkers and building predictive models for cancer diagnosis [14, 15]. Cancer diagnosis using gene expression data represents a paradigm shift in oncology, offering deeper insights into the molecular basis of cancer and paving the way for personalized medicine. By leveraging the power of genomics and computational biology, researchers and clinicians can unravel the complexities of cancer biology and develop innovative strategies for early detection, accurate classification, and tailored treatment approaches. Continued advancements in technology and methodology are crucial for realizing the full potential of gene expression profiling in improving cancer outcomes and ultimately transforming the landscape of cancer care.

The contributions of this paper are manifested below,

- This work initiates with the meticulous selection of pertinent gene expression microarray datasets, followed by rigorous pre-processing steps like normalization and missing value handling. These steps ensure high data quality, essential for robust cancer diagnosis by minimizing noise and inconsistencies.
- This paper employs Principal Component Analysis (PCA) to tackle the challenge of high-dimensional data. PCA reduces dataset complexity while retaining crucial information, facilitating more efficient analysis and enhancing the interpretability of gene expression patterns in cancer classification.
- This work integrates the Hybrid Grey Wolf Red Piranha (HRPGW) algorithm for feature selection. HRPGW combines the optimization strengths of Grey Wolf Optimization with the exploratory capabilities of Red Piranha, enabling the identification of the most discriminative genes for accurate cancer classification.
- This work employs advanced classification models such as Artificial Neural Networks (ANN) and Long Short-Term Memory Networks (LSTM), specifically Artificial Long Short-Term Memory (ALSTM). These models are chosen for their ability to handle complex relationships in gene expression data, ensuring precise cancer diagnosis and facilitating early intervention strategies.

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 2023, Mondol et al. [16] developed a computationally efficient approach, hist2RNA, inspired by bulk RNA sequencing, to predict the expression of 138 genes (from 6 commercial molecular profiling tests), including luminal PAM50 subtype, from H&E-stained whole slide images (WSIs). Using TCGA data (n = 335), aggregate features from a pretrained model to predict gene expression at the patient level from annotated H&E images. In 2020, Zhang et al. [17] proposed a new image processing-based method utilizing an optimized Convolutional Neural Network (CNN) for early detection. The CNN is optimized using the improved whale optimization algorithm. Comparative evaluations on two datasets show the proposed method's superiority over other approaches, demonstrating its potential for enhanced diagnostic accuracy in medical applications. In 2020, Chen et al. [18] developed a Pathomic Fusion, a method for multimodal fusion of histology and genomic features to predict survival outcomes. Utilizing the Kronecker product and gating-based attention, this interpretable approach models feature interactions across modalities. Validated on glioma and renal carcinoma datasets from TCGA, Pathomic Fusion improves prognostic accuracy compared to unimodal models, offering a robust framework for integrating diverse biomedical data in disease prediction. In 2019, Gobin et al. [19] explores the diagnostic potential of MMP across various cancers by examining their dysregulation. MMPs, often upregulated in cancer, contribute to disease progression, angiogenesis, invasion, metastasis, and immune evasion. Using The Cancer Genome Atlas (TCGA) data, the study evaluates the diagnostic and prognostic roles of 24 MMPs in fifteen cancer types through differential expression, hierarchical clustering, and ROC analysis. In 2020, Mostavi et al. [20] introduced three Convolutional Neural Network (CNN) models (1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN) to classify tumor and non-tumor samples from The Cancer Genome Atlas (TCGA). Trained on 10,340 samples of 33 cancer types and 713 normal tissues, the models achieved 93.9-95.0% accuracy.

In Haznedar et al. [21] proposed a hybrid approach using adaptive neuro-fuzzy inference system (ANFIS), fuzzy c-means clustering (FCM), and simulated annealing (SA) algorithm. Applied to five cancer datasets, this method achieved an average accuracy of 96.28%, outperforming other algorithms like Bayesian networks, support vector machines, and J48 decision trees. In 2022, Chandrashekar et al. [22] developed for exploring, analyzing, and visualizing cancer genomic, transcriptomic, and proteomic data. Utilizing data from The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), UALCAN facilitates the evaluation of gene expression, promoter DNA methylation, and patient survival across 33 cancer types. Updated since its 2017 release, UALCAN now includes microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and enhanced visualization features, supporting cancer researchers worldwide in making discoveries and generating testable hypotheses. In 2019, Park et al. [23] offered a robust platform for exploring gene expression profiles across diverse normal and tumor tissues. Updated with Apache Lucene indexing for rapid search and a user-friendly interface using Google Web Toolkit (GWT), GENT2 now includes over 68,000 samples across 72 tissues, up from 57 in its predecessor. It supports analysis of differential expression by tumor subtypes and provides prognostic insights and meta-analysis of survival data, making it indispensable for researchers worldwide.

2.1 Problem Statement

Cancer diagnosis using gene expression microarray data faces several challenges that impact its effectiveness and reliability. Firstly, the data often suffer from noise due to variations in experimental conditions, sample preparation techniques, and inherent biological variability among patients. This noise can obscure meaningful patterns in gene expression profiles and lead to erroneous conclusions. Secondly, gene expression microarray data are characterized by high dimensionality, where the number of genes measured far exceeds the number of samples available for analysis. Managing this high-dimensional data requires sophisticated computational methods for feature selection and dimensionality reduction to extract relevant information and avoid overfitting. Additionally, ensuring data quality through rigorous pre-processing steps such as normalization and handling missing values is crucial. Variability in data quality can introduce biases that affect the accuracy and reproducibility of diagnostic results. Addressing these challenges is essential for improving the accuracy and reliability of cancer diagnosis based on gene expression microarray data, thereby enhancing early detection and treatment planning strategies.

3. PROPOSED METHODOLOGY

Cancer diagnosis using gene expression microarray data involves analysing gene activity patterns to detect cancer early and plan effective treatments. This process faces challenges such as data noise, high dimensionality, and variability across samples. To address these issues, advanced pre-processing, dimensionality reduction, feature selection, and classification algorithms are employed. These methods enhance diagnostic accuracy and reliability by mitigating noise and overfitting. Robust algorithms are crucial for handling complex data pre-processing and ensuring reproducible results, ultimately facilitating early detection and personalized treatment strategies for cancer patients. Fig. 1 depicts the overall proposed architecture.

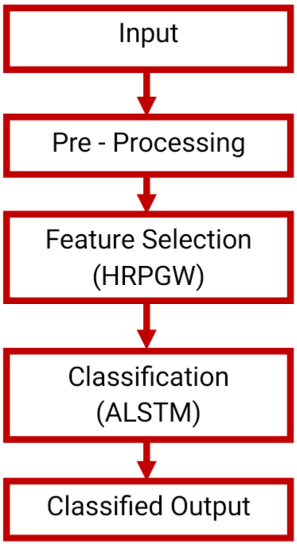


Figure 1: Proposed Architecture

3.1 Pre Processing

In the pre-processing phase, the data undergoes normalization to ensure consistency in scale and handling missing data techniques are applied to address any gaps or null values present in the dataset.

3.1.1 Normalization

Normalization is a crucial step in data pre-processing aimed at standardizing the scale of features within a dataset. It ensures that all variables contribute equally to the analysis, regardless of their original units or scales. The process involves transforming the numerical values of features to a common scale, typically between 0 and 1 or -1 and 1. One common normalization technique is Min-Max scaling, where each feature's values are transformed proportionally to fit within a specified range. This is achieved by subtracting the minimum value from each observation and dividing by the range (the maximum value minus the minimum value). 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 with a standard deviation of 1. Normalization ensures that features with larger scales do not dominate those with smaller scales during analysis, thus improving the performance and convergence of algorithms.

3.1.2 Handling Missing Data

Handling missing data is a critical aspect of data pre-processing aimed at addressing the absence of values within a dataset. Missing data can arise due to various reasons such as measurement errors, data corruption, or intentional non-response. Failing to address missing values can lead to biased analyses and inaccurate results. There are several strategies for handling missing data, including deletion, imputation, and prediction. Deletion involves removing observations or variables with missing values, but this can lead to loss of valuable information and reduced sample size. Imputation methods replace missing values with estimated values based on statistical measures such as mean, median, or mode. However, imputation can introduce bias and distort the distribution of the data. Prediction methods use machine learning algorithms to predict missing values based on other variables in the dataset

3.2 Dimensionality Reduction

Dimensionality reduction aims to streamline modelling by reducing the number of variables. It encompasses feature selection, which involves choosing significant variables, and feature extraction, which transforms high-dimensional data into fewer dimensions. This process accelerates model training and enhances accuracy by mitigating overfitting. This study used PCA for dimensionality reduction.

3.2.1 PCA

PCA is a statistical technique used for dimensionality reduction and data compression. It works by transforming high-dimensional data into a lower-dimensional representation while preserving most of the important information. The main idea behind PCA is to identify the directions (or principal components) along which the data varies the most. These principal components are linear combinations of the original features and are orthogonal to each other. PCA accomplishes this by finding the eigenvectors and eigenvalues of the covariance matrix of the data. The eigenvectors represent the principal components, and the eigenvalues indicate the amount of variance explained by each principal component. After computing the principal components, PCA projects the original data onto these components, effectively reducing the dimensionality of the data. This projection retains the maximum amount of variability present in the original data. Here's a brief explanation of PCA.

Covariance Matrix Calculation

Given a dataset X with n observations and p features, calculate the covariance matrix C is expressed as per Eq. (1).

$$\mathrm{C} = \frac{1}{\mathrm{n}-1}\left(\mathrm{X} - \underline{\mathrm{X}}\right)^{\mathrm{T}}(\mathrm{X} - \underline{\mathrm{X}}) \qquad (1)$$

Eigen value Decomposition

As per Eq. (2), compute the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and corresponding eigenvectors $(v_1, v_2, \ldots, v_p)$ of the covariance matrixC.

$$\mathrm{C}v_i = \lambda_i v_i \qquad (2)$$

Sort Eigen values

Sort the eigenvalues in descending order and rearrange the corresponding eigenvectors accordingly.

Select Principal Components

Choose the first k eigenvectors corresponding to the k largest eigenvalues to form the transformation matrix M.

Project Data

Project the original dataX onto the new subspace defined by M to obtain the transformed datasetY. PCA reduces the dimensionality of the dataset while preserving the maximum variance. It finds the directions (principal components) along which the data varies the most and projects the data onto these components, effectively capturing the essential information in a lower-dimensional space.

3.3 Feature Selection

In this study, HRPGW is employed to enhance classification accuracy by selecting relevant features and refining the collected data, ultimately improving model performance. The original ovarian cancer dataset contains 15,154 features derived from various measurements. For classification, a subset of key features was selected based on their significance, reducing dimensionality for improved performance. The selected features are representative of the diverse and crucial attributes needed for accurate ovarian cancer detection and classification.

3.3.1 HRPGW

Red Piranha Optimization (RPO) emulates the predatory behavior of red piranha fish, known for their ferocious and omnivorous nature. Divided into three phases: Searching, Encircling, and attacking, the algorithm mimics the maneuvers piranhas use while hunting. Scouts lead the swarm in random movement until prey is detected, triggering the Encircling phase. Upon encircling the prey, the swarm initiates a feeding frenzy, guided by a Prey Encircling Signal (PES) and a Frenzy Signal (FS), allowing each fish to attack and escape in a coordinated manner until the prey is consumed.

Prey Searching

During the searching phase, red piranha fish organize into a flock, with weaker individuals positioned in the middle for protection, while stronger one's act as scouts on the outskirts. Scouts ensure safety, alert the group to potential prey, and initiate encirclement. Once prey is surrounded, an attack commences, with fish following an Attack-Then-Escape (ATE) behavior until the prey is consumed. This searching behavior enhances the algorithm's exploration ability. To facilitate exploration, scouts are randomly selected to guide the search, ensuring a global exploration. The total iterations are distributed evenly among the three phases: Searching, Encircling, and Attacking. The number of leading scouts is determined, and random individuals are selected as scouts, while the remaining individuals are categorized into clusters, each with its own scout. Individual positions are updated based on their cluster's scout during exploration and represented using Eq. (3) to Eq. (6).

$$\overrightarrow{dpm} = \left|\vec{c}\vec{X}(t) - \overrightarrow{Xpm}(t)\right| \qquad (3)$$
$$\overrightarrow{Xpm}(t+1) = \vec{X}(t) - \vec{a} \cdot \overrightarrow{dpm} \qquad (4)$$
$$\vec{a} = \overrightarrow{rv_1} * (-2 + \overrightarrow{rv_2}) + (1 - \overrightarrow{rv_1})(1 + \overrightarrow{rv_3}) \qquad (5)$$
$$\vec{c} = 2 \cdot \overrightarrow{rv_4} \qquad (6)$$

where $\overrightarrow{dpm}$ is the distance between the mth piranha fish (mth solution) and the prey, $\overrightarrow{Xpm}(t)$ is the position vector of the scout of the ith cluster, $\overrightarrow{rv_1}$, $\overrightarrow{rv_2}$, $\overrightarrow{rv_3}$and $\overrightarrow{rv_4}$are random vectors in which $\overrightarrow{rv_1}$, $\overrightarrow{rv_2}$, $\overrightarrow{rv_3}$and $\overrightarrow{rv_4}$∈[0,1], $\vec{a}$ and $\vec{c}$ are coefficient vectors. In RPO algorithm, during the searching phase, the distance between each piranha fish and the prey is calculated. The position vectors of the scouts and random vectors are used to update the position of each piranha fish iteratively. The value of a coefficient vector $\vec{a}$ decreases over successive iterations, allowing piranhas to move away from their reference scout, exploring new regions in the solution domain. This adaptive variation in search vectors ensures high exploration ability, as piranhas scan the solution space effectively. During the attacking phase, the value of $\vec{a}$is restricted to the interval [-1, 1], promoting exploitation and convergence towards the prey's position. After each iteration, the location vector of each piranha fish is updated, and the objective function is calculated to reflect its proximity to the prey. Greedy selection is then applied to accept the best position for each fish, ensuring convergence towards the prey.

Prey Encircling

During the searching phase in RPO, individual movement is guided by scouts but remains inherently random, promoting exploration of new areas within the search domain. When certain fish, known as leaders or alpha fish, detect a potential prey, they issue PES, prompting others to surround the prey to immobilize it. The encircling phase utilizes a logarithmic spiral as the primary position update mechanism for herd individuals, ensuring efficient movement towards the prey's location. The algorithm selects a subset of alpha individuals, positioning the prey hypothetically at the midpoint between them. Each individual's position is then updated based on its distance from the predicted prey location, following the spiral equation. This approach facilitates effective exploitation during the encircling phase, enhancing the RPO's overall performance using Eq. (7) to Eq. (10).

$$\overrightarrow{Xprey}(t) = \frac{1}{f}\left(\frac{\frac{\sum_{i=1}^{f} x_{1i}}{\sum_{i=1}^{f} x_{2i}}}{\overset{\dots}{\sum_{i=1}^{f} x_{ui}}}\right) \qquad (7)$$
$$\vec{d} = \left|\overrightarrow{Xprey}(t) - \overrightarrow{Xpm}(t)\right| \qquad (8)$$
$$\overrightarrow{Xpm}(t+1) = \vec{d} \cdot e^{bls} \cos\cos(2\pi ls) + \overrightarrow{Xprey}(t) \qquad (9)$$
$$ls = 1 - \frac{2t}{z_{enc}} \qquad (10)$$

where, $\overrightarrow{Xprey}(t)$ is the predicted location of the prey at the iteration t, $\overrightarrow{Xpm}(t)$ is the position of the mth search agent, $\vec{d}$is the distance between the mth search agent and the prey, b is a constant defines the shape of the logarithmic spiral, and l is a number in the interval [−1,1].

Prey Attacking

At the culmination of the encircling phase, the prey is tightly surrounded, rendering escape impossible. Alpha fish then emit a (FS), initiating the attack phase. In this state of frenzy, the herd competes to reach the prey, following the alpha fish closely. During the attack phase, each search agent's position is updated using Eq. (11) to Eq. (15). Initially, the n alpha individuals are identified, and the predicted prey location is calculated. Subsequently, the position of the remaining herd members is determined by calculating their distance from the prey and updating their positions accordingly.

$$\overrightarrow{dpm} = \left|\vec{c}\overrightarrow{Xprey}(t) - \overrightarrow{Xpm}(t)\right| \qquad (11)$$
$$\overrightarrow{Xpm}(t+1) = \overrightarrow{Xprey}(t) - \vec{a} \cdot \overrightarrow{dpm} \qquad (12)$$
$$\vec{a} = 2\vec{g} \cdot \overrightarrow{rv_1} - \vec{g} \qquad (13)$$
$$\vec{c} = 2 \cdot \overrightarrow{rv_2} \qquad (14)$$
$$g = 2 - t * \frac{2}{z_{att}} \qquad (15)$$

Where, $\overrightarrow{Xprey}(t)$ is the predicted location of the prey at the iterationt, $\overrightarrow{Xpm}(t)$ is the position of the mth search agent, $\overrightarrow{dpm}$ is the distance between the mth search agent and the prey, $\overrightarrow{rv_1}$ and $\overrightarrow{rv_2}$ are random vectors ∈ [0,1], $\vec{g}$vector decreases linearly from 2 to 0 over the course of iterations, $\vec{a}$ and $\vec{c}$ are coefficient vectors. During the attacking phase, $\vec{a}$ is set to a random value in [−1,1], hence, new search agent position will lie somewhere between the prey's position and the search agent's present position.

3.3.2 Hybrid Red Piranha Grey Wolf Optimization

Incorporating the grey wolf hunting strategy with the prey searching strategy of RPO enhances the algorithm's ability to escape local optima. In the natural world, grey wolves exhibit coordinated hunting behavior led by the alpha, with occasional participation from the beta and delta wolves. Similarly, in RPO, the scouts play a crucial role in guiding the search process randomly. By adapting the grey wolf strategy, the best solutions obtained (α, β, γ) serve as guides for other search agents. This means that while the alpha, beta, and delta wolves have better knowledge of prey location, the rest of the pack adjusts their positions based on these leaders. In the context of RPO, this translates to updating the positions of search agents based on the positions of the best solutions found so far. By leveraging this collective intelligence, RPO improves its exploration ability, enabling it to effectively navigate the search space and avoid getting trapped in local optima as per the proposed Eq. (16) to Eq. (18). This integration combines the strengths of both strategies, resulting in a more robust optimization algorithm capable of achieving higher-quality solutions.

$$\overrightarrow{\mathrm{dpm}_\alpha} = \left|\vec{c}_1\vec{X}_\alpha - \overrightarrow{\mathrm{Xpm}}\right| \quad ,\overrightarrow{\mathrm{dpm}_\beta} = \left|\vec{c}_2\vec{X}_\beta - \overrightarrow{\mathrm{Xpm}}\right|, \overrightarrow{\mathrm{dpm}_\gamma} = \left|\vec{c}_3\vec{X}_\gamma - \overrightarrow{\mathrm{Xpm}}\right| \qquad (16)$$

$$\overrightarrow{\mathrm{Xpm}} = \vec{X}_\alpha - \vec{a}_1 \cdot \overrightarrow{\mathrm{dpm}_\alpha} \quad , \overrightarrow{\mathrm{Xpm}_2} = \vec{X}_\beta - \vec{a}_2 \cdot \overrightarrow{\mathrm{dpm}_\beta}, \overrightarrow{\mathrm{Xpm}_3} = \vec{X}_\gamma - \vec{a}_3 \cdot \overrightarrow{\mathrm{dpm}_\gamma} \qquad (17)$$

$$\overrightarrow{\mathrm{Xpm}}(t+1) = \frac{\overrightarrow{\mathrm{Xpm}\ _1} + \overrightarrow{\mathrm{Xpm}\ _2} + \overrightarrow{\mathrm{Xpm}\ _3}}{3} \qquad (18)$$

3.4 Classification

In this study, the classification phase utilizes advanced models including ANN and LSTM, specifically designed as ALSTM. These models are chosen for their capability to effectively analyze complex gene expression patterns, ensuring precise and reliable cancer diagnosis for early intervention and treatment planning. In this study, the classification phase utilizes advanced models tailored for analyzing gene expression data, crucial for precise cancer diagnosis. ANN are employed to process raw features through layered computations, optimizing classification accuracy by capturing complex relationships within the data. ANN excel in initial feature extraction and nonlinear mappings, enhancing the understanding of gene activity patterns crucial for tumor detection. Also, LSTM networks are integrated to handle sequential dependencies inherent in gene expression sequences. LSTM effectively capture temporal dynamics, ensuring robust detection of long-term patterns that signify cancer development or progression. Together, these models offer a synergistic approach: ANN for comprehensive feature analysis and initial classification, and LSTM for nuanced understanding of sequential data, thus improving diagnostic accuracy and enabling personalized treatment strategies based on individual gene expression profiles.

3.4.1 ANN

- Input Layer

The input layer forwards raw data or features to hidden layers for processing, with one neuron per feature. No computation occurs within this layer.

- Hidden Layers

Hidden layers carry out the primary computation in neural networks. Neurons within a hidden layer receive inputs from the preceding layer, compute a weighted sum, and apply an activation function to generate an output. The configuration of hidden layers is tailored to problem complexity and calculated as per Eq. (19).

$$a_{ij} = f\left(\sum_{k=1}^{n_{i-1}} w_{ik} a_{ik} + b_{ij}\right) \qquad (19)$$

The output of each neuron in a hidden layer is calculated using a weighted sum of inputs and a bias term.

- Output Layer

The output layer generates the neural network's final output, with the number of neurons determined by the problem type.

3.4.2 LSTM

Recurrent neural networks (RNN) of the Long Short-Term Memory (LSTM) type were created to solve the problem of identifying long-term dependencies in sequential input. Because of their memory blocks, which allow them to retain information over extended periods of time, they are particularly well-suited for tasks involving time-series data and natural language processing.

- Input Layer: Receives sequential data input.
- Hidden Layer: Contains LSTM units responsible for processing and retaining information.
- Output Layer: Produces the final output based on the processed information.

LSTM replaces the basic units of regular RNNs with memory cells, which allow them to retain information over long sequences. LSTM units have three main gates: input gate, forget gate, and output gate.

- Input Gate: Regulates how fresh data enters the memory cell.
- Forget Gate: Selects the data from the memory cell to remove.
- Output Gate: Adjusts the output according to the input's previous state and present value.

The activation of each LSTM unit at time $l_t$is calculated using Eq. (20):

$$l_t = \sigma(wm_{i,l} \cdot x_t + wm_{h,l} \cdot l_{t-1} + bi) \qquad (20)$$

Where,$l_t$and l_{t-1}represent the activation at time respectively, σ is a non-linear activation function,$wm_{i,l}$ is the input-hidden weight matrix,$wm_{h,l}$ is the hidden-hidden weight matrix, bi is the hidden bias vector, and $x_t$is the input at time t. LSTM networks excel at capturing long-term dependencies in sequential data. They mitigate the problem of gradient vanishing, allowing for more effective learning over longer sequences. ALSTM algorithm integrates attention mechanism with LSTM, enhancing sequence prediction by focusing on relevant information in algorithm 1.

Algorithm 1: ALSTM

```
Start
Define preprocess_data (data)
   Normalize data by subtracting the mean and dividing by the standard deviation
   Return normalized data
ENDDEF
Define Function build_ann_model (input_shape)
   Create Sequential model
   Add Input layer with shape input_shape
   Add Dense layer with 128 units and 'relu' activation
   Add Dropout layer with rate 0.5 for regularization
   Add Dense layer with 64 units and 'relu' activation
   Return ann_model
ENDDEF
Define Function build_lstm_model (input_shape)
   Create Sequential model
   Add LSTM layer with 128 units and 'relu' activation, return_sequences=True, input_shape=input_shape
   Add Dropout layer with rate 0.5 for regularization
   Add LSTM layer with 64 units and 'relu' activation, return_sequences=False
 Add Dense layer with 32 units and 'relu' activation
   Add Dense layer with 1 unit and 'sigmoid' activation for binary classification
   Return lstm_model
ENDDEF
Define Function train_and_evaluate (data, labels)
   Preprocess data using preprocess_data function
   Split data into training, validation, and test sets
   Build ANN model using build_ann_model function
   Train ANN model on training data
   Extract features from the final hidden layer of ANN model
   Reshape extracted features to fit LSTM input requirements
   Build LSTM model using build_lstm_model function
   Train LSTM model on extracted features from ANN
   Evaluate LSTM model on test set
   Return evaluation results
ENDDEF
END
```

4. RESULTS AND DISCUSSION

The proposed model is implemented using the Python platform and benchmarked against existing models like Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Multiple Linear Regression (MLR). 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. This comparative 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.

The Ovarian cancer dataset [24], used in this research, comprises 15,154 genes (features) and 253 observations divided into two classes: 162 individuals with ovarian cancer and 91 healthy individuals. It features 15,154 numeric (continuous) attributes and one categorical attribute, generated using the WCX2 protein chip, distinguishing it from the Ovarian cancer dataset. This dataset is crucial for researchers aiming to explore the genetic basis of ovarian cancer, offering a detailed genetic profile that can aid in developing predictive models and identifying potential biomarkers. Its comprehensive nature makes it an invaluable resource for advancing diagnostic and therapeutic strategies in ovarian cancer research. The rich genetic information it provides is instrumental in understanding the disease's mechanisms and improving patient outcomes through targeted interventions.

4.1 Performance Metrics

In optimizing cancer diagnosis, accuracy, recall, precision, and F1-score assess classification algorithms' ability to predict cancer cases effectively. These metrics gauge overall correctness, sensitivity to detecting cancer, precision in positive predictions, and a balanced measure of both precision and recall, critical for enhancing early detection and treatment strategies.

- Accuracy

  Accuracy measures the proportion of correctly classified instances among all instances. It is calculated as per Eq. (21).

$$\text{Accuracy} = \frac{\text{Totalnumberofinstances}}{\text{TruePositives} \quad +\text{TrueNegatives}} \qquad (21)$$

- Precision

  Precision measures the proportion of instances classified as positive that were actually positive. It is calculated as per Eq. (22).

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} \quad +\text{FalsePositives}} \qquad (22)$$

- Recall

  Recall measures the proportion of actual positive instances that were correctly identified by the classifier. It is calculated as per Eq. (23).

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} \quad +\text{FalseNegatives}} \qquad (23)$$

- F1-score

  F1-score is the harmonic mean of precision and recall, providing a balanced measure of a classifier's performance. It is calculated as per Eq. (24).

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \quad \times\text{Recall}}{\text{Precision} \quad +\text{Recall}} \qquad (24)$$

These metrics help evaluate the performance of models, with accuracy providing an overall measure of correctness, recall assessing the model's ability to capture all positive instances, precision measuring the model's accuracy in labelling positive instances, and F1-score offering a balanced measure considering both precision and recall.

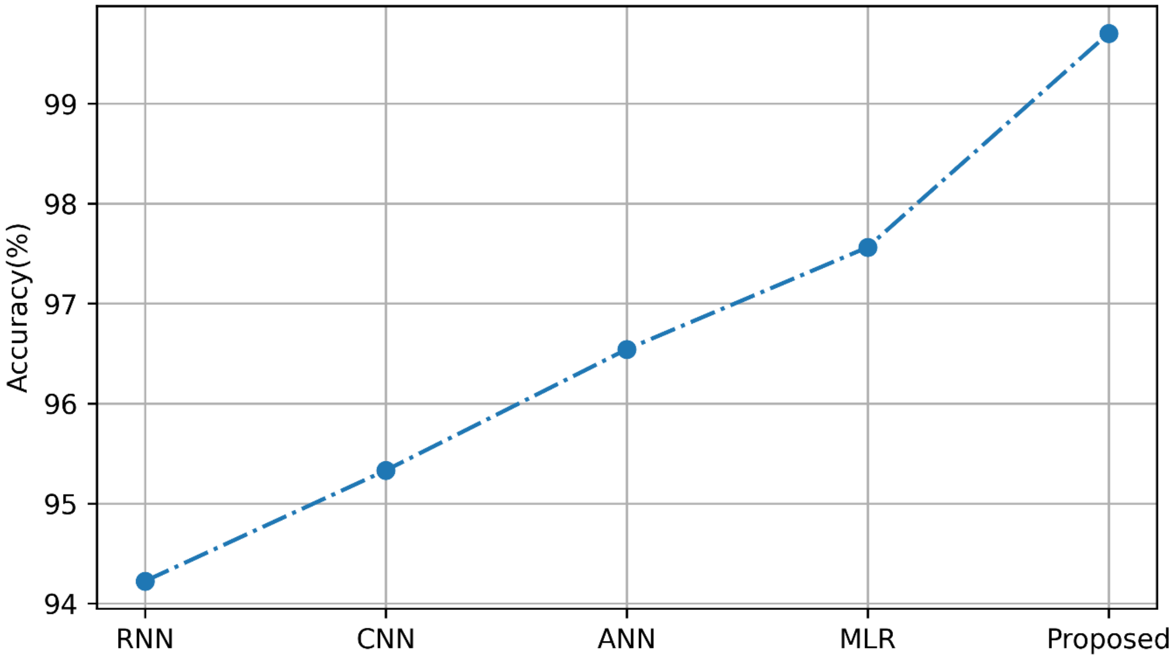
4.2 Performance Analysis

The performance analysis of existing models and the proposed model for cancer diagnosis is summarized in Table 1. The table presents four key performance metrics: accuracy, precision, recall, and F1-score. These metrics are crucial in evaluating the effectiveness of the models in diagnosing cancer. Accuracy measures the proportion of true results among the total number of cases examined. The proposed method is compared with RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), ANN (Artificial Neural Network) and MLR (Multiple Linear Regression). The proposed model achieves an impressive accuracy of 99.704%, significantly higher than RNN (94.223%), CNN (95.332%), ANN (96.543%), and MLR (97.564%). Precision indicates the proportion of true positive results among the total positive results predicted by the model. The proposed model reaches nearly perfect precision at 99.999%, outperforming RNN (92.453%), CNN (94.356%), ANN (95.345%), and MLR (97.543%). Recall measures the proportion of true positive results correctly identified by the model out of all actual positive cases. The proposed model scores a remarkable 99.965%, which is superior to RNN (92.765%), CNN (94.346%), ANN (96.745%), and MLR (97.765%). F1-score is the harmonic mean of precision and recall, providing a balance between the two. The proposed model achieves an F1-score of 99.954%, well above RNN (93.654%), CNN (95.654%), ANN (96.975%), and MLR (97.769%). The proposed model demonstrates superior performance across all metrics compared to the existing models. Its high accuracy, precision, recall, and F1-score highlight its effectiveness and reliability in cancer diagnosis. These results indicate that the proposed model not only enhances diagnostic accuracy but also ensures that the diagnoses are both precise and comprehensive. Consequently, this model shows great potential for improving early detection and personalized treatment strategies for cancer patients

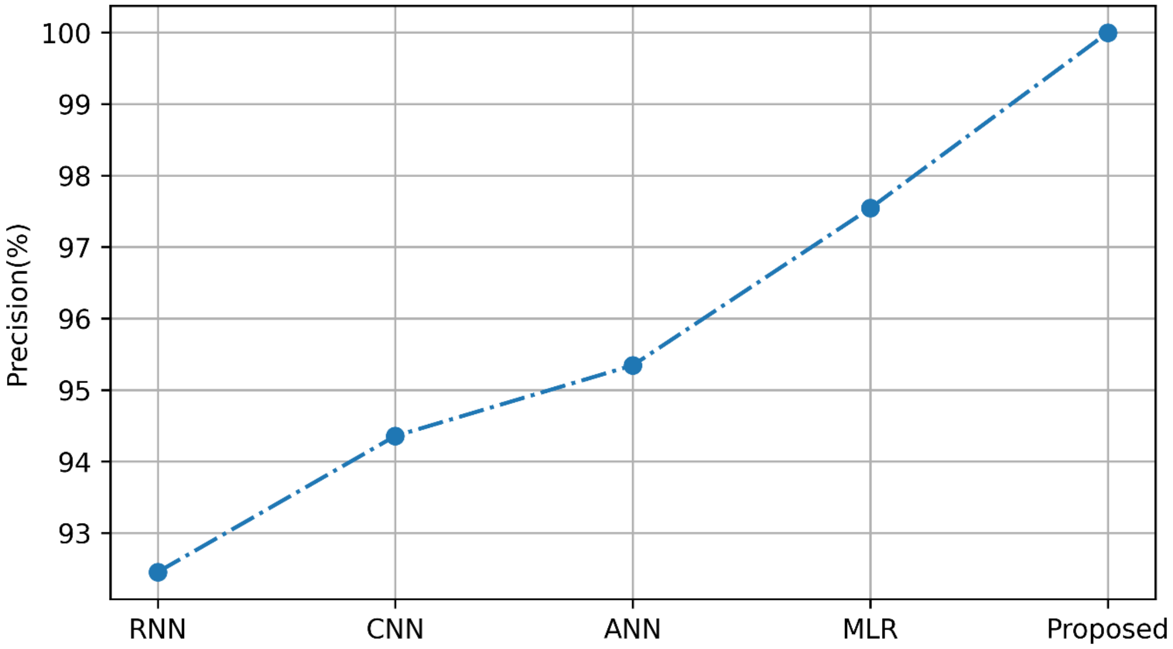
Table 1: Performance Analysis

Methods	RNN	CNN	ANN	Proposed
Accuracy	94.223	95.332	96.543	99.704
Precision	92.453	94.356	95.345	99.999
Recall	92.765	94.346	96.745	99.965
F1-score	93.654	95.654	96.975	99.954

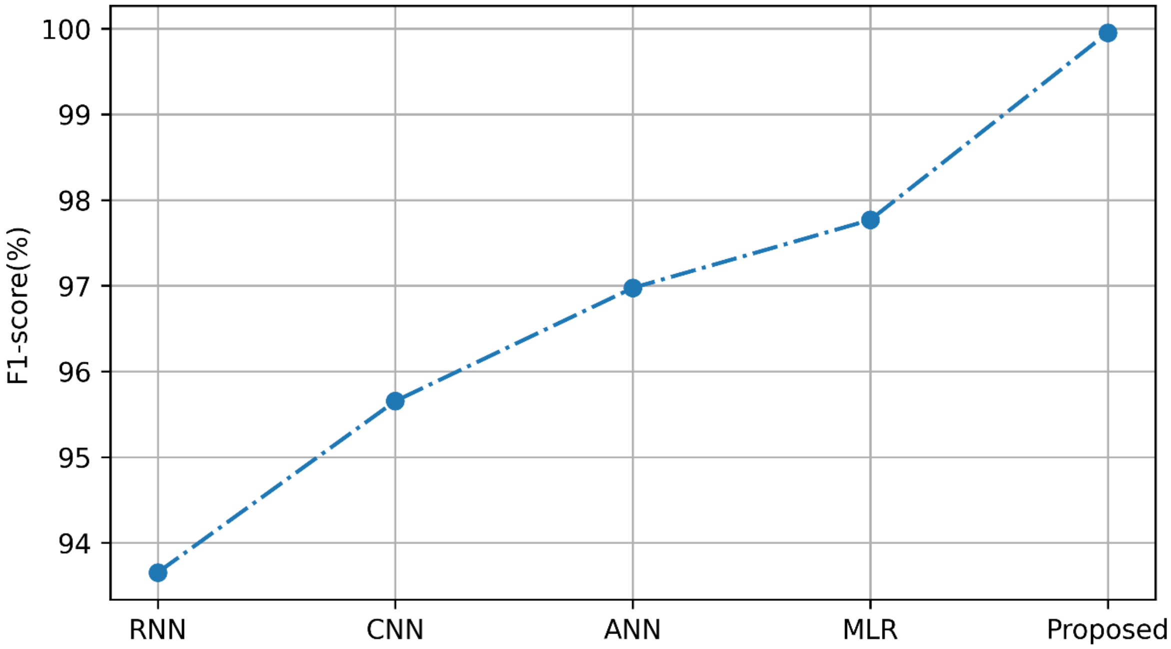
4.3 Graphical Representation



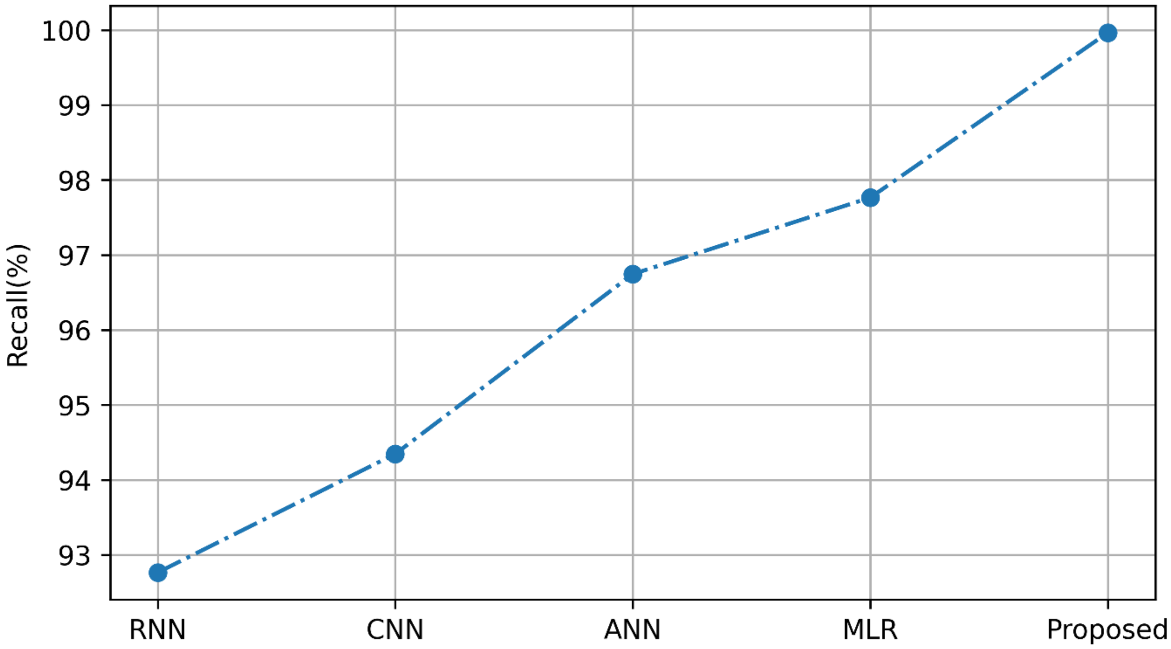
(a)



(b)



(c)



(d)

Figure 2: Graphical Representation of Existing and Proposed Model (a) Accuracy (b) Precision (c) F1-Score (d) Recall

Fig. 3 provides a graphical representation comparing the performance of existing models (RNN, CNN, ANN, MLR) and the proposed model across four metrics: (a) Accuracy, (b) Precision, (c) F1-Score, and (d) Recall. The graphs clearly illustrate the superior performance of the proposed model, which achieves higher scores in all metrics. This visual comparison underscores the model's effectiveness and reliability in cancer diagnosis, showcasing its potential for better early detection and personalized treatment compared to traditional methods.

5. CONCLUSION

The process of classifying samples as cancerous or non-cancerous by the analysis of gene expression microarray data helped in the early detection and treatment planning of a variety of cancer types. The study encountered several obstacles, such as fluctuations and noise in the data, selecting features from large-scale datasets, overfitting, unequal distribution of classes, and the requirement for reliable algorithms to successfully differentiate between samples with and without cancer. In order to analyze microarray data for cancer detection, the study used a thorough methodology. Prioritizing pertinent microarray data was the first step, and then crucial data pre-processing procedures such handling missing values and standardization were performed to guarantee the quality of the data. To minimize the complexity of the dataset, dimensionality reduction techniques were used, especially PCA. HRPGW algorithm was then used to select the features. It merged the exploration potential of Red Piranha with the optimization strength of Grey Wolf Optimization. Following that, classification models were applied for precise cancer diagnosis, such as ALSTM, which includes ANN and LSTM. By ensuring that the most pertinent features were retrieved from the data and minimizing the effects of noise and high dimensionality present in microarray datasets, this integrated strategy improved classification performance and, in the end, the precision and dependability of cancer detection.

Conflicts of interest

The authors have no conflicts of interest to declare.

REFERENCES

[1] Zhang S, Rong F, Guo C, Duan F, He L, Wang M, Zhang Z, Kang M, Du M. Metal–organic frameworks (MOFs) based electrochemical biosensors for early cancer diagnosis in vitro. Coordination Chemistry Reviews. 2021 Jul 15;439:213948.

[2] Schoots IG, Padhani AR, Rouvière O, Barentsz JO, Richenberg J. Analysis of magnetic resonance imaging–directed biopsy strategies for changing the paradigm of prostate cancer diagnosis. European urology oncology. 2020 Feb 1;3(1):32-41.

[3] Yu K, Tan L, Lin L, Cheng X, Yi Z, Sato T. Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health. IEEE Wireless Communications. 2021 Jun;28(3):54-61.

[4] Li H, Wu P, Wang Z, Mao J, Alsaadi FE, Zeng N. A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis. Computers in biology and medicine. 2022 Dec 1;151:106265.

[5] Horie Y, Yoshio T, Aoyama K, Yoshimizu S, Horiuchi Y, Ishiyama A, Hirasawa T, Tsuchida T, Ozawa T, Ishihara S, Kumagai Y. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointestinal endoscopy. 2019 Jan 1;89(1):25-32.

[6] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine. 2021 Dec;13:1-7.

[7] Sayed S, Nassef M, Badr A, Farag I. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications. 2019 May 1;121:233-43.

[8] Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications. 2020 Mar;79(11):7731-62.

[9] Weng, Shengquan. "Convergence theorems of modified Ishikawa iterations in Banach spaces." Results in Nonlinear Analysis 2.3 (2019): 125-135.

[10] Racle J, Gfeller D. EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. Bioinformatics for cancer immunotherapy: methods and protocols. 2020:233-48.

[11] Moody L, Mantha S, Chen H, Pan YX. Computational methods to identify bimodal gene expression and facilitate personalized treatment in cancer patients. Journal of Biomedical Informatics. 2019 Jan 1;100:100001.

[12] Bosma SC, Hoogstraat M, van der Leij F, de Maaker M, Wesseling J, Lips E, Loo CE, Rutgers EJ, Elkhuizen PH, Bartelink H, van de Vijver MJ. Response to preoperative radiation therapy in relation to gene expression patterns in breast cancer patients. International Journal of Radiation Oncology* Biology* Physics. 2020 Jan 1;106(1):174-81.

[13] Sharifi M, Avadi MR, Attar F, Dashtestani F, Ghorchian H, Rezayat SM, Saboury AA, Falahati M. Cancer diagnosis using nanomaterials based electrochemical nanobiosensors. Biosensors and Bioelectronics. 2019 Feb 1;126:773-84.

[14] Ren H, Zhu J, Yu H, Bazhin AV, Westphalen CB, Renz BW, Jacob SN, Lampert C, Werner J, Angele MK, Bösch F. Angiogenesis-related gene expression signatures predicting prognosis in gastric cancer patients. Cancers. 2020 Dec 8;12(12):3685.

[15] Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC bioinformatics. 2019 Dec;20:1-6.

[16] Zhou R, Zhang J, Zeng D, Sun H, Rong X, Shi M, Bin J, Liao Y, Liao W. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I–III colon cancer. Cancer immunology, immunotherapy. 2019 Mar 13;68:433-42.

[17] Mondol RK, Millar EK, Graham PH, Browne L, Sowmya A, Meijering E. hist2rna: an efficient deep learning architecture to predict gene expression from breast cancer histopathology images. Cancers. 2023 Apr 30;15(9):2569.

[18] Zhang N, Cai YX, Wang YY, Tian YT, Wang XL, Badami B. Skin cancer diagnosis based on optimized convolutional neural network. Artificial intelligence in medicine. 2020 Jan 1;102:101756.

[19] Chen RJ, Lu MY, Wang J, Williamson DF, Rodig SJ, Lindeman NI, Mahmood F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging. 2020 Sep 3;41(4):757-70.

[20] Rahim, Robbi. "Quantum Computing in Communication Engineering: Potential and Practical Implementation." Progress in Electronics and Communication Engineering 1.1 (2024): 26-31.

[21] Gobin E, Bagwell K, Wagner J, Mysona D, Sandirasegarane S, Smith N, Bai S, Sharma A, Schleifer R, She JX. A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. BMC cancer. 2019 Dec;19:1-0.

[22] Mostavi M, Chiu YC, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC medical genomics. 2020 Apr;13:1-3.

[23] Haznedar B, Arslan MT, Kalinli A. Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. Medical & Biological Engineering & Computing. 2021 Mar;59:497-509..

[24] Chandrashekar DS, Karthikeyan SK, Korla PK, Patel H, Shovon AR, Athar M, Netto GJ, Qin ZS, Kumar S, Manne U, Creighton CJ. UALCAN: An update to the integrated cancer data analysis platform. Neoplasia. 2022 Mar 1;25:18-27.

[25] Park SJ, Yoon BH, Kim SK, Kim SY. GENT2: an updated gene expression database for normal and tumor tissues. BMC medical genomics. 2019 Jul;12:1-8.

[26] Dataset taken from: “https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer”, dated 1/6/2024.