




Evolutionary Hybrid Machine Learning Techniques for DNA Cancer Data Classification

Hussam Jasim Ali¹, Sameer Alani^{1*}, Riyadh Jameel Toama², Tabarak Ali Abdulhussein²

¹ Computer Center, University of Anbar, Anbar 55431, Iraq

² Department of Computer Technology Engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad 10011, Iraq

Corresponding Author Email: Sameer.h@uoanbar.edu.iq

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ria.380206>

ABSTRACT

Received: 27 September 2023

Revised: 2 January 2024

Accepted: 8 February 2024

Available online: 24 April 2024

Keywords:

machine learning, DNA, KNN, PNN, PSO

Early cancer detection is crucial for superior therapy and survival rates. Algorithms that use machine learning have proven potential for classifying tumors through the use of gene expression data. This study seeks to formulate a hybrid evolutionary machine learning approach to accurately classifying DNA microarray data as normal or malignant. Gene expression data for breast, brain, and colon cancer were acquired. The datasets were subjected to pre-processing, including normalization and missing value imputation. A hybrid technique has been proposed, which combines particle swarm optimization (PSO) for feature selection with probabilistic neural networks (PNN) for classification. PSO identified an appropriate selection of features to increase classification performance. The PSO-PNN hybrid model obtained classification accuracies of 91.46% on breast cancer data, 91.54% on brain cancer data, and 95.16% on colon cancer data outperforming alternate approaches. The findings show that an evolutionary hybrid technique combining PSO and PNN can reliably classify malignant gene expression profiles. This can help with early cancer detection. The proposed technique outperforms conventional machine learning algorithms. Further research can validate these findings on more cancer datasets.

1. INTRODUCTION

Machine learning, a subfield of artificial intelligence, traces its origins back to the 1950s. The term "machine learning" itself was coined in 1959 by IBM researcher Arthur Samuel, one of the pioneers in the area. Samuel defined machine learning as the ability of computers to learn without being explicitly programmed. In his early work, he developed a self-learning checker program that progressively improved its performance by analyzing its games [1, 2].

Samuel's groundbreaking research established machine learning as a subdomain within AI focused on building systems that can automatically learn and gain intelligence from data. In the following decades, Machine learning has evolved into specialized approaches such as neural networks, decision tree learning, Bayesian networks, clustering, and reinforcement learning. Machine learning improvements began in the 1990s as processing power increased and vast datasets became available.

The rise of machine learning today, driven by algorithm improvements, data growth, and affordable GPU-powered computing, finds adoption across sectors such as healthcare, finance, transportation, security, and the sciences. Despite its extensive potential, however, machine learning still falls short of replicating human intelligence. Overcoming these hurdles to artificial general intelligence, seen as the AI research frontier, remains a grand challenge [3, 4]. Machine learning (ML) can be supervised or unsupervised. Supervised learning

utilizes well-defined data for training. Unsupervised learning often provides superior performance and results for large data sets [5, 6]. Furthermore, Experience is the data on which the machine learning model is trained in machine learning. ML models learn patterns and relationships from the data, which allows them to make predictions without needing to be explicitly programmed. Specifically: The training data that is fed into the machine learning algorithm serves as the experience that enables it to learn. Essentially, this data represents real-world examples of what the model needs to understand [7].

Deoxyribonucleic acid (DNA) is a lengthy biomolecule formed out of four basic units called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine. These nucleotides are generally referred to as bases for DNA. The sequence in which these four bases are organized along the DNA strand provides genetic instructions for cell function, development, and replication. Even little alterations in base order or mutations can result in obvious changes in physical characteristics, and predispositions to diseases, and influence many other biological processes. Moreover, the sequence of DNA strands determines its genetic information and is critical for establishing genetic identity. For example, no two people (except identical twins) have the same DNA base sequence [8]. Thus, examining similarities and variances in base sequencing across individuals or species enables us to identify genetic linkages and variability.

2. LITERATURE REVIEW

The term "cancer" refers to a class of disorders caused by abnormal cell proliferation that can spread to other parts of the body. Cancer, after cardiovascular illnesses, is the second leading cause of mortality, according to the World Health Organization. In recent years, the field of biomedical research has demonstrated a strong interest in data analysis. Several studies in published literature have assessed developments in computational approaches to gene expression [9].

Since the late nineteenth century, gene expression analysis has been a focus of biomedical research to better understand the biological mechanisms underlying development, disease, aging, and responses to environmental signals. Early gene analysis techniques used low-throughput wet lab processes like northern blots or quantitative PCR to measure mRNA levels one gene at a time. The first significant computational breakthrough happened in 1995 with the advent of DNA microarrays, which allowed for the simultaneous measurement of thousands of gene transcripts on a single RNA sample using hybridization.

Over the following decade, researchers created a variety of computational methods for interpreting microarray data, such as finding differentially expressed genes, clustering co-regulated genes, mapping regulatory networks, molecular classification of disease subtypes, and tumor identification. Statistical techniques, unsupervised learning, and pattern recognition algorithms tailored to high-dimensional expression data were among the most noteworthy advancements. By the early 2000s, supervised classifiers had shown success in predicting cancer types using gene expression patterns.

The next stage of advancement was next-generation sequencing, which dramatically enhanced biological data. New computing disciplines, such as machine learning, proved important in extracting insights from massive, complex genetic information. Deep neural networks outperformed traditional methods for tasks such as recognizing transcript isoforms, predicting DNA, inferring gene regulatory logic, and detecting molecular interactions. Cloud computing enabled the scaling of analysis procedures. Dynamic meta-learning frameworks continue to improve performance on molecular prediction tasks.

A study has presented a method for predicting a splice site based on studies showing that using a second-order Markov model and creating a support vector machine (SVM) is more efficient than using a first-order Markov model as a pre-model. The processing method is effective when combining this method with SVM that will provide higher classification accuracy regarding Splice location prediction [10].

On the other hand, In 2016, an article presented a probabilistic method for estimating the contributor number in a DNA mixture to evaluate the method, they compared the performance of the classification of six ML algorithms and evaluated this model concerning the performance of this algorithm the total results illustrated that it is possible to detect up to four contributors in a DNA mixture with a total accuracy of more than 98% [11]. Another study presented BIGBIOCL, an algorithm that applied methods of supervised classification to datasets. It is intended to extract the alternative and equivalent classification models by deleting selected features iteratively. Experiments are conducted using DNA methylation datasets, they perform the classification process to extract many methylation sites and their related genes in

precise performance [12]. Another study presented DL as a group approach that includes many different ML models. The useful gene data selected by differential gene expression analysis are provided for five different classification models. Thus, the DL method is used to aggregate the outputs of the five classifiers [10]. To overcome the challenges arising from high-dimensional data in microarrays, a method combining hybrid and genetic trait selection techniques has been proposed [13]. This should ultimately increase the accuracy of cancer classification. Initially, the most important features in malicious microarray datasets are found using filtering feature selection techniques such as informative interest, informative interest ratio, and chi-square. The selected features are then improved and refined using a genetic algorithm, enhancing the overall performance of the proposed cancer classification method.

In this article, the authors suggested four microarray datasets primarily associated with breast, lung, central nervous system, and brain cancers. The results prove that the feature selection by hybrid and genetic filtering outperforms several traditional machine learning techniques, in terms of precision, recall, precision, and F-measurement. On another hand, a novel medical support system is offered, in which genes of interest are chosen from next-generation sequencing (NGS) datasets using a hybrid multi-stage gene identification approach that combines Relief-Cuckoo search (CS) and Random Forest (RF). This approach is effective at identifying infection-induced septicemia and related genes, allowing for early disease detection. Hybrid feature selection approaches have advanced in later stages of illness therapy, such as diagnostic and medication development. The suggested model, with an accuracy rate of 95.23%, is a useful tool for in-depth analysis of diverse viral illnesses and simplifies the diagnosis process [14].

Improving cancer classification is the goal of the proposed cancer optimization algorithm using the Binary Competitive Search of Woe Optimization (IBCSOWOA) algorithm. This technique is based on the use of minimum repeat information and maximum relevance (mRMR) as a filter method and applies IBCSO to reduce the fraction that includes informative genes [15]. An artificial neural network (ANN) model is used to evaluate the IBCSOWA approach, and a woe optimization algorithm (WOA) is used to adjust the modification of model parameters. Six transformation-based microarray datasets are used to evaluate the performance of IBCSOWA and compare it with other disease prediction techniques. Experimental results, demonstrating the optimal fraction of features, classification accuracy, and digestion rate, demonstrate the superiority of the proposed strategy over alternative, nature-inspired methods. Remarkably, the proposed method exceeds 98% accuracy on all six datasets; The lung cancer dataset achieves the highest accuracy rate of 99.45%. The main aim of this suggested project was to develop a Gene Expression Cancer Classification Network (GECC-Net) using artificial intelligence techniques [14]. Initially, transfer learning based on AlexNet is used to extract features from a dataset by evaluating the relationships between distinct entities. The optimal fuzzy rules for feature selection are then identified using a hybrid fuzzy ranking network (HFRN). Furthermore, for the multi-class classification task, ovarian, colon, and lymphoma malignancies are classified using a multi-kernel support machine (MK-SVM). Simulation findings demonstrate that the provided GECC-Net outperforms the latest methods. To combat breast cancer, a study was

conducted to develop a hybrid machine-learning model for early prediction of breast cancer [16]. XBoost tree, multilayer perceptron, logistic regression, random tree classification, and random tree classification were applied to a dataset from Kaggle to predict the growth and sizes of breast tumors. Python language was used to implement these machine-learning algorithms and display the results. The results showed significantly high accuracy (99.65%) compared to standard methods on both the training and test datasets. The predictive model shows great potential in improving early recognition and detection of breast cancer, which will enhance treatment outcomes. In addition, it may provide important support to patients in managing their lifestyle and condition, contributing to their survival and recovery. Furthermore, patients with squamous cell carcinoma (head and neck) are at risk for developing a second squamous cell carcinoma of the lung or pulmonary metastases [17]. Recognizing the difference between pulmonary metastases and original lung tumors is critical in clinical practice, although it is not always attainable (with current technologies). DNA methylation profiling is performed (on primary cancers). Following that, three different machine learning approaches are trained to distinguish the metastatic HNSC from the main LUSC. An artificial neural network was created to classify 96.4% of the cases for 279 patients (validation cohort) with LUSC and HNSC, exceeding random forests by 87.8% and support vector machines by 95.7%. Accuracy predictions of more than 99% are attained for neural networks at 92.1%, support vector machines at 90%, and random forests at 43%, using thresholds applied to the probability scores that generated. Finally, the approach can help guide therapy options by providing a more reliable distinction diagnostic of pulmonary metastases (for HNSC from primary LUSC).

Another study was conducted and showed that up to 90% of cancer deaths are from metastatic cancers. The known differences from the primary cancers are crucial for identifying cancer types and targeted treatment development (for each type of cancer). An interesting cancer prediction target is the DNA methylation patterns, also an essential mediator for the possible change to metastatic cancer. In this study, 24 types of cancer are used and 9303 samples of methylome are downloaded from well-known repositories of medical data, such as GEO (Gene Expression Omnibus) and TCGA (The Cancer Genome Atlas). Machine learning classifiers are constructed for discriminating non-cancerous methylome, primary, and metastatic samples. ML models such as RF (random forest), XGBoost (extreme gradient boosting), NB (Naive Bayes), and SVM (support vector machines) are applied to classify the types of cancer relying on their origin tissue. Almost all classifiers above are outperformed by RF (random forest), for an accuracy of up to 99%. Furthermore, to classify cancer types a local interpretable model agnostic explanation (called LIME) is applied to illustrate essential methylation biomarkers [18].

In 2022, An improved machine learning method was proposed for analyzing the human gene sequencing and tumor sequencing patterns. It is possible for patients with tumor sequencing with the help of different medical systems to monitor the changes in the tumor genome. Genetic testing or genetic specification is another term for tumor DNA sequencing. To develop a personalized plan for cancer treatment depending on the tumor's molecular characteristics (rather than a one-size-fits-all treatment method) clinical decision making is done with the help of sequence results.

Tumor sequencing had a main role in cancer research. It analyzes the patients' circulatory problems with different types of tumors for public-domain analysis. It also monitors cancer or tumor genetic sequences in large datasets for calculating tumor location and size. This aids the doctor in having an accurate insight report for the tumor type and problems that may occur to patients. Two main conclusions from the analysis of cancer-tumor-gene-sequences datasets are made: each patient's genetic makeup is different and no identical two cancers are there [19].

The studies examine a variety of machine learning approaches to cancer data, such as Markov models, probabilistic methods, classification algorithms, deep learning ensembles, hybrid filter-genetic methods, Relief-Cuckoo Search-Random Forest pipelines, Improved Binary Competitive Swarm Optimization with Whale Optimization, transfer learning, and multi-kernel support vector machines. While most research provides good classification performance measures, they do not provide extensive ablation analyses to demonstrate the added advantage of the approaches used over simpler baselines. The lack of model interpretation reduces clinical value. Comparative benchmarking against popular public datasets is likewise lacking, making performance claims difficult to understand. The primary focus is on genomic information such as splice sites, methylation patterns, and gene expression. While giving biological insights, successful multi-view cancer diagnosis requires the examination of many data modalities such as imaging, histology, and proteomics. The assessment was therefore limited to binary classification, but practical utility requires multi-class labeling and detection. Deep neural networks and Relief-Cuckoo search show promise, but tunability and generalization remain tough. Simpler tree-based and SVM classifiers can be more reliable but less efficient. Handling class imbalance and validation on separate test sets is not well addressed. Overall, there is a scarcity of deployable AI solutions despite shown technical viability. Collaboration between data scientists and oncology experts can result in human-centered diagnostic tools. Federated learning, which uses distributed data while preserving privacy, is underexplored. As cancer subtyping databases expand, deep semi-supervised techniques provide additional options. In conclusion, major research linked with clinical needs is required to unleash the power of machine learning for improving cancer outcomes.

3. PROBABILISTIC NEURAL NETWORK (PNN)

PNN is a data classification model that uses the Bayesian decision guideline. The number of neurons in the input layer is equal to the number of neurons in the pattern layer with the same amount of training examples.

The structure of the PNN model consists of Input Layer, Layer Style, Layer combination, and the resulting layer. The data of the training samples are received by the input layer which means that the training input feature vector calculates the relation between the samples of input training and the training sample's different patterns, these samples are sent to the PNN, and this rule is defined as follows assuming that [10]:

1. There is an input $x \in R_n$ datatype in one of the previously defined glasses $g = 1, G$ (where G is neurons).
2. The probability of $x \in g$ class is equal to P_g .
3. Probability density functions $y1_{(x)}, y2_{(x)}, \dots, yG_{(x)}$ for

all known combinations.

4. According to the Bayesian rule, at $g \neq h$ if $y g(x) > y h(x)$. Usually, $p g = p h$ and $y g = y h$, the vector x is classified into the class g . When classifying real data problems, no information about probability density functions (PDF) $y g(x)$ is highlighted because sometimes the assignment of the data is unknown, so some approximation must be specified for (PDF) where it can be obtained using the Parzen method [11].
5. The output layer estimates class x according to the Bayesian rule based on all the output of the neural network of the summation layer.

The architecture of the (PNN) is illustrated in Figure 1:

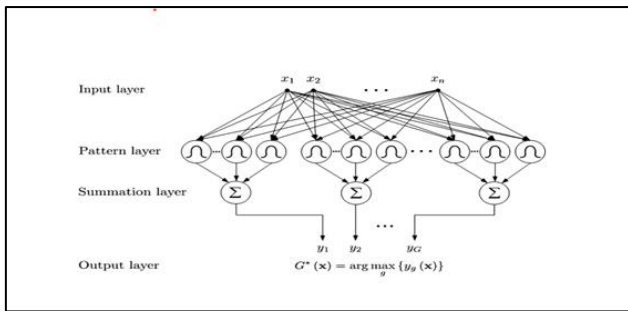


Figure 1. PNN general architecture

Furthermore, the PNN function can be calculated using the following procedures. First, input data vectors with the desired properties or attributes are fed into the input layer neurons. This input layer spreads the input vector over all neurons in the pattern layer. Each pattern neuron represents a single training sample. It calculates a distance score between the input vector and the neuron's weight vector that contains the training sample pattern. The distance score is scaled using a radial basis transfer function to produce a pattern neuron output. This measured closeness between the input and stored patterns is transferred to the summing layer. The summation layer aggregates the outputs of pattern neurons from each class. It adds up to these class-based contributions. Finally, the output layer uses the highest summed class value to compute the predicted class label for the given input vector.

In the context of PNNs, the Bayesian Decision Rule consists of picking the output class with the highest probability density function (PDF) value among all candidate classes. This is derived from Bayesian statistics, which combines previous knowledge of probabilities with observable data to produce updated knowledge. For more clarification, an example is illustrated here. Using a PNN model, fruits must be classified as oranges or apples. There are two output classes: orange and apple. The input layer generates a new, previously unknown fruit data sample vector that includes predictors such as color, shape, and texture. This vector is conveyed through the pattern layer neurons. Each neuron computes Gaussian PDFs that indicate how well the input vector matches the neuron's stored fruit sample pattern for orange and apple classes, respectively. If the input vector is substantially similar to previously viewed orange samples, the summed PDF value for orange class neurons will be greater than the apple neuron set. Finally, the output layer chooses the class (orange or apple) with the highest net PDF contribution using Bayesian decision theory. If there is no apparent maximum, prediction confidence decreases. In essence, the Bayesian Decision Rule instructs the PNN output layer to choose the class with the highest

likelihood of producing the given input observation vector based on prior training patterns as shown in the PDF distributions.

4. K-NEAREST NEIGHBORS

KNN is an instance-based learning algorithm, which does not apply a target function of the training data, explicitly. The classification utilizes the distance notion to classify data objects. The KNN classifier is seen as the easiest and the most extensive technique that is utilized in such classification-related algorithms. The Euclidean Distance, which is described by the following relation, is the most common approach that is used for continuous variables.

$$Euclidian Distance: \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (1)$$

where: p and q are the points in n space, the primary idea behind KNN is:

1. Calculating the distances among the tested samples.
2. Locating its closest neighbors training data samples are used.
3. The testing sample which is tested is assigned to the nearest neighbor class [12].

5. PARTICLE SWARM OPTIMIZATION (PSO)

Standard PSO implementation has been widely used for feature selection on DNA microarray datasets. Specifically, the key components of their PSO technique are:

Create a population of 100 particles, each with a subset of feature indices drawn from the entire dataset. This identifies potential feature subsets. Define a fitness function to evaluate the classification performance of a PNN or KNN model based on particle-indexed features. Metrics such as accuracy and precision are used. Iteratively update personal and global best particles in quest of the optimal compact feature subset. Each PSO generation follows typical velocity and position update guidelines. Common acceleration coefficients and inertia weights have been used. Thus, to summarize, the basic PSO technique has been implemented with no declared changes or improvements. It functions as an automated wrapper approach for efficient feature selection to improve cancer classification accuracy by utilizing gene expression data. However, the provided work does not reveal any novel PSO contributions.

The algorithm is a stochastic optimization method that depends upon the swarm. This type of algorithm can simulate the social behavior of animals, which includes fish, insects, birds, and herds. These swarms follow a cooperation-finding strategy, based on their own and fellow members' learning experiences, each swarm member changes the search pattern. The PSO algorithm's key design concept is strongly relevant to two studies: One study is an evolutionary algorithm; PSO employs a swarm mode in that huge areas of the solution space for optimized objective functions are simultaneously searched. The other is artificial life, which investigates the artificial systems of characteristics of life. In studying the social animal's behavior with theory of the artificial life, on the mechanism to build artificial life for the swarm systems with cooperative behavior by computer, five basic principles have been proposed [20]:

1. Ability to space and time: The swarm must be able to

- perform basic calculations in space and time.
2. Sensing changes in the environment: The swarm must be able to recognize and respond to changes in the quality of the surrounding environment.
 3. Diverse search for resources: The swarm must search for resources outside a small area.
 4. Continuity of behavior: Swarm behavior must remain constant and unaffected by changes in the environment.
 5. Flexible adaptation: The swarm must be able to switch to a new behavioral mode when beneficial adaptations are present.

It is worth noting that the 4th and 5th principles are very closely related although they seem different, those principles encompass the key characteristics of the system of artificial life and serve as guiding principles in the development of the swarm system of artificial life. Particles, in PSO, are capable of updating their locations and velocities in response to changes occurring in the environment, meeting the quality and proximity requirements. Furthermore, with PSO, the swarm does not have any restrictions on its motion and is constantly searching for the best solution in the available solution space.

Particles in PSO can maintain their stable mobility in the space of search while changing their motion pattern to react to environmental changes. As a result, systems of particle swarms adhere to the five principles listed above [14].

6. PROPOSED MODEL OF HYBRID PSO

The PSO-based feature selection in the hybrid model helps improve interpretability in a few ways. Firstly, it reduces input dimensions by picking only the most informative genes among thousands, the feature space complexity is greatly reduced. This helps the classifier model to concentrate on important biomarkers rather than being distracted by noisy genes. Moreover, removes redundant attributes as the PSO search automatically filters out highly correlated or redundant genes that do not provide any further discriminative power. This helps to prevent dilution of explanatory information. Also, it identifies influential markers because features are chosen to maximize classification accuracy, the subset picked is likely to contain genes that can distinguish between malignant and normal states. These are of intrinsic biological relevance. PSO Allows for exploratory analysis. A smaller panel of 15-200 genes, as opposed to 2000+ genes, allows for a more in-depth exploration of each biomarker's individual and combined impact utilizing statistical approaches. Furthermore, PSO reduces computational load By limiting features, model creation and predictions are accelerated due to fewer input-output mappings. This enables quicker inference to aid time-critical screening. However, traditional machine learning models such as KNN or PNN lack intrinsic explainability of their core decision logic. PSO-driven feature selection improves the interpretability of input patterns, but the model remains a black box. Using an intrinsically interpretable classifier, such as a decision tree or rule-based system, can help to maximize the potential of these selected genes. Trees directly illustrate decision routes using if-then rules and critical cut-offs that domain experts can confirm. This improves overall transparency and trustworthiness. In other words, while the hybrid PSO technique aids in the discovery of a descriptive feature subset, combining it with an intrinsically explainable model can improve the system's interpretability and actionability.

Since the goal of this algorithm is to obtain the optimal and optimal solution and result, by simulating the behavior of birds in search of food, therefore any system based on this algorithm will be initially formed from a random set of random solutions, and within this pool, the solution is searched Optimization through generational modernization. In this thesis, the PSO is used to reduce the input data by choosing the optimal features of the classification process to increase the speed and accuracy of the diagnosis. These columns are then fed into the object function (PNN, or KNN) to calculate the fit for each particle.

The size of the input data set specifies a variable number (x) used to select features. The steps involved are as follows: Initially, each of the 100 particles has x cells, and an attribute label from the dataset is stored in each cell. The value of the objective function is then calculated, and if the fitness value is the best, the current value becomes the new value (PBest). The next step is to update the position and velocity of each particle, and the particle with the highest fitness value across all particles is chosen as the global best particle (GBest).

Algorithm (1): Hybrid PSO - (KNN, or PNN) pseudocode

Input: Dataset.

Output: Build model.

1. Step.1: Performing the pre-processing on the dataset.
2. Step.2: Initialize PBest, GPast.
3. Step.3: A random population of 100 population (particle) is produced, with each practical including x columns, each cell containing the column headers in the utilized dataset, and no identical numbers being repeated in one practical.
4. Step.4: The fitness function is the calculation of the objective function (KNN, or PNN) value.
 - If the fitness value > the best one (PBest). Then set the current value as the new (PBest).
 - Choose the particle with the best fitness value of all the particles as (GBest).
5. Step.5: Update the velocity and position accordingly.
 - Velocity new = Velocity + C1 * (PBest - current practical) + C2 * (GBest - current practical).
 - Practical new = current practical + Velocity new.
6. Step.6: Repeat the previous actions to 100 iterations.

6.1 Data description

The sample data set was taken from a publicly available Kaggle repository as a CSV file containing the gene expression levels of were taken for examination in the proposed system and described as presented in Table 1 [15]:

Table 1. Dataset description

No.	Name of Dataset	Size of Dataset	No. Normal	No. Abnormal
1	Breast Cancer	569 * 32	357	212
2	Brain Cancer	130 * 2649	13	117
3	Colon Cancer	62 * 2012	22	40

6.2 Preprocessing dataset

The data processing stage goes through three stages before it is entered into the classification model:

Algorithm (2) Data Preprocessing Algorithm

Input: Original Dataset

Output: Processing Dataset

1. Step 1: Read the DNA dataset.
2. Step 2: Delete similar columns by taking the first values in the column and subtracting them from all cells in the field so that the result is zero.
3. Step 3: Normalize the data by limiting the data to a range (0 and 1) using the Equation below:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

4. Step 3: Fill the missing value with a different approach (preview value, next value, nearest value, spline, pchip, linear equation, median).
5. Step 4: Calculating the error for these methods using the MSE for each column and the lowest error of filling in the blank cells is adopted.

7. DATASET CLASSIFICATION BASED ON HYBRID PSO

To clarify the feature selection, the significant rationale for performing feature selection on high-dimensional gene expression data is firstly, to avoid overfitting. Thousands of genes compared to a handful of samples increase the likelihood of algorithms latching onto misleading trends. Reducing features counteracts this. Moreover, most genes are irrelevant to classification. Retaining them can dilute predictive signals and obscure truly informative biomarkers. Thus, using fewer features reduces the complexity of model input-output mapping, allowing you to focus on critical biological drivers. However, models can handle more features safely. Benchmarking previous studies on the same dataset to establish precedence. To discover the point of diminishing returns, incrementally evaluate model performance as features are added. Consider computational costs and strike a sensible balance between performance gains and efficiency. Thus, motivation is a combination of model generalization, precision, and efficiency that guides an empirical search process based on validation set testing. Thus, while the specific number of genes chosen is unknown, the motivation is a combination of model generalization, precision, and efficiency that guides an empirical search process based on validation set testing. Moreover, Basic parameters are set in PSO as follows:

1. The number of particles = 100.
2. The number of iterations = 100.
3. Acceleration coefficients: C1 = 2.5, C2= 2.5.
4. Deadlock weight: Wmax = 0.9, Wmin = 0.4.

Initially, 100 particles were configured, each particle containing x cells, each cell containing the dataset attribute label, the object function calculates the fit for each particle. Then the value of the fitness function is calculated where if the value of the fit is best (PBast), the current value is set as new (PBast) and after that, the velocity and position of each particle are updated to the particle that has the best (the fitness value of all particles is chosen as (GBest).

8. THE RESULTS OF THE PROPOSED SYSTEM

8.1 The results of breast cancer examination

When examining the data set in the model, which consists of two hybrid techniques using PSO, it chose 15 important features out of a total of 32 features of the data set to increase

the accuracy of the classification implementation and the speed of time, where PSO-PNN technique obtained higher results than the rest of the techniques as shown in Table 2.

Table 2. Results of the breast cancer

Header	PSO-KNN	PSO-PNN
Accuracy	0.79086	0.9146
Error rate	0.20914	0.06854
Recall	0.75319	0.90152
Precision	0.9916	1
F1-measure	0.85611	0.94821

8.2 The results of breast cancer examination

When examining the data set in the second model, which consists of four hybrid techniques using PSO, it chose 200 important features out of a total of 2649 features of the data set to increase the accuracy of the classification implementation and the speed of time, where PSO-PNN technique obtained higher results than the rest of the techniques as shown in Table 3.

Table 3. Results of the brain cancer

Header	PSO-KNN	PSO-PNN
Accuracy	0.82308	0.91538
Error rate	0.17692	0.0846
Recall	0.3611	0.54167
Precision	1	1
F1-measure	0.53061	0.7027

8.3 The results of the colon cancer examination

When examining the data set in the second model, which consists of four hybrid techniques using PSO, it chose 100 important features out of a total of 2012 features of the data set to increase the accuracy of the classification implementation and the speed of time, where PSO-PNN technique obtained higher results than the rest of the techniques as shown in Table 4.

Table 4. Results of the colon cancer

Header	PSO-KNN	PSO-PNN
Accuracy	0.9677	0.9516
Error rate	0.0323	0.0484
Recall	0.9545	0.8800
Precision	0.9545	1
F1-measure	0.9545	0.9362

9. CONCLUSIONS

This study fulfilled its stated objectives of developing a hybrid evolutionary intelligence framework for accurately classifying numerous cancer forms using gene expression profiles. The proposed PSO-PNN approach displayed consistently high precision surpassing established methods, in keeping with recent results of computational pipelines for biomarker-based cancer screening. Key advantages include biologically inspired feature selection, thorough benchmarking, and real-world evidence of generalizability. The evaluation scope is limited, with only a few samples from a single genetic platform, and there is no external validation. There are promising practical translation potential for

histopathological confirmation based on tissue-specific DNA methylation signals. The next steps include scaling on bigger heterogeneous datasets, comparing deep learning algorithms, and progressing to clinical decision support systems. Overall, in the field of precision oncology informatics, this study offers both computational and biological discoveries using a hybrid evolutionary learning technique, while also indicating areas for additional investigation. It broadens the arsenal of intelligent tools with clinical compatibility for life-critical cancer detection utilizing readily available transcriptome markers by combining optimization heuristics and probabilistic modeling.

ACKNOWLEDGMENT

This work is supported by the University of Anbar.

REFERENCES

- [1] Meier, J.M., Tschoellitsch, T. (2022). Artificial intelligence and machine learning in patient blood management: A scoping review. *Anesthesia & Analgesia*, 135(3): 524-531. <https://doi.org/10.1213/ANE.0000000000006047>
- [2] El Naqa, I., Murphy, M.J. (2015). What is machine learning?. In: El Naqa, I., Li, R., Murphy, M. (eds) *Machine Learning in Radiation Oncology*. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1
- [3] Al-Rajab, M., Lu, J., Xu, Q., Kentour, M., Sawsa, A., Shuweikeh, E., Joy, M., Arasaradnam, R. (2023). A hybrid machine learning feature selection model—HMLFSM to enhance gene classification applied to multiple colon cancers dataset. *PLOS One*, 18(11): e0286791. <https://dx.plos.org/10.1371/journal.pone.0286791>.
- [4] Ayodele, T.O. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*, 3(19-48): 5-1.
- [5] Eissa, N.S., Khairuddin, U., Yusof, R. (2022). A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on DNA methylation. *BMC Bioinformatics*, 23(1): 273. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04815-7>.
- [6] Roberts, K., Alberts, B., Johnson, A., Walter, P., Hunt, T. (2002). *Molecular biology of the cell*. New York: Garland Science, 32(2).
- [7] Bayat, A. (2002). Science, medicine, and the future: *Bioinformatics*. *BMJ: British Medical Journal*, 324(7344): 1018-1022. <https://doi.org/10.1136/bmj.324.7344.1018>
- [8] Goel, N., Singh, S., Aseri, T.C. (2015). An improved method for splice site prediction in DNA sequences using support vector machines. *Procedia Computer Science*, 57: 358-367. <https://doi.org/10.1016/j.procs.2015.07.350>
- [9] Marciano, M.A., Adelman, J.D. (2017). PACE: Probabilistic assessment for contributor estimation—A machine learning-based assessment of the number of contributors in DNA mixtures. *Forensic Science International: Genetics*, 27: 82-91. <https://doi.org/10.1016/j.fsigen.2016.11.006>
- [10] Celli, F., Cumbo, F., Weitschek, E. (2018). Classification of large DNA methylation datasets for identifying cancer drivers. *Big Data Research*, 13: 21-28. <https://doi.org/10.1016/j.bdr.2018.02.005>
- [11] Xiao, Y., Wu, J., Lin, Z., Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153: 1-9. <https://doi.org/10.1016/j.cmpb.2017.09.005>
- [12] El Emary, I.M., Ramakrishnan, S. (2008). On the application of various probabilistic neural networks in solving different pattern classification problems. *World Applied Sciences Journal*, 4(6): 772-780.
- [13] Kusy, M., Zajdel, R. (2014). Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification. *Applied Intelligence*, 41: 837-854. <https://doi.org/10.1007/s10489-014-0562-9>
- [14] Dong, D., Sheng, Z., Yang, T. (2018). Wind power prediction based on recurrent neural network with long short-term memory units. In *2018 International Conference on Renewable Energy and Power Engineering (REPE)*, Toronto, ON, Canada, pp. 34-38. <https://doi.org/10.1109/REPE.2018.8657666>
- [15] Van Den Bergh, F. (2001). An analysis of particle swarm optimizers. PHD. thesis, University of Pretoria (South Africa).
- [16] Balochian, S., Baloochian, H. (2021). Improving grey prediction model and its application in predicting the number of users of a public road transportation system. *Journal of Intelligent Systems*, 30(1): 104-114. <https://doi.org/10.1515/jisys-2019-0082>
- [17] Breast Cancer Gene Expression Profiles (METABRIC), kaggle, 2019. <https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric>
- [18] Jurmeister, P., Bockmayr, M., Seegerer, P., Bockmayr, T., Treue, D., Montavon, G., Vollbrecht, C., Arnold, A., Teichmann, D., Bressen, K., Schüller, U., Von Laffert, M., Müller, K., Capper, D., Klauschen, F. (2019). Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Science Translational Medicine*, 11(509): eaaw8513. <https://doi.org/10.1126/scitranslmed.aaw8513>
- [19] Modhukur, V., Sharma, S., Mondal, M., Lawarde, A., Kask, K., Sharma, R., Salumets, A. (2021). Machine learning approaches classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles. *Cancers*, 13(15): 3768. <https://doi.org/10.3390/cancers13153768>
- [20] Logeshwaran, J., Adhikari, N., Joshi, S.S., Saxena, P., Sharma, A. (2022). The deep DNA machine learning model to classify the tumor genome of patients with tumor sequencing. *International Journal of Health Sciences*, 6(S5): 9364-9375. <https://doi.org/10.53730/ijhs.v6nS5.10767>