Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P.7019-7033 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Hybrid Optimized Algorithm-Based Frequent Pattern Mining for GA-ANN Prediction Model on Genomes

E. Sheeba Sugantharani[1*] and Dr. A. Subramani[2]

[1]Department of Computer Science, Mother Teresa University, Kodaikanal, India
[1]Department of Computer Science, Lady Doak College, Madurai, India
[2]Department of Computer Science, M.V. Muthiah Govt. Arts College for Women, Mother Teresa University, Kodaikanal, India

How to cite this article: E. Sheeba Sugantharani, A. Subramani (2024). Hybrid Optimized Algorithm-Based Frequent Pattern Mining for GA-ANN Prediction Model on Genomes *Library Progress International,* 44(3), 7019-7033.

ABSTRACT
In genomic research, establishing a relationship among variables is usually of interest. Genomic research usually links factors. DNA microarray gene expression data introduces novel molecular biology and medical data analysis issues. In scientific data, Frequent Pattern Mining successfully applied association patterns discovered in microarray gene expression analysis. Discretization and association rule mining plays an important role in bioinformatics. Traditional pattern-finding approaches need many DNA sequence scans. The issue is recognizing intriguing patterns to make time-consuming and disagreeable judgments. The paper proposes a novel technique for identifying frequent patterns in DNA sequences and analyzing microarray gene expression profiling data. It first discovers frequent patterns from DNA sequences, then performs association rule mining and clustering discretization on the gene expression data using Fuzzy C-Means and PBMF index. A hybrid Diff-Eclat algorithm is employed to generate strong association rules and improves performance by microarray gene expression data's prediction. Finally, a GA-ANN (Genetic Algorithm-Artificial Neural Network) model is developed to predict biological knowledge from the discriminant rules. The method is implemented on a gene expression dataset and shows improved performance compared with SVM(Support Vector Machine), CNN(Convolutional Neural Network), RF(Random Forest) based on various metrics. The frequent patterns found in DNA sequence that were discovered during this approach have significant implications for medical data analyses like disease etiology, treatment analysis, mutation, and genetic analysis.

Keywords: Bioinformatic, Genomic research, DNA microarray, Gene expression analysis, Association rule mining, Clustering discretization.

I. INTRODUCTION

A massive amount of data has been accumulated in various fields due to network developments and the prevalent use of massive storage devices in big data [1, 2]. This information contains a wealth of knowledge and information. The main objective of bio sequence data mining is to identify basic functions, subsequences biological functions were predicted, and mutual functions and sequence interactions were identified [3, 4].

Bio sequence patterns typically reflect critical functional (or structural) components in bio sequence, like repetitive patterns [5]. Accordingly, bio sequence pattern mining is an important research area and technique for gene and protein fold recognition, sequence interaction explanation, and bio sequence functional prediction [5].The most important research object in bioinformatics was biological sequence data. These three types of sequences are included they are protein sequence, RNA sequence, and DNA sequence [6].

Protein Sequence Analysis is the method of studying a protein or peptide sequence using one of a variety of analytical techniques [7]. RNA-seq (RNA-sequencing) is a technique that examines the RNA sample quantity and sequences using next-generation sequencing (NGS) [8]. DNA is famously composed of four different types of nucleotide bases, and the precise sequence of these bases along DNA determines the individuality of an individual gene [9]. In the evolution process, a more repetitive sequence was produced due to the gene's replication [10]. This process contributes to newgene production and it is also very important in the study of genetic variation [11]. Diseases caused by mutations in repetitive sequences include thymus hypoplasia syndrome, Williams syndrome, and muscle atrophy [11].

Microarray technologies offer a powerful tool for tracking numerous genes for their expression patterns simultaneously, with applications ranging from cancer diagnosis to drug response [12]. Gene expression is the transcriptional transformation of DNA sequences into mRNA sequences, which are then converted into amino acid sequences known as proteins. The biggest challenge with microarray data is the high density of data. The information gathered from Microarray experiments was frequently conducted in the form of an expression-level matrix. And there is always a research question of whether the repetition of sequence of gene patterns or change in occurrence of sequence patterns contribute for particular genetical disorders.

To overcome those issues, initially, frequent patterns are discovered in the DNA sequence. Patterns that occur repeatedly in a data set are known as frequent patterns [13]. Discovering frequent patterns is critical in mining associations, correlations, and other interesting data connections. Then the profiling data of microarray gene expression was analyzed using association rule mining and clustering discretization was carried out. In its most fundamental form, utilizing machine learning methods, association rule mining examines data for patterns or co-occurrences in a database. Clustering discretization methods are utilized to transfer a continuous function into a discrete function with known solution values at all points in space and time. Finally, a strong association rule is generated that enhances the performance. DNA sequence and Frequent Patterns found during this research have significant implications for mutation, genetic analysis, diagnosing genetical disorders caused by mutations in repetitive sequences include thymus hypoplasia syndrome, Williams syndrome, and muscle atrophy [14].

The major contribution of this research is as follows,

- Initially, discovering frequent patterns from DNA sequences.
- Then, association rule mining and clustering discretization are carried out by using the Fuzzy C-Means PBMF index for analyzing microarray gene expression profiling data.
- Next, to improve the performance a hybrid Diff-Eclat method is proposed which generates a strong association rule.
- At last, to predict biological knowledge, GA-ANN is used.

The remainder of this research is split into the sections listed below: Section 2 discusses data mining works. Section 3 describes the concept of the proposed algorithm. Section 4 assesses the effectiveness of the proposed scheme and discussion. At last, section 5 ends with the outcome and future objectives.

II. LITERATURE REVIEW

Iqbal [15] presented a novel genetic algorithm-based feature reduction method to rectify the problems of scalability. To improve accuracy and scalability, an integrated model that connects the gap between lexicon-based and machine-learning strategies was proposed. Using this hybrid model, the feature set's size was reduced by up to 42% without impacting accuracy. Principal component analysis (PCA) and latent semantic analysis were utilized to compare the feature reduction methods (LSA). This method indicates a higher accuracy of up to 15.4% over PCA and a higher accuracy of up to 40.2% over LSA.

Farheen [16] developed a hybrid method to predict the node's probable location called the multi-parameter spatial temporal modeling method. For enhancing routing performance without increased packet overhead, A multi-path routing protocol with path diversion at critical points along the proposed path was established based on estimated probability locations. The observed results show that the suggested solution was discovered to have a higher packet delivery ratio than traditional methods.

Li [17] conducted studies on the support vector machine (SVM) method, which combines three heuristic algorithms, such as genetic algorithms, particle swarm optimization, and slap swarm algorithm (SSA) in which the fiber-reinforced CPB's strength was predicted. The test outcomes revealed that polypropylene fibers boost CPB strength, whereas straw fibers reduce CPB strength in some circumstances.

Ai [18] surveyed on a variety of association rule mining algorithms used on high-dimensional datasets. Based on the previous studies, the algorithm's relative metrics and their main characteristics are described. To make the adaption to the high dimensional dataset better, the optimization area and the improved areas are pointed out. In general, association rule mining algorithms that integrate multiple optimization methods with advanced computer methods can enhance scalability and interpretability.

Hemeida [19] made research to investigate the evolutionary optimization algorithms (EOAs) implementation in machine learning by employing four distinct optimization techniques for mining two well-known data sets. The suggested optimization algorithm was evaluated by using the selected dataset such as the breast cancer dataset and the Iris dataset. The neural network (NN) is employed in this paper's classification problem as well as four optimization techniques: dragonfly algorithm (DA), multiverse optimization (MVA), grey wolf optimization

(GWO), and whale optimization algorithm (WOA).

Azimi [20] developed a new method that predicts the peak particle velocity named a genetic algorithm used to optimize a new hybrid evolutionary artificial neural network. The suggested GA-ANN proposes a systematic and automated approach for selecting an appropriate ANN architecture, which contains the number of epochs, training algorithm, activation functions, and several neurons. To evaluate the suggested method, among monitoring and blasting stations available at the Sungun Copper Mine site in Iran, a data set consisting of radial distance (RD), horizontal distance (HD), maximum charge weight per delay, and a new modified radial distance (MRD) was used.

Cai [21] developed a hybrid rating-based many-objective recommendation technique to make diverse and accurate results by optimizing the recommendations such as recall, diversity, accuracy, and novelty simultaneously. To improve the many-objective evolutionary algorithms (MaOEAs) and the effectiveness of the model's suggestions, a novel generation fitness evaluation method and a partition-based knowledge mining method are also proposed.

Zhang [29] introduced a novel hybrid optimization algorithm CSDE on the basics of cuckoo search CS and differential evolution DE that resolves the constrained engineering issues. Both of these algorithms [29] are highly appropriate for problems with engineering. This study separates the population into two groups and employs CS and DE for every subgroup separately. Because of the numerous design variables and constrained engineering circumstances, because a single optimizer was not able to fulfill the precision requirement, the use of hybrid optimization algorithms (such as CSDE) appears to be the most promising approach for completing this work.

III. PROPOSED METHODOLOGY

Mining association rules is a crucial method of data mining for several applications today and this method is employed on a microarray dataset to derive attractive relations among gene sets. Associations and correlations among items are discovered by frequent itemset mining in big transactional. The association rule mining’s goal is, to extract frequent patterns by utilizing the intervals of gene expression. In this proposed work the Gene expression of data is converted from continuous value to discrete value before frequent pattern mining and they are replaced by the gene intervals. The suggested method uses association and clustering rules along with gene intervals on gene expression data’s frequent pattern mining and they are depicted in Figure 1.

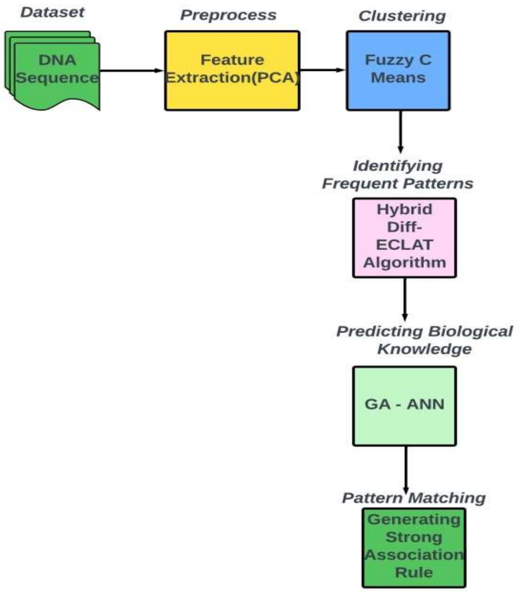
Dataset
DNA Sequence
Preprocess
Feature Extraction(PCA)
Clustering
Fuzzy C Means
Identifying Frequent Patterns
Hybrid Diff-ECLAT Algorithm
Predicting Biological Knowledge
GA - ANN
Pattern Matching
Generating Strong Association Rule

Fig. 1. Framework of proposed methodology

A. Dataset Description

The gene expression dataset [22] comprises training (38 samples) and independent (34 samples) datasets from a 2022 study by Golub et al. It focuses on classifying acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) cases through gene expression monitoring using DNA microarrays. The measurements are derived from bone marrow and peripheral blood samples, with re-scaled intensity values. Golub's study demonstrated the potential of gene expression monitoring for cancer classification, providing a foundation for identifying and categorizing diverse cancer classes based on molecular characteristics.

B. DNA sequence concept definition

(1) The DNA sequence contains four letters in series they are P, Q, R, and S. This sequence is represented as C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 }, in which, ci € {P, Q, R, S} ($0\leq j \leq n-1$).

(2) (DNA sequence length): For C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 } for DNA sequence, n represents the DNA sequence length. sequence is given as C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 }, in which, ci € {P, Q, R, S} ($0\leq j \leq n-1$).

(3) (DNA's sub sequence): A fragment in a DNA sequence is the sub sequence. For this sequence C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 }.For DNA sequence C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 }, if there is a fragment$C^{'} = \{c_a, c_{a+1}, \dots \dots, c_{a+b}\}$satisfying $a \geq 0$, $b \geq 0$, and $a + b \leq n - 1$,then$C^{'}$ is designated as a DNA subsequence of C.

(4) It is a DNA super sequence., if a DNA sequence exists C= {c0; c1; c2; ; ; ; ; ; ; ; cn-1 }, C_1' and C_2' are two sub sequence of C, $C_1' = \{c_a, c_{a+1}, \dots \dots, c_{a+b}\}$, $C_2' = \{c_{a'}, c_{a'+1}, \dots \dots \dots \dots \dots .., c_{a'+b'}\}$. While $a' \leq a$

And $a + b \leq a' + b'$, then the super sequence of c_1' is c_2'.

(5) The C corresponding DNA pattern is the DNA pattern for the DNA sequence C.

(6) DNA pattern length, DNA pattern L its length is the number of letters in L.

(7) The sub pattern DNA, while a subsequence of C is $C^{'}$, DNA pattern's sub pattern $C^{'}$ corresponding to C.

(8) DNA super pattern, for DNA sequence C, $C^{'}$ is the super sequence, the corresponding DNA pattern C's super pattern

(9) DNA pattern's prefix, the DNA pattern's length L is greater than 1, then L values prefix is a substring that eliminates the last letter. The length of the DNA pattern is one, and there is no prefix. L's prefix is shown as prefix (L).

(10) DNA pattern's postfix, the DNA patterns length L is greater than 1, then the post fix of L is the substring that removes the first letter. The length of the DNA pattern 1 does not have a prefix. The L is the postfix that is depicted as postfix(L).

(11) Joinable, L1 and L2 are two pattern's that are joinable if the following constraints are satisfied:

(a) Length (L1) = length (L2) =1, L1 and L2 are joinable.

(b) Length (L1) = length (L2) >1, L1 and L2 are joinable if postfix(L1) = prefix (L2).

(12) String concatenation, for two patterns L1 and L2 string concatenation, is depicted as L1+L2, here L1 and L2 are sequential concatenations.

(13) Joining of two patterns, L1 and L2 are two patterns that are joinable, the joining of L1 and L2 are depicted as join (L1;L2), those values is based on various conditions such as

(a) If Length (L1) = Length (L2) =1 then, join (L1;L2) = L1+L2, here L1 and L2 are string concatenations.

(b) If Length(L1) = Length (L2)>1, then join (l1;L2) = L1 +(L2-prefix(L2)), here L1 string concatenation and L2 is the last letter.

(c) When Length(L1)=Length (L2)=1, L1 and L2 are joinable, the result is L1+L2; here L@ and L1 are also connectable, so the results joined is L2+L1.

C. DNA position information

(14) The DNA patterns position L in sequence C is the index set for all L occurrences in C, and it is depicted as $position_{L;C}$.

In the DNA sequence, the first letter's index is specified as 0.

(15) +n set, if set D= $\{d_i\}$, satisfies $1 \leq i \leq |D|$, next, +n set of D is referred as D^{+n}= $\{d_i + n\}$,

In this any natural number is denoted as n.

(16) -n set, -n set, if set D= $\{d_i\}$, satisfies $1 \leq i \leq |D|$, next, +n set of D is referred as D^{-n}= $\{d_i - n \,|d_i - n \geq 0\}$, in this any natural number is denoted as n.

D. DNA sequence scanning

DNA sequence scanning analyzes a DNA molecule's nucleotide sequence in detail. Both computational and experimental approaches are used in genomics, molecular biology, and bioinformatics. Computational DNA sequence scanning uses advanced algorithms to find patterns, motifs, and functional components[23]. This may require identifying genes, regulatory areas, binding sites, or DNA variants. Large-scale genomic investigations require computer methods to effectively sort genetic data. Experimental DNA sequence scanning uses DNA sequencing. This approach examines a DNA molecule's full nucleotide sequence, revealing its genetic composition. DNA sequencing helps discover genetic variants and mutations by revealing nucleotide order. DNA sequence scanning aims to read the genetic code, comprehend gene functions, find regulatory elements that govern gene expression, and detect disease-associated genetic variants. This approach advances customized medicine, genetic diagnostics, and genetic engineering by helping us grasp DNA's complicated genetic information. The combination of computational and experimental DNA sequence scanning methods advances genomics and molecular research. To identify entire frequent patterns, only one scan for the DNA sequence is performed. Every letter's location in the DNA sequence is also positioned at the same time. The location of the data is stored in another simple hash table. Figure 2 depicts the DNA letter-position structure plotted in a hash table.

Key (DNA letter)	Value (Position of DNA letter)

Fig. 2. Position of DNA letter structure

E. Preprocessing

Preprocessing is carried out before frequent pattern mining, in this each genome or sequence is processed. Initially, the dataset's letters are replaced by numbers. Consequently, when the length of the DNA sequence is long, in this sequence time and memory consumption are reduced. The sequence is divided into blocks and each consists of an equal number of bases. This preprocessing is not like the FPE (Format-Preserving Encryption) method because even a single base sequence is not discarded while blocking species sequence patterns. The block length is selected.

F. Gene Clustering Employing Fuzzy C Means and PBMF Index

For gene clustering, similar and dissimilar features are identified largely by employing fuzzy c-means cluster analysis[24]. In a fuzzy clustering algorithm, a similar expression pattern is used to divide genes. Additionally, various clusters show diverse, expression patterns that are well-separated. Because genes belong to more than one cluster in fuzzy clustering, it is possible to identify genes that are conditionally co-regulated or co-expressed. In the following way, acted genes beyond one transcription factor are identified and multifunctional proteins are encoded. The PBMF index is used to evaluate the results of gene clusters. On the contrary side, the PBMF-index stands for fuzzy PBM-index utilizing fuzzy c-means soft clustering analysis. The PBMF-index design assures reduced resulting cluster numbers. These processes are stated briefly and are shown below. The objective function minimization Zm supports a large part of FCM:

$P_n = \arg \min_{[y_i \in cluster j]} \sum_{i=1}^{M} \sum_{j=1}^{D} v_{ij}^{n} y_i - D_j^2$ (1) Here, n is an actual number higher than 1, $V_{ij}$is the membership degree of y_i , y_i is the i^th^ gene expression in d dimensions, D_j is the gene center for d-dimensional clusters, and ||*|| expresses the similarity among centers and any measured levels of gene expression. Iterative optimization is used to achieve fuzzy partitioning of P_n, with $V_{ij}$and $D_j$updated by:

$$V_{ij} = \left(\sum_{a=1}^{D} \left(\frac{y_i - d_j}{y_i - d_a} \right)^{\frac{2}{n-1}} \right)^{-1} \quad (2)$$

$$d_j = \frac{\sum_{i=1}^{M} V_{ij}^{n} \cdot y_i}{\sum_{i=1}^{M} V_{ij}^{n}} \quad (3)$$

In FCM cluster analysis, the quality is verified using PBMF-index and it is referred to as three factors' products. The maximization product ensures that the partition contains a small number of compact clusters separated by a large distance among at least two clusters. The PBMF index is numerically denoted as follows:

$$U_{PBMF}(A) = \left(\frac{1}{A} \cdot \frac{F_1}{W_n} \cdot C_a \right)^2 \quad (4)$$

Here, K denotes the number of gene clusters. The VPBMF(A) factors follow the actual notations and are described as follows. The factor F_1 is the total amount of each sample's distance from the entire center d_0. This factor is calculated independently of the number of clusters and is as follows:

$$F_1 = \sum_{i=1}^{M} y_i - d_0 \quad (5)$$

In the FCM algorithm, the P_n factor is similarly processed. Here, D_K denotes the greatest any two gene clusters are separated:

$$C_A = \max_{1 \le i,j \le A} D_i - D_j \quad (6)$$

The PBMF-index optimization is dependent on the lower cluster number, the minimized measure of P_n, and the greater evaluation of D_K.

FCM groups together genes with similar data structures, co-expressed levels, or proximity, and the PBMF index guarantees a lower cluster number.

G. Frequent Pattern Mining

For the given set of elements $J = \{j_1, j_2, j_3, \ldots\ldots, j_n\}$ and $S = \{s_1, s_2, s_3, \ldots\ldots, s_m\}$ represents the set of transactions, frequent is a subset of J⊆ $, if support(T)≥ minimum support, here, user-defined threshold provides minimum support.

H. Eclat Algorithm

The ECLAT algorithm is a depth-first search algorithm. It employs a vertical database design, which means that rather than explicitly displaying all transactions; every item is placed together along with a cover called titlist, and utilizes the intersection-based method to determine the item set's support. If there is a small item set, then this algorithm is better than the Apriori algorithm because it needs only less space. It is better suited to small datasets and takes less time to generate frequent patterns than the Apriori algorithm. There are numerous governed data mining methods, one of which is affiliation rule mining. ECLAT is a depth-first pursuit calculation that employs a set crossing point. It is a generally rich calculation appropriate for both effective and similar execution with location upgrading qualities. To accomplish itemset mining, the ECLAT calculation is used. Itemset mining enables us to identify continuous examples in data. The fundamental concept behind the ECLAT calculation is to utilize Tid set crossing points to perform the assistance of an applicant itemset while avoiding the age of subsets that do not exist in the prefix tree. ECLAT is an information mining method that was designed for market bin analysis. Following this set mining aims to identify procedures in the shopping habits of grocery store customers, mail-order companies, and online retailers. It tries to differentiate sets of items that are frequently purchased together here. Once identified, related item arrangements are used to enhance the offered items association on the grocery store racks or in the list of mail-request pages or web shop, and it suggests which items are packed differently, or allows to recommend different items to customers. ECLAT is based on two fundamental advancements: applicant age and pruning. During the applicant age step, each n-itemset applicant has created from two regular n-1 itemsets and after that, its help is tallied; if its help is less than the limit, it is discarded; otherwise, it is repeated thing sets and used to produce n+ 1 itemset. Because ECLAT employs an upward design, checking support is limited. The age of the applicant is undoubtedly a hunt in the inquiry tree. The basic thing with 1-itemset is a similarity class with the prefix, and this similarity class is similar to the fundamental transfer database in vertical format. Due to its profundity first inquiry, ECLAT does not fully exploit the descending summary property.[30]

Algorithm 1: Pseudocode for Eclat Algorithm

```
Input: $DE((s_1, u_1), \ldots (s_n, u_n) \,|Q);\ r_{min}$
Output: $DF(DE, r_{min})$
For all $s_j$ occurring in DE do
    Q:= Q ∪ $s_j$
    Init (DE') // initialize a new equivalence class with the new prefix Q
    For all $s_k$ occurring in DE such that k>j do
        $u_{tmp} = u_j \cap u_k$
        If $|u_{tmp}| \geq r_{min}$ then
            $DE' := DE \cup (s_k, u_{tmp})$
            $DF = DF \cup (s_k \cup Q)$
        $end\ if$
    $end\ for$
    $if\ DE' \neq \{\,\} then$
        $Eclat(DE', r_{min})$
    $end\ if$
```

```
$end\ for$
```

I. Hybrid Diff-Eclat Algorithm

The diffset arrangement (the difference between two sets) has significantly reduced the Eclat algorithm's run time and memory usage. A diff-Eclat algorithm is essentially the Eclat algorithm with a diff set arrangement. It is similar to Eclat, except that in Diff-Eclat, tid sets are arranged in ascending, and diff sets are arranged in descending depending on their size. Diff-run Eclat's time and memory usage is reduced by sorting diffsets and tidsets. The Diff-Eclat algorithm's pseudo code is shown below:

Algorithm 2: Pseudocode for Hybrid Diff-Eclat algorithm

```
Input: $DE((s_1, u_1), \dots (s_n, u_n) \,|Q);\ r_{min}$
Output: $DF(DE, r_{min})$
Sort $s_j$ in descending order, with frequency reference.
For all $s_j$ occurring in DE do
    Q:= Q ∪ $s_j$
    Init (DE') // initialize a new equivalence class with the new prefix Q
    For all $s_k$ occurring in DE such that k>j do
        $u_{tmp} = u_j \cap u_k$
        If $|u_{tmp}| \geq r_{min}$ then
            $DE' := DE \cup (s_k, u_{tmp})$
            $DF = DF \cup (s_k \cup Q)$
        $end\ if$
    $end\ for$
    $if\ DE' \neq \{\ \} then$
        $Eclat(DE', r_{min})$
    $end\ if$
end for
```

J. Development Of GA-ANN Prediction Model for Biological Knowledge

Biological knowledge is required to extract efficient and significant patterns from discriminant rules to reveal fatal and critical causes of diseases[25]

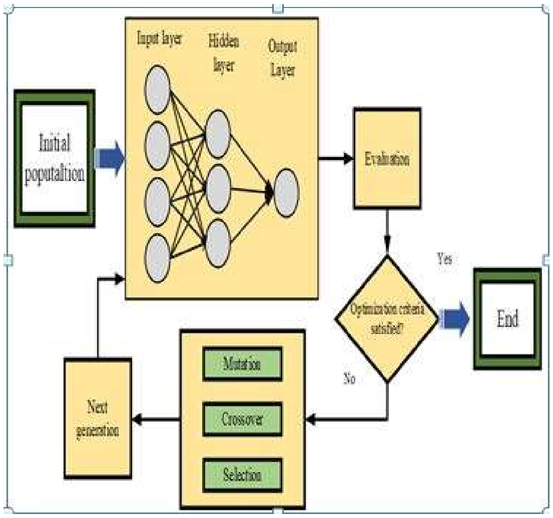

Fig. 3. Architecture of GA-ANN

Following the acquisition of the DEG list, the ANN architecture is built in PYTHON, the input variable is the expression levels of up- and down-regulated probes, and the output variable is the microarray sample diagnosis.A training set of microarray samples was created at random, and the remaining microarray samples represented a control set. The ANN model is composed of three layers: an input layer with n nodes, and an output layer with one node[26]. Then the mean square error's threshold and the maximum recursive time are set. Tansig is used as the transfer function between the input and hidden layers, and the weight-corrected learning rate is utilized, whereas purelin is set up as a transfer function between the hidden and output layers. In this, ANN is trained by using GA[27]. The initial number of populations and the maximum evolutionary generation are set for optimization by GA. GA-ANN has chosen useful input variables at random mostly during every round of computation, guaranteeing consistent computational accuracy. Time-series analysis, RNNs, hidden Markov models, CNNs, and differential equations can simulate dynamic biological processes like RNA-seq data temporal variations. These approaches reveal temporal relationships, hidden states, and complicated patterns, improving our knowledge of dynamic biological events.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In the proposed technique, the dataset is selected. Initially, for the selected DNA sequence dataset, frequent patterns are identified. Then, for analyzing the gene expression, the Fuzzy C-Means PBMF index is used, in which association rule mining and clustering discretization are carried out. Next, to improve the performance a strong association rule is generated. Finally, GA-ANN is used for predicting biological knowledge. For this purpose, the Gene expression dataset [22] is utilized. Their performance is compared with other state-of-arts. The implementation is done on the PYTHON platform.

A. PERFORMANCE METRICS

- **Accuracy:**

To calculate accuracy, divide the total number of forecasts by the number of correct predictions.

$$Accuracy = \frac{(Tp\ +Tn)}{(Tp+Tn+Fp+Fn)} \qquad (7)$$

- **F-measure:**

The F-measure is determined by combining precision and recall.

$$F - Measure = \left(\frac{(2\times Preci\times Recal}{Preci+Recal}\right)(8)$$

- **MCC:**

MCC is a correlation coefficient with four values: True-Positive (Tp), False Positive (Fp), True-Negative (Tn), and False Negative (Fn).

$$MCC = \left(\frac{((Tp\ \times Tn)-(Fp\times Fn))}{\sqrt{((Tp+Fp)(Tp+\quad)(Tn+F\ \ )}(Tn+Fn)}\right) \qquad (9)$$

- **FPR**

The number of times a test result was determined as the ratio of data that was incorrectly defined as positive to data that was correctly defined as negative is referred to as the "false positive rate.

$$FPR = \left(\frac{Fp}{Fp+Tn}\right) \qquad (10)$$

- **FNR**

FNR is an abbreviation for false-negative rate. FNR denotes the proportion of false negative categorization to all data categorization.

$$FNR = \left(\frac{Fn}{Fn+T}\right) \qquad (11)$$

- **Negative Prediction Value (NPV)**

NPV describes the performance of a diagnostic test or other quantitative metric.

$$NPV = \frac{Tn}{Tn+Fn} \qquad (12)$$

- **Precision:**

It calculates the ratio of truly predicted positives to total expected positives.

$$Precision = \left(\frac{Tp}{Tp+Fp}\right) \qquad (13)$$

- **Sensitivity:**

The sensitivity measure assesses and predicts the model's performance in each category.

$$Sensitivity = \left(\frac{Tp}{(Tn+Fp)}\right) \quad (14)$$

- **Specificity:**

The specificity metric evaluates the model's ability to forecast true negatives in all categories.

$$Specificity = \left(\frac{Tn}{(Tn+Fp)}\right) \quad (15)$$

B. PERFORMANCE ANALYSIS

In this section, the experimental results of Microarray gene expression analysis utilizing the Fuzzy C-Means PBMF index and a hybrid Diff-Eclat method are carried out. Table 1 explains the sample data for microarray

Table I. Sample Data for Microarray

Gene	G1	G2	G3	G4	G5
S1	-0.8615	-0.0331	-0.3517	-0.80574	-0.16842
S2	-0.16772	1.0374	0.13913	0.87657	0.1146
S3	0.415047	1.35855	-2.4959	1.10088	-0.90791
S4	-0.13072	0.48876	2.42972	-1.46141	-1.30048

The sample microarray gene expression dataset was used, where each column denotes a gene and each row denotes a sample. The Fuzzy C-Means PBMF index is used to transform the gene expression values into discretized values. Where k represents the number of clusters, and k is supplied by the user. The user passes the k value 2 in this experiment.

Gene expression values are replaced by gene intervals after clustering and discretization. Table 2 displays the experimental results of discretization processes.

Table II. Gene Values are Discretized

Gene/ Sample values	S1	S2	S3	S4
G1	[-0.26 1.46]	[0.53 -0.19]	[0.72 1.85]	[0.23 0.62]
G2	[-0.67 -1.16]	[-2.67463]	[2.23 -0.83 -0.48]	[-0.52 1.26 -0.60]
G3	[-0.27 1 .40]	[0.09 -0.36]	[1.91 2.07]	[0.19 0.97]
G4	[-2.28 0 .04 0.41]	[-2.42168]	[-2.11 1.25 -0.47]	[1.15 0.62 0.93]

The discretized gene expression data was transformed into transactional data. TIDs depict transactions, and itemsets represent gene expression values with gene intervals. Table 3 shows a sample transaction dataset. The Diff-Eclat algorithm is then used to discover frequent patterns by association rules.

Table III. Frequent Item sets are grouped

S.No	Itemset	Support
0	{[0.29, -0.06, -0.43, -0.40, 0.19, 1.15, 0.18]}	0.028409
1	{[2.22, -0.16]}	0.020834
2	{[-0.50, -0.24, 0.24, 0.70, -1.47, -0.56, -0.10}}	0.019157
3	{[0.73, 0.35, -0.32]}	0.046594
4	{[-0.02, -0.13, -0.69, -0.32, 0.03, 1.11, 1.06]}	0.043632
5	{[-0.89, -0.36]}	0.042658
6	{[0.26, 1.02, 0.27, -0.39]}	0.026056
7	{[-2.16, -0.80, -1.49, -0.53, -1.25, -0.48, 0.05, -0.44, -0.43]}	0.017983
8	{[0.49, -0.10, -0.46, -0.43, -0.82, 1.28, 0.85, 0.96, -1.33, -0.95, -0.54, -0.11, -1.68, -0.28]}	0.019123
9	{[0.40, -0.21, 0.78, 0.71, -0.11, 0.26, 0.58, 0.12, -0.004, -1.26, -0.72, -0.80, -0.69, -0.24, -0.46]}	0.013779

At last, identify the discriminatory rules that ensure a minimum level of support and confidence. Table 4 displays the significance association rule. Table 5 displays the outcomes of discriminant association rules. Their performance is analyzed and compared with several existing methods such as SVM, CNN, and RF.

S.No	Rules	Support
0	[-1.86 0.24 -0.19 0.085 -2.60 -2.37 -2.07] → [2.57 1.34 2.23]	0.021434
1	[-0.83 1.46]→ [-2.68 -1.17]	0.014281
2	[-0.01 0.18 0.68 2.08 2.09 0.02 0.14]→ [-0.27 -0.39]	0.021078
3	[-0.67 -0.87 -0.40]→ [-0.72 -1.16 -0.54 -1.14]	0.035986
4	[-0.73 1.66 0.55 -0.14, -0.27 0.17 -0.46]→ [0.10 0.46]	0.012746
5	[-0.50 -0.48]→ [0.47 0.48]	0.014487
6	[-0.89 -0.36, -0.77 -0.03]→ [-1.30 -0.36]	0.042455
7	[1.45, -0.23]→ [-0.60 -0.93]	0.032008
8	[-1.02 -0.23 4.25 0.12 0.16 -0.86, -1.36 -0.86 -0.95 -0.90 0.90 -2.16 1.38-0.82]→ [0.62 0.50]	0.015007
9	[-1.66 -0.69, 0.08 0.002 -0.38 -1.31 -0.95 -2.55, 0.31 -1.27 0.20 0.57 0.39 -1.03 -1.06]→ [0.86 1.01 0.46]	0.02047

Table IV. Performance Metrics Comparison

Performance metrics	Hybrid Diff-Eclat (Proposed)	SVM	CNN	RF
Accuracy	0.892898	0.858507	0.819254	0.833201
Precision	0.885548	0.851450	0.784749	0.874365
Sensitivity	0.900407	0.865736	0.861117	0.789276
Specificity	0.885489	0.851374	0.777946	0.876545
F-measure	0.865077	0.831767	0.794625	0.805259
MCC	0.938811	0.896376	0.824741	0.825152
NPV	0.900396	0.865707	0.860585	0.799748
FPR	0.019576	0.026218	0.024581	0.020108
FNR	0.003431	0.005614	0.006585	0.004163

In Table 5, the suggested Hybrid Diff-Eclat algorithm and the existing techniques SVM, CNN, and RF performance comparison are compared based on several performance metrics such as precision, accuracy, specificity, sensitivity, NPV, F-measure, MCC, FPR, and FNR. Performance demonstrates that Hybrid Diff-Eclat algorithm (Proposed) outperforms SVM, CNN, and RF.

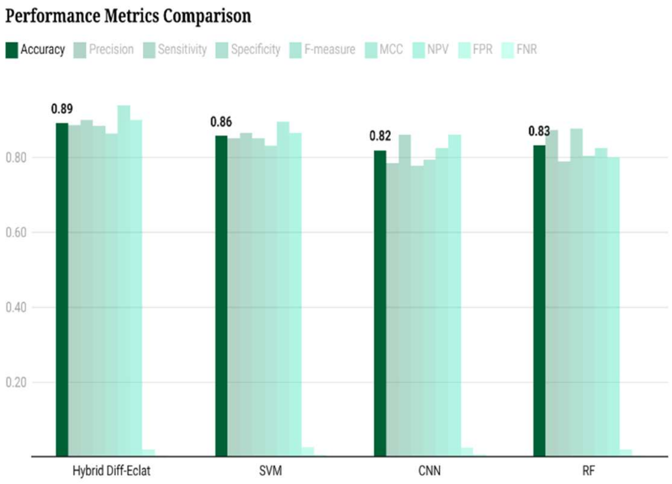

Fig. 4. Performance Metrics Comparison - Accuracy

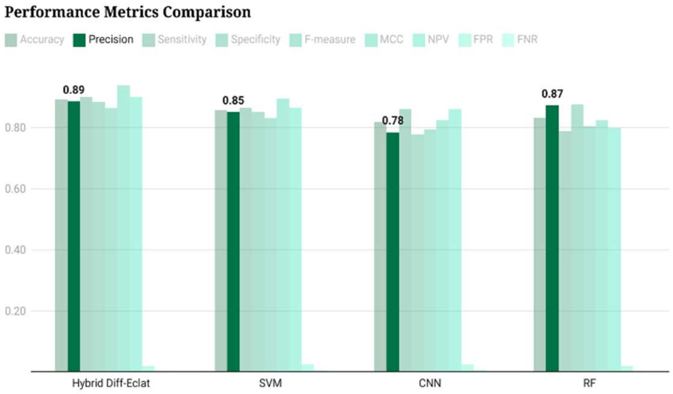

Fig. 5. Performance Metrics Comparison - Precision

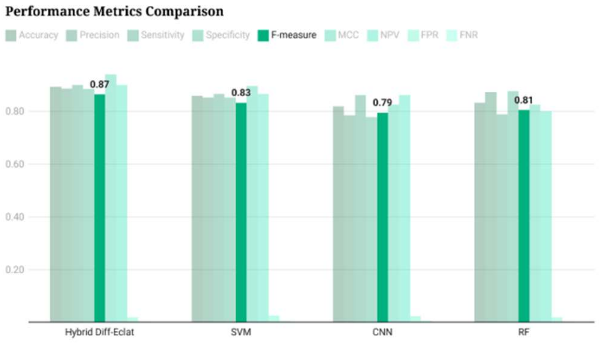

Fig. 6. Performance Metrics Comparison –F- Measure

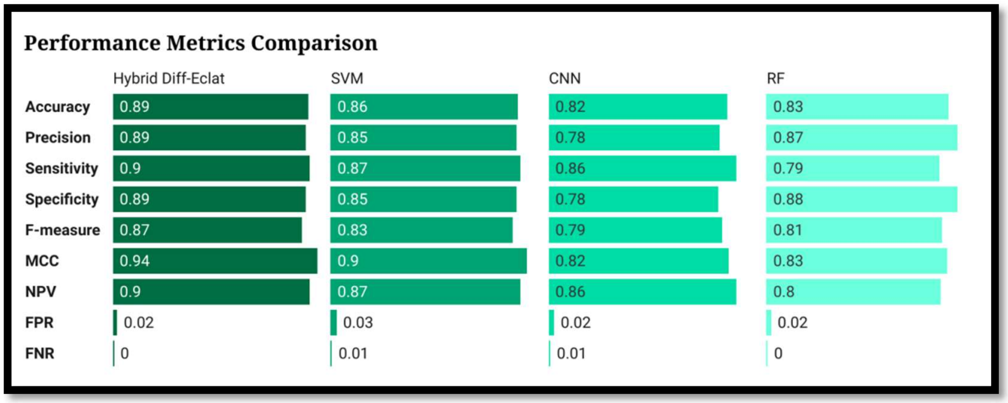

Fig.7. Performance metrices of Accuracy, Precision, Sensitivity, Specificity, F-Measure, MCC, NPV, FPR, FNR

C. *COMPARATIVE ANALYSIS*

The suggested approach has 0.892898 accuracy, greater than the previous methods' 0.858507, 0.819254, and 0.833201 demonstrated in Figure 7. F-measure performance metrics compare proposed and current approaches. The suggested technique yields 0.865077 F-measure, greater than 0.831767, 0.794625, and 0.805259. Comparison of suggested and current FNR approaches on performance. Comparison demonstrates that the suggested method's 0.003431 FNR is lower than the existing methods' 0.005614, 0.006585, and 0.004163 FNR. Comparison of planned and current specificity approaches. Comparison demonstrates that the suggested technique has 0.885489 specificity, greater than 0.851374, 0.777946, and 0.876545. Performance demonstrates that Hybrid Diff-Eclat algorithm outperforms SVM, CNN, and RF.

D. *DISCUSSION*

t-SNE and PCA visualize genomics models. Researchers use SHAP and integrated gradients to interpret models' latent representation. Biological validation comes from pathway analysis and data comparison. Model performance can be improved using biologically inspired synthetic data[28]. Sharing biological representations without patient data is conceivable using synthetic data. Biological similarity, data variety, and model flexibility determine model transferability. Validation, ethics, and domain adaptability are crucial to model implementation.

Synergy explains rules without language or visuals. For complicated genetic model assumptions, rule mining contextualizes gradient characteristics. This technique streamlines genetic data processing and shows intricate relationships for better decision-making.Explore activation properties and idea weighting to explain model rationales beyond quantitative measures.

Mining and deep learning are used in hybrid approaches. Transfer learning with pre-trained models enhances generalization. The methods find non-linear correlations in complicated genomic and biological datasets. Various ways assess genomic model interpretability. These approaches preserve complicated model structure patterns. To improve interpretability, researchers assess genetic feature relevance, information transfer, and model behavior. Counterfactuals and statistics boost genetic and bioinformatics disclosure and trust. Rule mining can be subtle with gradients or deep model activation patterns. Rule mining gradients and interpretability inform model conclusions. Validated and contextualized models predict biology broadly and precisely. Require extensive cross-validation and external validation.

Bioinformatics and genomics use t-tests, ANOVA, correlation studies, and enrichment tests for computational predictions. This result demonstrates how model features activate, stressing their decision-making importance. Analyze major idea weighting to demonstrate the model's attention on essentials. This qualitative analysis provides greater insights than numerical measurements on how the model understands and prioritizes input.

V. CONCLUSION

A novel method for detecting frequent patterns in DNA sequences was suggested in this research. Here, a novel method was described for analyzing microarray gene expression profiling data using clustering discretization and association rule mining. Fuzzy C-Means and the PBMF index are utilized throughout discretization to transform gene expressions into gene expression intervals and discrete data, and common patterns are found by mining association rules to discover significant relationships among microarray genes. To develop strong association rules, a Hybrid Diff-Eclat algorithm is used. It enhances the prediction of microarray gene expression data and evaluates the algorithm's sensitivity, precision, specificity, performance coefficient, and error rate. The overall performance demonstrates that the proposed technique performed better. In future research, to overcome computational time as well as memory explosion issues with gene expression datasets the proposed algorithm has to be enhanced. Numerical performance improvement data in the conclusion are essential for comparing progress to previous efforts. Accuracy (0.892898), precision (0.885548), and F1 score (0.865077) demonstrate model efficacy. Making improvements to computational efficiency, training time, or other benchmarks supports the conclusion. This quantitative data highlights the model's superiority and shows development in genomics and bioinformatics research, boosting its legitimacy and influence.

Conflicts of Interest
The authors declare no conflict of interest.

Author Contributions: The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, writing review and editing, visualization, have been done by 1st author. The supervision and project administration have been done by the 2nd author.

REFERENCES

[1] Liu, W., Kong, C., Niu, Q., Jiang, J., computer-integrated, X. Z.-R, (2020) “A method of NC machine tools intelligent monitoring system in smart factories,” Elsevier. Robotics and Computer-Integrated Manufacturing, Volume 61, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2019.101842.

[2] Zhang, Z., Ding, S., Intelligence, W. J.-E. A. of A., (2019) “A hybrid optimization algorithm based on cuckoo search and differential evolution for solving constrained engineering problems,” Elsevier. Engineering Applications of Artificial Intelligence, Volume 85, Pages 254-268, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2019.06.017.

[3] Lv, Z., Ding, H., Wang, L., Q. Z.,(2022) “A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome,” Elsevier. Neurocomputing, Volume 422, 2021, Pages 214-221, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.09.056.

[4] Zhang, H., Li, Y., Lv, Z., … A. S.-I. J. (2020), “A real-time and ubiquitous network attack detection based on deep belief network and support vector machine,” IEEE/CAA Journal of Automatica Sinica, vol. 7, no. 3, pp. 790-799, May 2020, doi: 10.1109/JAS.2020.1003099.

[5] Mittempergher, L., Delahaye, L., A. W.-T. J. of M.,(2019) “MammaPrint and BluePrint molecular diagnostics using targeted RNA next-generation sequencing technology” Elsevier The Journal of Molecular Diagnostics vol 21(5), pp. 808-823, https://doi.org/10.1016/j.jmoldx.2019.04.007

[6] Sivanathan, A., Gharakheili, H., … F. L.-I. T., (2019), “Classifying IoT devices in smart environments using network traffic characteristics” IEEE Transactions on Mobile Computing, vol. 18, no. 8, pp. 1745-1759, doi: 10.1109/TMC.2018.2866249.

[7] Borisov, N., Sorokin, M., Garazha, A., and, A. B.-N. acid detection,(2020) “Quantitation of molecular pathway activation using RNA sequencing data” Springer. Nucleic Acid Detection and Structural Investigations: Methods and Protocols, pp. 189-206, https://doi.org/10.1007/978-1-0716-0138-9_15

[8] Butt, A., Rasool, N., reports, Y. K.-M. biology, (2018) “Predicting membrane proteins and their types by extracting various sequence features into Chou’s general PseAAC” Springer, Molecular biology reports, vol. 45(6), pp. 2295-2306, https://doi.org/10.1007/s11033-018-4391-5

[9] Hopman, R., & M’charek, A. (2020) “Facing the unknown suspect: Forensic DNA phenotyping and the oscillation between the individual and the collective” Springer BioSocieties, vol.15(3), pp. 438-462, https://doi.org/10.1057/s41292-020-00190-9

[10] Lee, H., Zhang, Z., & Krause, H. M. (2019). “Long noncoding RNAs and repetitive elements: junk or intimate evolutionary partners?” Elsevier. TRENDS in Genetics, 35(12), pp. 892-902, https://doi.org/10.1016/j.tig.2019.09.006

[11] Wekesa, J., Meng, J., & Luan, Y. (2020) “Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction” Elsevier. Genomics, vol 112(5), pp. 2928-2936, https://doi.org/10.1016/j.ygeno.2020.05.005

[12] Menyhárt, O., & Győrffy, B. (2021). “Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis”, Elsevier Computational and structural biotechnology journal, vol 19, pp. 949-960. https://doi.org/10.1016/j.csbj.2021.01.009

[13] Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., ... & Soranzo, N. (2020). “The polygenic and monogenic basis of blood traits and diseases,” Elsevier Cell, 182(5), 1214-1231. https://doi.org/10.1016/j.cell.2020.08.008

[14] Jyonouchi, S., Graham Jr, J. M., & Ming, J. E. (2020). “Genetic syndromes with evidence of immune deficiency.” Elsevier In Stiehm's Immune Deficiencies, second edition (pp. 61-97). Academic Press. https://doi.org/10.1016/B978-0-12-816768-7.00002-8

[15] Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). “A hybrid framework for sentiment analysis using genetic algorithm based feature reduction.” IEEE Access, 7, 14637-14652. https://doi.org/10.1109/ACCESS.2019.2892852

[16] Farheen, N. S., & Jain, A. (2022). “Improved routing in MANET with optimized multi path routing fine tuned with hybrid modeling.” Elsevier Journal of King Saud University-Computer and Information Sciences, 34(6), 2443-2450. https://doi.org/10.1016/j.jksuci.2020.01.001

[17] Li, E., Zhou, J., Shi, X., Jahed Armaghani, D., Yu, Z., Chen, X., & Huang, P. (2021). “Developing a hybrid model of salp swarm algorithm-based support vector machine to predict the strength of fiber-reinforced cemented paste backfill.” Springer. Engineering with Computers, 37, 3519-3540. https://doi.org/10.1007/s00366-020-01014-x

[18] Ai, D., Pan, H., Li, X., Gao, Y., & He, D. (2018). “Association rule mining algorithms on high-dimensional datasets.” Springer. Artificial Life and Robotics, 23, 420-427. https://doi.org/10.1007/s10015-018-0437-y

[19] Hemeida, A. M., Alkhalaf, S., Mady, A., Mahmoud, E. A., Hussein, M. E., & Eldin, A. M. B. (2020). “Implementation of nature-inspired optimization algorithms in some data mining tasks.” Elsevier Ain Shams Engineering Journal, 11(2), 309-318. https://doi.org/10.1016/j.asej.2019.10.003

[20] Azimi, Y., Khoshrou, S. H., & Osanloo, M. (2019). “Prediction of blast induced ground vibration (BIGV) of quarry mining using hybrid genetic algorithm optimized artificial neural network.” Elsevier Measurement, 147, 106874. https://doi.org/10.1016/j.measurement.2019.106874

[21] Cai, X., Hu, Z., & Chen, J. (2020). “A many-objective optimization recommendation algorithm based on knowledge mining.” Elsevier Information Sciences, 537, 148-161. https://doi.org/10.1016/j.ins.2020.05.067

[22] Dataset based on https://www.kaggle.com/datasets/crawford/gene-expression, Golub et al "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" 2022-10-18.

[23] Robson, B. (2020). “The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target.” Elsevier Computers in Biology and Medicine, 125, 103963. https://doi.org/10.1016/j.compbiomed.2020.103963

[24] Abualigah, L., & Dulaimi, A. J. (2021). “A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm.” Springer Cluster Computing, 24, 2161-2176. https://doi.org/10.1007/s10586-021-03254-y

[25] Robson, B. (2020). “Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus.” Elsevier Computers in biology and medicine, 119, 103670. https://doi.org/10.1016/j.compbiomed.2020.103670

[26] Veeranjaneyulu, K., Lakshmi, M., & Janakiraman, S. (2024). “Swarm Intelligent Metaheuristic Optimization Algorithms-Based Artificial Neural Network Models for Breast Cancer Diagnosis:” Springer Emerging Trends, Challenges and Future Research Directions. Archives of Computational Methods in Engineering, 1-18. https://doi.org/10.1007/s11831-024-10142-2

[27] Zhao, Z., Bao, Y., Gao, T., & An, Q. (2024). “Optimization of GFRP-concrete-steel composite column based on genetic algorithm-artificial neural network.” Elsevier Applied Ocean Research, 143, 103881. https://doi.org/10.1016/j.apor.2024.103881

[28] Qiu, Y., Zhou, J., He, B., Armaghani, D. J., Huang, S., & He, X. (2024). “Evaluation and Interpretation of Blasting-Induced Tunnel Overbreak: Using Heuristic-Based Ensemble Learning and Gene Expression Programming Techniques.” Springer Rock Mechanics and Rock Engineering, 1-29. https://doi.org/10.1007/s00603-024-03947-x

[29] Zhang, Z., Ding, S., Intelligence, W. J.-E. A. of A., & 2019. A hybrid optimization algorithm based on cuckoo search and differential evolution for solving constrained engineering problems. Elsevier. Retrieved November 15, 2022, from https://www.sciencedirect.com/science/article/pii/S0952197619301563

[30] Kumari P., (2024). Predicting the Severity of Diabetes Using ECLAT Algorithm in Data Mining. Springer nature. https://doi.org/10.2991/978-94-6463-433-4_26