Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SIMILARITY-BASED GENE DUPLICATION PREDICTION IN PROTEIN-PROTEIN INTERACTION USING DEEP ARTIFICIAL ECOSYSTEM NETWORK

SRINATH DOSS[1], JOTHI PARANTHAMAN[2], VINSTON RAJA R[3], JOHN ANAND G[4]

Professor, Faculty of Engineering and Technology, Botho University, Botswana.[1]

Lecturer, Faculty of Engineering and Technology, Botho University, Botswana[2]

Assistant Professor, Department of Information Technology, Panimalar Engineering College[3]

Fellow, Faculty of Engineering and Technology, Botho University, Botswana[4]

E-mail: srinath.doss@bothouniversity.ac.bw[1] ,jothi.paranthaman@bothouniversity.ac.bw[2], rvinstonraja@gmail.com[3]

john.anand@bothouniversity.ac.bw[4]

ABSTRACT

In the living organism, almost entire cell functions are performed by protein-protein interactions. As experimental and computing technology advances, yet more Protein-Protein Interaction (PPI) data becomes processed, and PPI networks become denser. The traditional methods utilize the network structure to examine the protein structure. Still, it consumes more time and cost and creates computing complexity when the system has gene duplications and a complementary interface. This research uses gene expression patterns to introduce a deep artificial ecosystem for gene duplication counting and cancer cell prediction. The main objective of this research is to predict the MYC proteins influence level, which is in charge of controlling cell growth and death in gene expression of lung cancer. Small body parts are responsible for these protein interactions, which are crucial for understanding life's activities. To achieve the research objective, a similarity-based clustering approach is employed for gene duplication counting, and Artificial Ecosystem Optimizer based Minimal Gated Recurrent Unit network (AEOMGRU) network-based approach is introduced to predict the cancer gene patterns. The proposed models' efficiency is compared to recently develop bio-inspired optimizer deep neural network techniques such as GAANN, PSOANN, and classic GRU. The efficiency of the proposed classifier shows the highest concerning the performance metrics weight average accuracy ratio of 99.08%, average precision rate of 99.2%, least root mean square error of 0.2%, and least mean absolute error of 0.5%.

Keywords: *Protein-Protein interaction, MYC Protein structure, Clustering, Gene duplication counting, Lung cancer, Minimal GRU network*

1. INTRODUCTION

Protein-protein interactions PPIs [1] are physical contacts of high specificity established between two are more protein molecules as an effect of biochemical reaction steered by interactions that include hydro bounding electrostatic forces and hydrophobic effect [2-3]. The PPIs play a significant role during the gene duplication process, and some of the influence of the protein to create disease in human beings MYC, FGFR1and ERBB2 [4-6]. Therefore, identifying

the PPIs is important for understanding the mechanism of life activity. The below figure 1 shows the structure [7] of MYC locus (top), gene (red line), and MYC protein organization (bottom). It shows that a variety of proteins interacts and regulates the MYC protein's activity. The Protein produces 439 amino acids. Gene duplication [8] is a key process that creates a new genetic organism during molecular development. The replication mechanism has been performed in many ways, such as the paralogs gene, orthologous gene, and d analog gene [9-10]. The paralogs [11] replicates genes within the same species.

Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

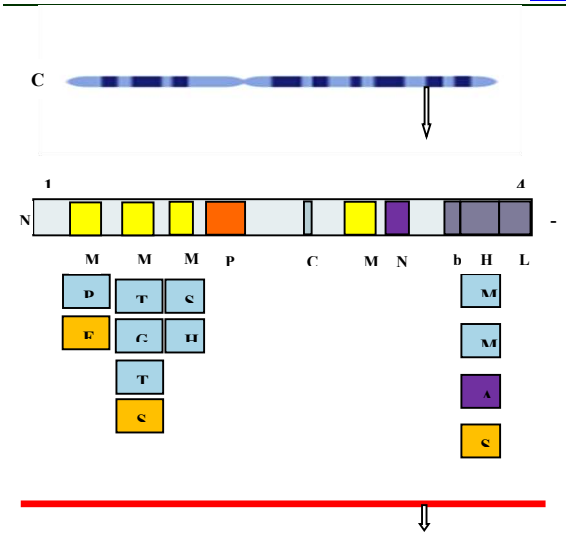

Fig. 1 General Structure of MYC Protein Interaction

In contrast, orthologous [12] replicates genes after the speciation event from the same parent gene. In analog replication, a gene with similar functions and characteristics and presents in different species is known as an analog gene [13]. The diagrammatical representation of this gene replication is shown in figure 2.

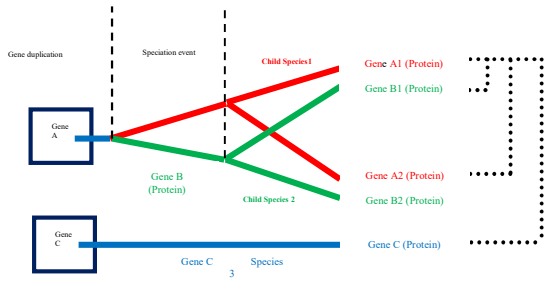

Fig. 2 General Mechanism of Gene Duplication

The replications of on cogenes [14] play a major role in the formative tithe of many cancers. In contrast, some abnormal proprotein applications counter many types of cancers, according to the earlier research [1he, MYC, CCND1, and ERBB2 proteins are replicated 20%, the FGFR1 and FGFR2 are replicated 12% in breast cancer cells. The HRAS, KRAS, and MYB proteins are replicated 30%, 20, and 15-20% respectively in colorectal cancer cells. The CCNE, KRA, S, and MET proteins are replicated 15%, 10%, and 1,0%, respectively, in gastric cancer [16-18]. The MYC, ERB,B2, and AKT2 proteins are replicated 20-30%, 15-30%,and 12% in ovarian cancer. According to earlier studies, the MYC protein is identified as an influential protein to cause cancers like lung cancer, breast cancer, gastric cancer, ovarian cancer, and so on [19-21]. Deep neural networks are effective in various fields, and their use in computational biology is growing by the day [22]. Several studies have employed deep learning algorithms to predict PPI labels [23-25].

As a result of the background analysis, this research has primarily focused on the MYC protein's gene expression pattern duplication counting based MYC gene influence in lung cancer gene expression patterns identification and introducing a classification tool for predicting cancers causing gene expression patterns and a clustering tool for gene duplication counting. There are currently only two patterns for gene clustering that can be reset or updated, whereas this manuscript proposes three patterns. This study uses a deep artificial ecology network to quantify gene duplications and forecast cancer cells based on gene expression patterns. A similarity-based clustering methodology is used to count gene duplications in this framework. An Artificial Ecosystem Optimizer-based Minimal Gated Recurrent Unit network (AEOMGRU) network-based technique is used to forecast cancer gene patterns. The main objective of this research is to predict the MYC proteins influence level, which oversees controlling cell growth and death in gene expression of lung cancer. Small body parts are responsible for these protein interactions, which are crucial for understanding life's activities. However, the advantage of AEOMGRU over existing methods is that the proposed method can model a collection of more efficient records and each pattern can be assumed to be dependent on previous ones.

The research paper has been organized in the following manner; section 1 describes the general introduction of the problem definition, section 2 details on the related research works, section 3 brief about the methodologies used for gene duplication counting and predicting cancers causing gene expression patterns, section 4 summarizes the evaluation results and discussions of the proposed approach. Finally, section 5 confers the conclusion of the research findings.

2. RELATED WORKS

The related research part discussed the earlier research on gene duplication problems, protein-protection problems, cancer protein identification problems and was utilized to support this research. Lai et al. developed a silicon docking approach to predict protein-protein interaction among 16 BdMAPKs and 86 BdPP2C2s in B. Furthermore, the prediction accuracy of the approach is investigated

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with docking site 3D protein structures. It reveals that the docking site also predicts 96 pair imparities two proteins and the. The fancy of the prediction approach is also evaluated with a staring database, which obtained a less false positive rate. Jiang et al. created a framework to predict patterns of repeated gene evaluation in duplicate genes. It analyzes genomes of 90 different eukaryotes and predicts the number of protein families' significant functional differentiation during gene duplication. Moreover, it can attribute about 6% of recurrent sequence evaluation between Paralogs. Dunk and Snel constructed a duplicate gene D, gene lossL, and horizontal gene transfer T-based DTL framework to count the evolution history of a gene family. It came to randomly produce histories for a specified size of two dissimilar species, the rooted caterpillar and complete binary tree. It can also compute the range of exponential growth she numbers of histories of random species trees size do 25. The evaluation results prove that the horizontal gene transfers in a dramatic increase in the amount of history. Chauve and Ponty presented mathematical models for genomic duplication problems. It provides ANN algorithmic ANN approach to solve the minimum episode ME clustering problem. It also combines the first linear time and space algorithm for the ME clustering problem at any interval. It is also generalized to allow every evolution scenario. Paszeka and Gorecki introduces a technique to predict the disruption of specific protein interactions in cancer patients using somatic mutation data and protein interaction networks. It uses a smoothing approach to score for edge nodes in the interaction network, which is used to qualify the proximity of each edge to somatic mutation in individual samples. Ruffalo and Bar-Joseph developed a tool to predict interacting prologs between the two protein families, maximizing the detectable co-evolutionary signals. Furthermore, this approach is generalized to predict on genomic co-localization of a gene coding for interacting proteins. Gueudre et al. developed a once-protein-protein interaction identification protocol. It observes the enhanced sensitivity of STK11 silenced lung cancer cells to the FDA-approved CDK4 based on the STK11-CDK4 connectivity. The OncoPPI approach is focused on finding the PPI resources that link cancer genes into a signaling network for predicting the tumor vulnerabilities for therapeutic examination. Li et al. presented similarity approach for PPIs network from the perspective of proteins complementary interface and gene duplication to improve the prediction accuracy. Chen et al. introduced a novel nature-inspired meta-heuristic, an artificial ecosystem-based optimization algorithm. This approach mimics the three unique behavior of living organisms such as production, consumption, and decomposition. The efficiency of the new optimizer is evaluated with benchmark optimization functions for eight real-world engineering problems. The evaluating result shows that the new optimizer outperforms than comparison algorithms. Zhao et al. Applied gated recurrent unit network model for wireless intrusion detection problem. The performance of the network classifier is tested with the NSL-KDD dataset. Also,a comparison has been performed with Artificial Neural Network, Feed Forward Neural Network, Long Sort Term Memory, Random Forest and Naive Bayes. The evaluation result shows that the GRU classifier obtained 99.35% validation accuracy, which is s maximum accuracy rate comparison approaches. Kasongo and Sun introduced a bio-inspired deep classifier, which integrates the genetic algorithm with the artificial neural network GAANN [37]. It replaces the two worst solutions for a population with two solutions for each population already stored in ANN. The classifier is designed to improve the performance as well as to reduce the computation time. Jose Anand et al. developed a recommended system for preference prediction in a multi-criteria recommendation system, which is utilized particle swarm optimization (PSO) to train ANN. The PSOANN integrates the multi-criteria rating information system and determining the preferences of users. Hamada and Hassan introduced an adaptive evolutionary algorithm for predicting negative linkages from PPI networks, optimized using the Minimum Weak Edge-Edge Domination (WEED) set. The approach could increase the quality of PPI data, according to the encouraging results achieved on the MINT dataset. Izudheen presented the dataset generation through Negatome, Random pair, and Recombine pair approaches were examined at three degrees of development. The N-Gram methods were used to accomplish feature extraction and feature selection. Support Vector Machine, Decision Tree, Neural Network, and Naive Bayes classifiers were used in ensemble classification and evaluation. The Genetic-PSO method offered an improved optimization technique represented through the search operation. Three network alignment algorithms based on distinct ideas were proposed in the reference [41]. They scored a PPI network alignment using sequence data, network topology,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and subnetwork module data. They then used efficient methods (heuristics and convex optimization) to generate alignments by maximizing the alignment scores. Ge et al. introduced a multi-level model LPPI to increase large-scale PPI accuracy and speed of large-scale PPI prediction. They created a weighted network by calculating node similarity using protein characteristics. Then, by lowering the size of the weighted graph, Graph Zoom was employed to speed up the embedding process. The rebuilt graph was then used to understand graph topology properties using graph embedding methods. Furthermore, the chance of two proteins interacting was predicted using the linear Logistic Regression (LR) model. Su et al, proposed a genetic algorithm based on community detection for feature selection and compare the efficacy of the proposed strategy. Rostami et al. presented classifying the data narrowed the possible values and eliminated ambiguity.The new method's performance was compared to that of well-known and state-of-the-art semi-supervised feature selection approaches on eight datasets. Rostami et al., the author proposed a graph embedding method called ExEm that uses dominating-set theory and deep learning approaches to capture node representations. The extracted expert embeddings can be used in various ways. A novel strategy uses expert vectors to calculate experts' scores and recommends experts extend these embeddings into the expert recommendation system.Nikzad-Khasmakhi et al. examined the accuracy-efficiency trade-off for various structured model pruning methods and datasets (CIFAR-10 and ImageNet) on TPUs using the VGG-16 model as an example (TPUs) and demonstrated that structured model pruning could significantly reduce model memory usage and speed on TPUs without compromising accuracy, particularly for small datasets.

Most of the research works discussed earlier in this section focus on partial parts alone, whether gene duplication prediction problems or protein-protein interaction prediction problems. All the above-related works failed to integrate similarity-based gene duplication prediction efficiently and the deep learning-based artificial ecosystem optimizer. GANN is only expressed in a particular tissue type and the early stages of development, whereas MYC is found throughout the body. MYC, for example, is concentrated in the newborn mouse's forebrain, kidney, and hindbrain, but it is absent from nearly all the animal's tissues in adults. Sun et al.discussed that coexist in mutualistic endosymbiosis with the roots of most vascular plants as Arbuscular Mycorrhizal (AM) fungi, which belong to an early branching fungal lineage called Glomeramycotina. With the inability of the fungi and plants to work together, they need to exchange phosphorus for carbon. This helps the plant's nutrition and ecosystem productivity. It is only through the exchange of signals with the root's cortical cells that they can grow into the root cortex and then differentiate into branching arbuscules. Skinnider et al. presented the arbuscules are housed in apoplastic compartments within the cortical root cells and are responsible for exchanging nutrients with the host.To combat the coming worldwide catastrophe of antibiotic resistance, we urgently require new antibiotics. Medical secondary metabolism has long been the principal source of clinically useful antibiotics. Bioinformatics has found many natural antibiotics that are still unidentified. Witch weeds, a group of parasitic plants of the genus Striga, is a major cause of crop loss in Sub-Saharan Africa and a global threat to agriculture. Due to a paucity of genetic information, it's been difficult to understand Striga parasite biology, which could lead to agricultural remedies. During the development of Striga's haustorium, genes involved in lateral root development were found to be co-opted, suggesting a partially co-opted pathway during the evolution of the haustorium. Yoshida et al. presented that most species have smaller genomes than expected based on polyploidy prevalence in their lineages, suggesting selection for genome shrinking. However, when ancestral GS is compared to the occurrence of ancestral polyploidy, it appears that DNA loss after polyploidy was extremely minimal.After polyploidy, selection may favor genome downsizing due to two hypotheses: reducing the cost of nucleic acid synthesis in both the nucleus and the transcriptome I by decreasing nitrogen (N) and phosphate (P) in the nucleus, and (ii) by reducing the effect of GS scaling effects on cell size, which affects CO2 uptake and water loss.The performance metric of many approaches demands improved results in protein-protein interaction prediction. Still,MYCresearches are considered as incomplete without interlinking concepts and in this research utilizes after protein-protein interaction the MYC protein's gene expression pattern duplication counting based MYC gene influence in lung cancer gene expression patterns identification has been mainly focused in this research to resolve the research gab in the earlier studies and also introduce a classification tool to predicting cancers causing gene expression patterns and a clustering tool for gene duplication counting.

Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The following steps archive the research motive. The lung cancer gene expression patterns are taken to count the gene duplication percentage using Similarity-based clustering. Anticancer gene patterns are utilized for training and predicting cancer by thee using the AEOMGRU network. The subsequent section describes the methodologies are used to archive the objective of this research.

3. PROPOSED METHODOLOGY

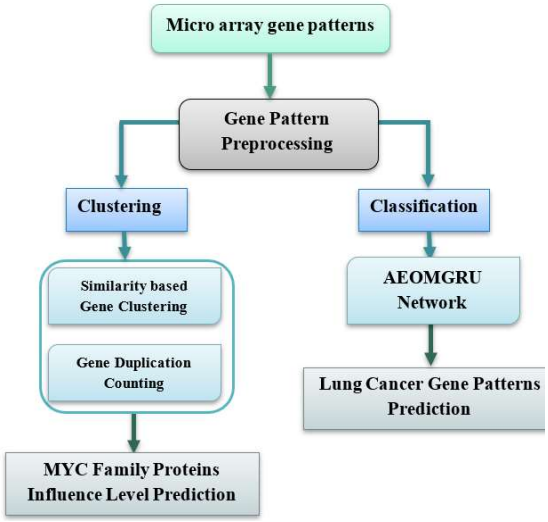

Fig. 3 Proposed Methodology Flow diagram

Figure 3 illustrates the workflow of the research work; it offers two different approaches functionality to solve the oncogene gene and protein prediction. Initially, the Gene patterns are collected from data sources, and the gene patterns are taken for preprocessing. The preprocessed gene patterns are taken as input to the clustering algorithm, the similarity approach utilized to form a gene cluster, and the clustered genes are ranked to count the gene duplication. The MYC proteins family influence levels in lung cancer genes are identified. The second approach is AEOMnetwork-based based lung cancer gene pattern classification. The detailed descriptions of the approaches are explained in subsequent sections.

Data source

This section discusses the data sources utilized to evaluate the efficiency of the classification and clustering tool. The MYC protein is identified as an influential protein to cause cancers like lung cancer, breast cancer, gastric cancer, ovarian cancer,etc. Therefore, the MYC protein's gene expression pattern dataset has been taken from this research's publicly available NCBI [43] database. The dataset contains a high volume of gene expression patterns with irrelevant information. During the preprocessing, the irrelevant details are removed before processing. The gene patterns are represented in text format. The classification model is trained with 85% of gene patterns and the remaining 15% of gene patterns taken to test network performance.

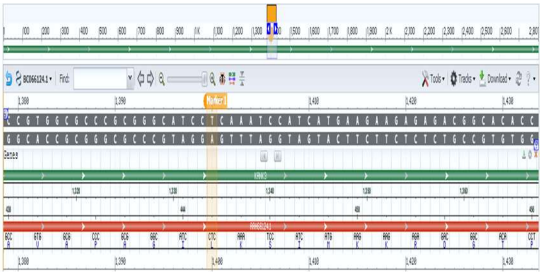

Fig. 4 Sample Genome Pattern and Data Viewer

Protein structure is determined by the sequence of amino acids and by local, low-energy chemical bonds between atoms in the polypeptide backbone and amino acid side chains. The structure of a protein is critical to its function; if a protein loses its shape structurally, it may no longer be useful. Figure 4 illustrates the genome data viewer, which contains details of cancer-causing proteins gene expression patterns, gene names, and location in the chromosome, etc. The clustered gene patterns are compared with the help of this information,an online genome data viewer tool offered by NCBI. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39.

Data Preprocessing and Normalization

The gene expression patterns collected from microarray have come in the text; therefore, processing the text data directly in any machine learning algorithm is possible. So the collected gene pattern is normalized using the bag of the word (bow) approach. The output values of the bow function have been normalized in standard form by using the following derivation,

$$GP_{ij} = normalize\ (GP_{ij})\ GP_{ij} = rand(1,4);$$
$$GP_{ij} = GP_{ij} * 20; \quad (1)$$

The above eq.(1) is used to normalize the converted gene pattern. This normalized gene vector is utilized for cluster-based gene duplication counting and AEOMGRU network-based cancer gene pattern

Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

prediction. The Similarity-based gene clustering for duplication count is described in the subsequent sections.

Similarity-based gene clustering for Duplication Count

The gene duplication counting approach contains two steps first is similarity gene clustering and ranking-based duplication counting.

Similarity-based gene Clustering

Clustering is an essential process in gene duplication counting. During the clustering process, similar patterns are clustered and ranked before gene duplication counting. The mathematical representation of gene similarity [44] based clustering is given as follows,

$$sim(\widehat{GP}, \hat{C}) = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|\widehat{GP}_{ij} - \hat{C}|}{\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{GP}_{ij} + \hat{C})} \quad (2)$$

Where$\widehat{GP}$Denotes gene features and$\hat{C}$Denotes the gene cluster's centroid value. Idenote the position of gene pattern, and j denotes the position of gene feature. The$\widehat{GP}_{ij}$Denotes the ith gene pattern'sjth feature value, and $sim(\widehat{GP}\hat{C})$ is used to calculate Similarity among gene patterns and a cluster centroid. if a gene value and centroid values get maximum Similarity, then that gene points can belong to that group, calculated using the eq. (2). Mathematical derivation to calculate centroid value as follows,

$$\hat{C} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{GP}_{ij}}{N} \quad (3)$$

Where$\widehat{GP}_{ij}$ denote the position of gene patterns belongs to kthcaluster, in eq. (3)$\hat{C}$It is used to calculate the centroid value for each cluster, and Ndenotes the total number of genes patterns in kthgene cluster. Mathematical derivation to calculate Similarity among centroids is represented as follows,

$$sim(\hat{C}_i, \hat{C}_j) = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}|\hat{c}_i - \hat{c}_j|}{\sum_{i=1}^{k}\sum_{j=1}^{n}(\hat{c}_i + \hat{c}_j)} \quad (4)$$

Where$\hat{C}_i$ denotes centroid value of ithgenecluster and$\hat{C}_j$ denotes the centroid value of jthgene cluster, in eq. (4) $sim(,\hat{C}_j)$ is used to calculate the Similarity among two-gene cluster centroid. If two gene clusters have maximum similarity, then these two clusters are merged and forma new cluster.

Initially, the cluster center is randomly selected to calculate Similarity among each gene pattern. Each gene cluster's centroid updates its value for each iteration by calculating the mean of clustered gene patterns value. This gene clustering process has been performed until cluster gene pattern values remain unchanged for two more iterations.

Gene Duplication Counting

Gene duplication counting is a significant process to identify influential proteins gene patterns. In this research similarity-based approach is utilized to count the influential gene pattern in lung cancer gene expression patterns. Mathematical derivation to calculate Similarity among ranked gene patterns with cancer-causing proteins gene patternsare represented as follows,

$$sim(\widehat{RGP}, GP) = 1 - \frac{\sum_{i=1}^{k}|\widehat{RGP}_i - GP|}{\sum_{i=1}^{k}(\widehat{RGP}_i + G\ )}\widehat{RGP} = max(\hat{C}_i) i \in 1, \dots k \quad (5)$$

Where in eq.(5$\widehat{RGP}$Denotes the ranked gene, and GP denotes the manually collected cancer-causing genepatterns, and $max(\hat{C}\ )$The gene clustering process has been performed to group similar gene patterns for gene ranking. It is used to rank gene patterns based on a maximum similarity value. The topmost ranked gene values were selected from each cluster after ranking.

Multiview Surveillance Techniques

F-DES-Fast and deep event summarization method successfully reduces video content while retaining important information such as events, according to the results of experiments. Real-time applications require the system to be up and running to do so.Equal partition-based approach for event summarization in videos:For video,to obtain the optimal number of key-frames without incurring additional computational costs by implementing Davies-Bouldin Index, a cluster validation technique.

Event bagging: A novel event summarization approach in Multiview surveillance videos:It used a meta approach to train the ensembles so that the interdependency and illumination changes of views have taken into account during the training phase.Deep event learning boost-up approach:The work demonstrates an efficient and accurate technique for detecting and summarizing the event in multi-view surveillance videos using boosting, a machine learning algorithm, as a solution. It is possible to capture interview dependencies across different video views by using weak learning classifiers in the boosting algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The model may perform better by SOMS: an efficient SOM technique for event summarization in multi-view surveillance videos for real-time applications, such as surveillance and security systems. Our proposed SOM-based summarization technique is compared to current state-of-the-art models using both qualitative and quantitative assessment.HDML: Habit detection with machine learning:The HDML model analyses our mood and suggests activities that improve our mood when we are in a bad mood or unproductive state. The model's overall accuracy is around 87.5 percent.

DCR-HMM: Depression detection based on Content Rating using Hidden Markov Model:The use of a Hidden Markov Model to detect depression using a new method based on how depressed the subject rates the content (HMM). The subject has shown a series of materials and, depending on how the subject reacts, predicts whethera subject is depressed.Stock Price Prediction Using Recurrent Neural Network and Long Short-Term Memory: The recommended model uses a different approach. Instead of using data for a specific model, latent dynamics of the data set are identified with the help of deep learning algorithms.

Text query-based summarized event searching interface system using deep learning overcloud:A deep learning framework extracts the features of moving objects in the frames. Local alignment captures the dependencies between different views of the video.Prediction of Liver Disease Using Grouping of Machine Learning Classifiers and many more: The Indian Liver Patient dataset was used. The results show that using grouping classification algorithms improves the accuracy of illness forecasting substantially.

This research identifies the MYC protein influence in gene duplication in lung cancer cells using gene duplication counting. These genes pattern values are compared with cancer-causing proteins gene patterns values. If a gene value gets maximum Similarity with the comparison gene pattern value, then the appropriate cluster's genes are taken for gene counting. Implementation of BLAST or FASTA algorithm for fast results for protein similarity searches, the BLAST algorithm does not produce the most accurate results possible; there is a significant risk that BLAST will miss a distant similarity between sequences that can be readily detected. During the evaluation process, the Similarity-based gene duplication counting approach is measured. The nearly 14% - 15% of MYC protein gene duplication pattern has been identified in the lung cancer gene expression dataset. The subsequent section discusses the AEOMGRU based cancer proteins gene patterns classification.

AEOMGRU based Cancer Proteins Gene Patterns Prediction

The gated recurrent unit GRU network is an add-on version of the Long Short Term Memory (LSTM) network. The GRU network and LSTM work similar, but it doesn't use a cell layer to transfer data. The GRU network has several benefits over LSTM, such as less expensive and faster performance. During the network train, the current input is learned from its previous hidden layer node's output. The minimal gated unit is similar to the fully connected gated unit. Still, thesenetworks work slightly differently than fully connected; GRU, the update and reset gate vector, is fused as forget gate. According to earlier research, the gated units perform well for polyploidy sequence-related dating problems. These proteins are frequently mixed structures for stereochemical, with the theory being that the helices link the parallel strands that form the sheet. Parallel processing can be performed based on the methods.

Therefore, the minimal gated recurrent unit has been utilized in this research to predict the cancer-causing protein's gene patterns. The classic minimal GRU network using the gradient descentoptimizer to update network weight by back-propagating the network, but this approach is simple. It consumesa lot of time to predict desired results. It is a global optimization approach, and it takes less computation time to reach global minima. Therefore, the artificial ecosystem optimizer has been utilized to resolve the issues mentioned above in the classic optimizer in the gated recurrent unit network model. The efficiency of the bio-inspired optimizer has already been used in related work parts.

The following derivations of the classifier areused to perform the gene pattern prediction process,

$$GP = \{GP_{11}GP_{12} \dots GP_{1j}GP_{21} \vdots GP_{22} \vdots \cdots GP_{2j} \vdots\vdots GP_{i1}GP_{i2} \dots GP_{ij}\} \quad (6)$$

Where in eq.(6), the GP denotes cancer-causing proteins gene pattern vector, which contains the number of gene patterns and j number of gene features andGP_{ij}Denotes the position of input gene pattern feature value. The minimal gated recurrent unit generally uses sigmoid and tangent activation functions to decide state activation for each node operation.

$$for_t = \sigma_g(W_{for}GP_t + U_{for}h_{t-1} + bi_{for})\sigma(ac) = (1 + e^{-pc})^{-1} \quad (7)$$

Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Ineq.(7), thefor_t denotes the forget gate value at time *t* and the symbolσ_g denotes the sigmoid activation function gate, ac denotes the actual gene pattern class value, and pcdenotes the predicted class value. The default threshold range of the sigmoid activation is denoted as sigmoid (0 1), *0* represents an unsuccessful node, and *1* represents a successful node. Still, the values between >0 and <1 are considered for back-propagation. The$W_{for}GP_t$ denotes the weight of each forgets gate node of gene pattern at time t, U_{for} denotes the learning rate of the forgets gate, theh_{t-1} denotes the currently hidden nodes input value and thebi_{for}The forget gate node calculates the sigmoid value for input gene patterns and currently hidden nodes and parameters values. Denotes bias assigned for each forget node. Finally, the sigmoid gate decides whether to keep node value or ignore based on the state activation threshold.

$$\hat{h}_t = \partial_h(W_hGP_t + U_h(for_t * h_{t-1}) + bi_h)\partial(av) = \frac{sinsinav}{coscosav} = \frac{e^{av}-e^{-pv}}{e^{av}+e^{-pv}} \quad (8)$$

In eq. (8)$\hat{h}_t$ denotes the candidate vector value at time *t* and the symbol∂_h denotes the tangent activation function gate for each candidate, ac denotes the actual glass value of neattern,d,andpcdenotes the predicted class value. This activation function decides the activation by using $\partial(av)$Value the, which is calculated by dividing t$sin\ sin(av)$ value and av) value. The default threshold range of the tangent activation is denoted as tangent (-1 1), *-1*represents an unsuccessful node,and *1* represents successful node b, at the values between >-1 and <1 is considered for back-propagation. $W_hGP_t$denotesthe weight of each hidden candidate output node's output of gene pattern at time *t*, and thebi_h Denotesassigned for each hidden candidate node's output. The candidate vector calculates the agent at value for each gene pattern along with the current hidden node's value and for gesereget t node's favor time stamp prime stamp. Finally, the tangent gate decides whether to keep node value or ignore based on the state activation threshold.

$$o_t = (1 - for_t) * h_{t-1} + for_t * \hat{h}_t \quad (9)$$

In eq.,(9) o_t denotes the output gate value at time *t* and the symbolfor_t denotes the candidate vector value at time *t*, theW_h denotes the weight of each hidden candidate output node's output, and tbi_h denotes bias assigned for each hidden candidate node output. Finally, the output gate node learns the successful gene patterns features value base for getting the candidate vector values after the element-wise visitations. The output node store stores predicted values for each pattern when the gemmoid gate and tangent gate's predicted value is near the activation threshold (gene patterns class value; otherwise, back-propagate the network using an artificial ecosystem optimizer until the specific time stamp gets over. The network weights have been updated using AEO operators such as producer, consumer, and decomposer. The production operator is represented as follows,

$$GP_1(t+1) = (1-lc)GP_i(t) + lcGP_{rand}(t)lc = \left(1 - \frac{t}{MI}\right)r_1$$
$$GP_{rand} = r(U-L) + L \quad (10)$$

wherein eq. (10) *i* denotes the size of the gene pattern population, MI denotes the maximum iteration L, and U denotes the lower and upper limits, r denotes the random vector. Its range is [0,1], the *lc* is denoted linear coefficient value. ThGP_{rand} is randomly calculated gene feature from the population.

The consumer factor operator is represented as follows,

$$GP_i(t+1) = GP_i(t) + CN * \left(GP_i(t) - GP_1(t)\right), i \in [2, \dots, n]$$
$$CN = \frac{1}{2} * \frac{v_1}{|v_2|} where v_1 \sim ND(0,1),\ v_2 \sim ND(0,1) \quad (11)$$

Where in eq. (11) the consumption factor is denoted as CN,GP_1*Denotes* the position of gene feature and the $ND(0,1)$Denotes the normal distribution range of meanv_1And standard deviationv_2.The consuming behavior of carnivores is represented as follows,

$$GP_i(t+1) = GP_i(t) + CN * \left(GP_i(t) - GP_j(t)\right), i \in [3, \dots, n]$$
$$j = rand([2i-1]) \quad (12)$$

wherein eq. (12) j denotes the random number generator, which is generated random number by calculating $2i-1$. The consuming behavior of omnivore is represented as follows,

$$GP_i(t+1) = GP_i(t) + CN * \Big(r_2 * \left(GP_i(t) - GP_j(t)\right)\Big) + (1-r_2)(GP_i(t) - GP_j(t))$$
$$where i \in [3, \dots, n] j = rand([2i-1]) \quad (13)$$

Where in eq. (13), ther_2IsIt also a random number generator, but the range value is between [0, 1]. The decomposition behavior is represented as follows,

$$GP_i(t+1) = GP_n(t) + DC * \left(e * GP_n(t) - h * GP_i(t)\right).$$
$$where,$$
$$DC = 3u,\ u \sim ND(0,1)$$
$$e = r_3 * randi([1,2]) - 1$$

Journal of Theoretical and Applied Information Technology
30th September 2022. Vol.100. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$h = 2 * r_3 - 1$ (14)

Where in eq. (14thee e, h, DC denotes the parameter to generate random numbers and normal distribution values.

The AEO starts the optimization by randomly generating a gene pattern population to update the network weight during the backpropagation. At each timestamp, the first gene pattern features update its weight based on eq. (10), and for the other gene pattern features same probability is used to choose among eq. (11), eq. (12) and eq. (13) to update their weights. If every gene feature gives the store the best value, it is accepted, as best, and each gene feature updates its weight based on eq. (14). When the predicted gene feature value goes out of the lower or upper threshold value of activation function during the network weight updating process, the optimizer randomly generates gene patterns. The networks node's weight updating process has been performed until it reaches the specified timestamp count.

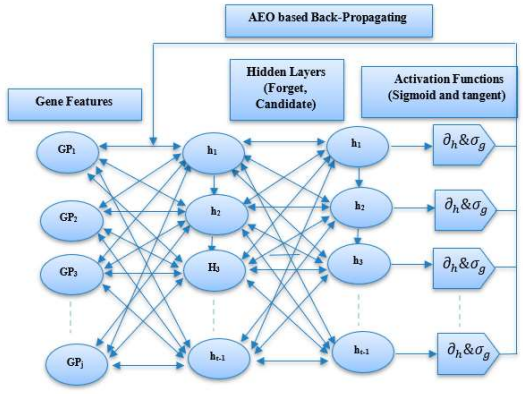

Fig. 5 General Architecture of AEOMGRU Network

Above figure 5 illustrates how the AEOMGRU network node's weights updating process is performed using the AEO algorithm. The step-by-step explanation of the network functions is given in the below algorithmic code.

AEOMGRU algorithm

Input: Protein gene pattern

t = GP_{ij}, $h = 0$, bi =,$U=0.02$ // Initialize the population, hidden layer, bias value and learning rate
FOR each i=1 to m
FOR each j=1 to n
IF(GPij==not null)
GP_{ij}=rand(1,4);
GP_{ij}=GP_{ij}*20;
GP_{ij} =normalize(GP_{ij}) //Normalize the input gene patterns
ELSE
GP_i = 0 // delete row
END IF
END FOR
END FOR
FOR each t=1 to m
FOR each h=1 to n
$\sigma_g(W_{for}GP_t + U_{for}h_{t-1} + bi_{for})$ //Compute the sigmoid value for forgetting gate value
$\partial_h(W_hGP_t + U_h(for_t * h_{t-1}) + bi_h)$ //Compute the tangent value for candidate gate value
IF ($\partial_h == 1$)&& ($\sigma_g == 1$)
for_t // forget state activate
$\hat{h}_t$ // candidate vector state activate
$o_t = (1 - for_t) * h_{t-1} + for_t * \hat{h}_t$ // update output date value
ELSE IF (∂_h> -1) && ($\sigma_g > 0$)
FOR each i=1 to m
FOR each j=1 to n
$GP_i(t+1)$ //update network nodes weight by using AEO based back-propagating(using eq(10) - eq(14))
END FOR
END FOR
for_t // forget state activate
$\hat{h}_t$ // candidate vector state activate
$o_t = (1 - for_t) * h_{t-1} + for_t * \hat{h}_t$ // update output date value
ELSE
$o_t = 0$ // ignore
END IF
END FOR
END FOR

Output: Predicted cancer gene patterns

Initially, the gene expression pattern is taken as input to the AEOMGRU network. The gene patterns are preprocessed before to AEOMGRU network's input layer. The normalized gene expression values are taken as input to the network, and the next important step is parameter initialization. Computes the forget node value and candidate vector value by calculating sigmoid value and tangent value for input node value and weight values hyperparameters. If the sigmoid value is lesser than a threshold, it updates the network weight by using AEO based back-propagation until it reaches the stopping criteria. Finally, the output gate updates its values based on the forget gate node and candidate vector node for classifying the cancer gene pattern. The efficiency of

the new classifier is evaluated with the lung cancer dataset. The evaluation results are explained in the subsequent section.

4. RESULTS AND DISCUSSIONS

This section discusses the evaluation results of the AEOMGRU classification tools to evaluate the performance high dimensional gene expression patterns are utilized, which is described in the data source section. The accuracy values for gene datasets are predicted by Metagenome Gene prediction. Classifier's efficiency is compared with recently introduced bio-inspires minimize deep neural new approach such as GAANN, PSOANN,and classic GRU are taken for comparison. The following evaluation metrics [45], [46] are utilized to evaluate the AEOMGRU classifier's performance.

$$Accuracy\ (Acc) = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalseNegative + FalsePasit} \quad (15)$$

$$Precision\ (Pr) = \frac{TruePositive}{TruePositive + FalsePasitive} \quad (16)$$

$$TruePositive\ (TP) = \frac{TruePositive}{TruePositive + FalseNegati} \quad (17)$$

$$FalsePasitive\ (FP) = \frac{FalsePasitive}{TruePositive + FalsePasi} \quad (18)$$

$$Rootmeansquareerror\ (r^2) = \sqrt{\frac{\sum_{r=1}^{N}(AV_r - PV_r)^2}{N}} \quad (19)$$

$$MeanAbslutePercentError\ (\alpha) = \frac{100}{n}\sum_{r=1}^{n}\frac{|AV_r - PV_r|}{AV_r} \quad (20)$$

Where in eq. (19) and eq. (20)AV_rDenotes the actual value of the class label in the r[th] position aPV_r Denotethe predicted value in the r[th] position. N and n denote the total number of gene patterns—the eq. (15) to eq. (17) are various denotes accuracy matrices used to calculate the accuracy values and eq. (18) to eq. (20) are various denotes error rate metrics, which are used to calculate the error values.

Table 1. Accuracy (%) values for Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	95.6	96.2	98.5	99.5
2	93.9	95.7	97.3	99.2
3	95.7	96.9	96.4	98.9
4	92.3	95.2	96.9	98.7
5	95.1	94.5	97.6	99.1

Table 1 contains the accuracy value obtained by the GAANN, PSOANN, GRU, and AEOMGRU fother lung cancer gene dataset. It shows that the Artificial Ecosystem Optimizer helps train gene patterns to the Minimal Gated Recurrent Unit network proficiently. The AEOMGRU outperforms comparison algorithms for all the iterations, which has obtained a maximum of 99.5 % accuracy rate.

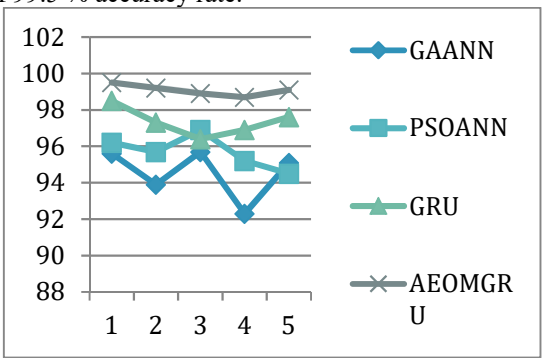

Fig. 6 Accuracy (%) Values For Lung Cancer Gene Dataset

Figure 6 demonstrates the accuracy value obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It clearly shows that AEOMGRU obtains the maximum accuracy.

Table 2. Precision (%) Values For Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	95.7	95.9	97.9	99.3
2	92.8	96.4	98.1	99.7
3	94.7	95.6	98.7	98.9
4	94.2	95.4	97.2	99.8
5	93.3	93.8	98.6	99.6

Table 2 contains the precision value obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer dataset. It shows that the Artificial Ecosystem Optimizer helps the Minimal Gated Recurrent Unit network in gene patterns proficiently. The AEOMGRU outer than the comparison algorithms for all the iterations s has obtained a maximum of 99.8 % precision rate.

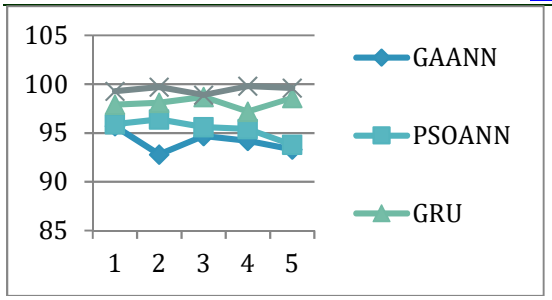

Fig. 7 Precision (%) Values for Lung Cancer Gene Dataset

The above figure 7 demonstrates the precision value obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It clearly shows that AEOMGRU obtains maximum precision.

Table 3. True Positive (%) Values for Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	94.9	95.1	98.9	99.8
2	93.5	93.9	99.1	99.5
3	95.6	94.2	98.5	99.2
4	93.8	95.4	96.2	98.9
5	94.2	94.8	98.4	98.6

Table 3 contains the True Positive rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. The AEOMGRU outperforms comparison algorithms for all the iterations. It has obtained a maximum of 99.8 % True Positive rate. It shows that the Artificial Ecosystem Optimizer help strain gene patterns to the Minimal Gated Recurrent Unit network proficiently.

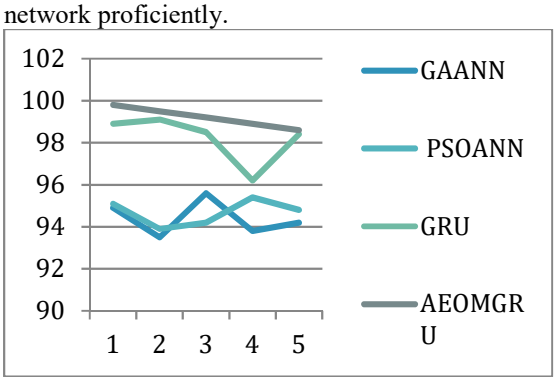

Fig 8 True Positive (%) Values for Lung Cancer Gene Dataset

Figure 8ddemonstrates the True positive rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU effort for the cancer gene dataset. It clearly shows that AEOMGRU obtains the maximum true positive rate.

Table 4. False Positive (%) Values for Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	5.1	4.9	1.1	0.2
2	6.5	6.1	0.9	0.5
3	4.4	5.8	1.5	0.8
4	6.2	4.6	3.8	1.1
5	5.8	5.2	1.6	1.4

Table 4 contains the false positive rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. The AEOMGRU outperforms comparison algorithms for all the i; it has obtained a minimum of 0.2 % False Positive rate. It shows that the artificial ecosystem optimizer helps train gene patterns to minimal gated recurrent unit networks proficiently.

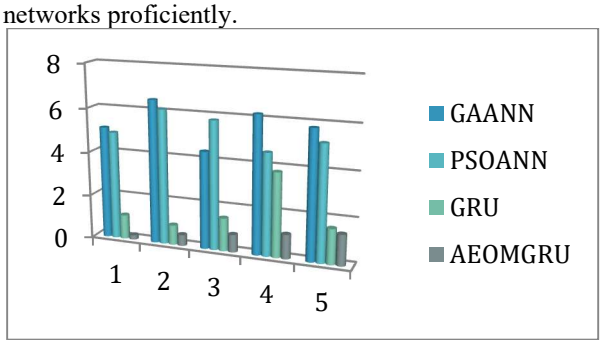

Fig 9. False Positive (%) Values for Lung Cancer Gene Dataset

Figure 9 demonstrates the false positive rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It clearly shows that AEOMGRU obtains the minimum false positive rate.

Table 5. Root Mean Square Error (%) Values for Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	4.3	4.1	2.1	0.7
2	7.2	3.6	1.9	0.3
3	5.3	4.4	1.3	1.1
4	5.8	4.6	2.8	0.2
5	6.7	6.2	1.4	0.4

Table 5 contains the Root mean square error rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. The AEOMGRU outperforms comparison algorithms for all the iterations; a minimum of 0.2 % Root mean

square error rate has been obtained. It shows that the artificial ecosystem optimizer helps proficiently train gene patterns in the minimal gated recurrent unit network.

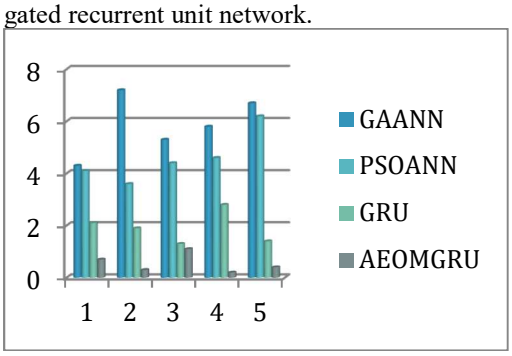

Fig. 10 Root Mean Square Error (%) Values for Lung Cancer Gene Dataset

Figure 10 demonstrates the Root mean square error rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It clearly shows that the minimum Root meansAEOMGRU obtains a square error rate.

Table 6. Mean Absolute Percent Error (%) Values for Lung Cancer Gene Dataset

Successful Rounds	GAANN	PSOANN	GRU	AEOMGRU
1	4.4	3.8	1.5	0.5
2	6.1	4.3	2.7	0.8
3	4.3	3.1	3.6	1.1
4	7.7	4.8	3.1	1.3
5	4.9	5.5	2.4	0.9

Table 6 contains the Mean absolute percent error rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It shows that the Artificial Ecosystem Optimizer helps train gene patterns to the Minimal Gated Recurrent Unit network proficiently. The AEOMGRU outperforms comparison algorithms for all the iterations, and it has been obtained a minimum of 0.5 % Mean absolute percent error rate.

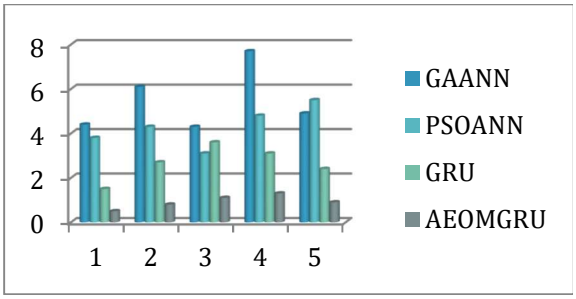

Fig. 11 Mean Absolute Percent Error (%) Values for Lung Cancer Gene Dataset

Figure 11 demonstrates the Mean absolute percent error rate obtained by the GAANN, PSOANN, GRU, and AEOMGRU for the lung cancer gene dataset. It clearly shows that AEOMGRU obtains the minimum Mean absolute percent error rate.

Table 7: Abbreviations for the Datasets provided

Datasets	Abbreviations
GAANN	Genetic Algorithm Artificial Neural Network
PSOANN	Particle Swarm Optimization Artificial Neural Network
GRU	Gated Recurrent Unit
PAR	Photosynthetic Active Radiation
MYB	Myeloblastosis
KRAS	Kirsten Rat Sarcoma virus oncogenic

The evaluation results prove that the AEOMGRU classification tool outperformed comparison algorithms. The classifier predicted the lung cancer genes with high throughput in less computation time for all the successful rounds. It indicates that the Artificial Ecosystem Optimizer helps train gene patterns to the Minimal Gated Recurrent Unit network proficiently during the lung gene pattern prediction. The effectiveness of the suggested models was compared to that of newly established bio-inspired optimizer deep neural network approaches as GAANN, PSOANN, and classic GRU. Each round's accuracy and precision were evaluated using the false positive, false negative, true positive, and true negative values. The true positives and true negatives were observed higher, and the false negatives and false positives have shown the least count. These counts resulted in the highest accuracy and precision for the proposed model compared to existing models.

5. CONCLUSION AND FUTURE SCOPE

Thus, the similarity cluster-based gene duplication counting approach helps identify the MYC proteins influence percentage in lung cancer gene patterns. The overall results and discussion show that the Artificial Ecosystem Optimizer helps train gene patterns to the Minimal Gated Recurrent Unit network proficiently during the lung gene pattern prediction. It depicts the combined features of minimized forget gate character of GRU network and the consumer, producer, and decomposer operator's behaviors of the AEO algorithms helps to learn the gene patterns efficiently. It illustrates that the

combined features of the network have minimized forget gate character and the consumer, producer, and decomposer operator behaviors which helped learn gene patterns efficiently. The AEOMGRU classifier outperforms for gene expression pattern text data. Therefore, the classifier's flexibility is generalized to handle all the data-based text problems. At present, this classification and clustering tool's efficiency is tested with healthcare problems. With an average accuracy ratio of 99.08 percent, an average precision rate of 99.2 percent, the least root means the square error of 0.2 percent, and a least mean absolute error of 0.5 percent, the suggested classifier has the maximum efficiency of performance measures. Future research focuses on integrating the CNN classifier with the proposed optimizer to achieve 100% efficiency in this domain and other domain problems.

CONFLICT OF INTEREST(COI)

The authors declare that they have no competing interests.

REFERENCES

[1] Chowdary, M. K., Nguyen, T. N., &Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human-computer interaction applications. Neural Computing and Applications, 1-18.

[2] Carabet L, Rennie P, Cherkasov A (2018) Therapeutic Inhibition of Myc in Cancer. Structural Bases and Computer-Aided Drug Discovery Approaches. International Journal of Molecular Sciences, 20 (1), 120. doi:10.3390/ijms20010120.

[3] Dash, R. K., Nguyen, T. N., Cengiz, K., & Sharma, A. (2021). Fine-tuned support vector regression model for stock predictions. Neural Computing and Applications, 1-15.

[4] Amanatidou A I, Dedoussis G V (2021) Construction and analysis of protein-protein interaction network of non-alcoholic fatty liver disease. Computers in Biology and Medicine, 131, 104243.

[5] Gheisari, M., Najafabadi, H. E., Alzubi, J. A., Gao, J., Wang, G., Abbasi, A. A., & Castiglione, A. (2021). OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city. Future Generation Computer Systems, 123, 1-13.

[6] Wang S, Zhu X-q, Cai X (2017) Gene Duplication Analysis Reveals No Ancient Whole Genome Duplication but Extensive Small-Scale Duplications during Genome Evolution and Adaptation of Schistosomamansoni. Front. Cell. Infect. Microbiol. 7:412. doi: 10.3389/fcimb.2017.00412.

[7] Billah, M. F. R. M., Saoda, N., Gao, J., & Campbell, B. (2021, May). BLE Can See: A Reinforcement Learning Approach for RF-based Indoor Occupancy Detection. In Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021) (pp. 132-147).

[8] De Kegel B, Ryan C J (2019) Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLOS Genetics, 15 (10), e1008466. doi:10.1371/journal.pg.1008466.

[9] Nichio BTL, Marchaukoski JN Raittz RT (2017) New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. Front. Genet. 8:165. doi: 10.3389/fgene.2017.00165.

[10] Poluri K M (2021) Protein-Protein Interactions: Principles and Techniques: Volume I. Springer Nature.

[11] Shirmohammady N, Izadkhah H, Isazadeh A (2021) PPI-GA: A Novel Clustering Algorithm to Identify Protein Complexes within Protein-Protein Interaction Networks Using Genetic Algorithm. Complexity.

[12] Armenia, Wankowicz S A M, Liu D, Gao J, Kundra R, Van Allen E M (2019) Publisher Correction: The long tail of oncogenic drivers in prostate cancer. Nature Genetics. doi:10.1038/s41588-019-0451-6.

[13] Omranian S, Angeleska A, Nikoloski Z (2021) PC2P: Parameter-free network-based prediction of protein complexes. Bioinformatics.

[14] charya D, Dutta T K (2021) Elucidating the network features and evolutionary attributes of intra-and interspecific protein-protein interactions between human and pathogenic bacteria. Scientific Reports, 11 (1), pp 1-11.

[15] Ascencio D, Diss G, Gagnon-Arsenault I, Dubé A K, DeLuna A, Landry C R (2021) Expression attenuation as a robustness mechanism against gene duplication. Proceedings of the National Academy of Sciences, 118 (6).

[16] Hemminki A, Hemminki K (2005) The Genetic Basis of Cancer. In: Curiel D.T., Douglas J.T. (eds) Cancer Gene Therapy. Contemporary Cancer Research. Humana Press. https://doi.org/10.1007/978-1-59259-785-7_2.

[17] Lepkes L, Kayali M, Blümcke B, Weber J, Suszynska M, Schmidt S, Ernst C (2021) Performance of In Silico Prediction Tools for the Detection of Germline Copy Number Variations in Cancer Predisposition Genes in 4208 Female Index Patients with Familial Breast and Ovarian Cancer. Cancers, 13(1), 118.

[18] Vazquez J M, Lynch V J (2021) Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife, 10, e65041.

[19] Zhu X, Tian X, Ji L, Zhang X, Cao Y, Shen C, Chen H (2021) A tumor microenvironment-specific gene expression signature predicts chemotherapy resistance in colorectal cancer patients. NPJ precision oncology, 5(1), pp 1-14.

[20] Sha K, Lu Y, Zhang P, Pei R, Shi X, Fan Z, Chen L (2021) Identifying a novel 5-gene signature predicting clinical outcomes in acute myeloid leukemia. Clinical and Translational Oncology, 23(3), pp 648-656.

[21] Abdelazim M A, Nasr M M, Ead W M. A Survey on Classification Analysis for Cancer Genomics: Limitations and Novel Opportunity in the Era of Cancer Classification and Target Therapies.

[22] Jia D, Chen C, Chen C, Chen F, Zhang N, Yan Z, Lv X (2021) Breast Cancer Case Identification Based on Deep Learning and Bioinformatics Analysis. Frontiers in Genetics, 12.

[23] Zhang J, Li, D, Zhang Y, Ding Z, Zheng Y, Chen S, Wan Y (2020) Integrative analysis of mRNA and miRNA expression profiles reveals seven potential diagnostic biomarkers for non-small cell lung cancer. Oncology reports, 43(1), pp 99-112.

[24] Bilal M, Raza S E A, Azam A, Graham S, Ilyas M, Cree I A, . Rajpoot N M (2021) Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. medRxiv.

[25] Lai Y H, Chen W N, Hsu T C, Lin C, Tsao Y, Wu S (2020) Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific reports, 10(1), pp 1-11.

[26] Zhao M, Chen Z, Zheng Y, Liang J, Hu Z, Bian Y, Wang Q (2020) Identification of cancer stem cell-related biomarkers in lung adenocarcinoma by stemness index and weighted correlation network analysis. Journal of cancer research and clinical oncology, 146(6), pp 1463-1472.

[27] Jiang M, Niu C, Cao J et al. (2018) In silico-prediction of protein-protein interactions network about MAPKs and PP2Cs reveals a novel docking site variants in Brachypodiumdistachyon. Sci Rep 8, 15083. https://doi.org/10.1038/s41598-018-33428-5.

[28] A von der Dunk, SH, Snel B (2020) Recurrent sequence evolution after independent gene duplication. BMC EvolBiol 20, 98. https://doi.org/10.1186/s12862-020-01660-1.

[29] Chauve C, Ponty Y Wallner M (2020) Counting and sampling gene family evolutionary histories in the duplication-loss and duplication-loss-transfer models. J Math Biol 80, pp 1353–1388. https://doi.org/10.1007/s00285-019-01465-x.

[30] J Paszek, P Górecki (2018) Efficient Algorithms for Genomic Duplication Models. in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 5, pp. 1515-1524, doi: 10.1109/TCBB.2017.2706679.

[31] Ruffalo M, Bar-Joseph Z (2019)Protein interaction disruption in cancer. BMC Cancer 19, 370. https://doi.org/10.1186/s12885-019-5532-5.

[32] Gueudre T, Baldassi C, Pagnani A, Weigt M (2020) Predicting Interacting Protein Pairs by CoevolutionaryParalog Matching. In: Canzar S., Ringling F. (eds) Protein-Protein Interaction Networks. Methods in Molecular Biology, vol 2074. Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9873-9_5.

[33] Li Z, Ivanov, A, Su R et al. (2017) The OncoPPi network of cancer-focused protein-protein interactions inform biological insights and therapeutic strategies. Nat Commun 8, 14356. https://doi.org/10.1038/ncomms14356.

[34] Chen Y, Wang W, Liu J, Feng J, Gong X (2020) Protein Interface Complementarity and Gene Duplication Improve Link Prediction of Protein-Protein Interaction Network. Frontiers in Genetics, 11. doi:10.3389/fgene.2020.00291.

[35] Zhao W, Wang L, Zhang Z (2019) Artificial ecosystem-based optimization: a novel nature-inspired meta-heuristic algorithm. Neural Computing and Applications. doi:10.1007/s00521-019-04452-x.

[35] Kasongo S M, Sun Y (2020) A Deep Gated Recurrent Unit based model for the wireless

intrusion detection system. ICT Express. doi:10.1016/j.icte.2020.03.002.

[36] N Nezamoddini, A Gholami (2019) Integrated Genetic Algorithm and Artificial Neural Network. 2019 IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing, New York, NY, USA, pp 260-262, doi: 10.1109/CSE/EUC.2019.00057.

[37] Jose Anand, J Raja Paul Perinbam, D Meganathan (2015) Design of GA-based Routing in Biomedical Wireless Sensor Networks. International Journal of Applied Engineering Research. 10 (4), pp 9281-9292.

[38] Hamada M, Hassan M (2018) Artificial Neural Networks and Particle Swarm Optimization Algorithms for Preference Prediction in Multi-Criteria Recommender Systems. Informatics, 5(2), 25. doi:10.3390/informatics5020025.

[39] Izudheen S (2021) Intelligent Exploration of Negative Interaction from Protein-Protein Interaction Network and its Application in Healthcare. Psychology and Education Journal, 58(2), pp 10637-10645.

[40] Lakshmi P, Ramyachitra D (2020) An Improved Genetic with Particle Swarm Optimization Algorithm Based on Ensemble Classification to Predict Protein-Protein Interaction. Wireless Personal Communications, 113(4), pp 1851-1870.

[41] Ge R, Wu Q, Xu J (2021) Computational Methods for Protein-Protein Interaction Network Alignment. In Recent Advances in Biological Network Analysis, Springer, Cham. pp. 45-63.

[42] Su X R, You Z H, Hu L, Huang Y A, Wang Y, Yi H C (2021) An Efficient Computational Model for Large-Scale Prediction of Protein-Protein Interactions Based on Accurate and Scalable Graph Embedding. Frontiers in Genetics, 12.Database: https://www.ncbi.nlm.nih.gov/geo/.

[43] Rostami, M., Berahmand, K., &Forouzandeh, S. (2021). A novel community detection-based genetic algorithm for feature selection. Journal of Big Data, 8(1), 1-27.

[44] Rostami, M., Berahmand, K., &Forouzandeh, S. (2020). A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. Journal of Big Data, 7(1), 1-21.

[45] Nikzad-Khasmakhi, N., Balafar, M., Feizi-Derakhshi, M. R., &Motamed, C. (2021). ExEm: Expert embedding using dominating set theory with deep learning approaches. Expert Systems with Applications, 177, 114913.

[46] Chen, K., Franko, K., & Sang, R. (2021). SWe has structured model Pruning of Convolutional Networks on Tensor Processing Units. arXiv preprint arXiv:2107.04191.

[47] Zhao C, Zang Y, Quan W, Hu X, Sacan A (2017) Hiv1-human protein-protein interaction prediction based on interface architecture similarity, in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (Hong Kong: IEEE), pp 97–100. doi: 10.1109/BIBM.2017.8217632.

[48] Naz H, Ahuja S (2020) Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders. doi:10.1007/s40200-020-00520-5.

[49] Giriprasad S, Mohan S, Gokul S (2018) Anomalies detection from video surveillance using support vector trained deep neural network classifier. International Journal of Heavy Vehicle Systems, 25 (3/4), 286. doi:10.1504/ijhvs.2018.094825.

[50] Sun, X., Chen, W., Ivanov, S., MacLean, A. M., Wight, H., Ramaraj, T., ...&Fei, Z. (2019). Genome and evolution of the arbuscularmycorrhizal fungus Diversisporaepigaea (formerly Glomusversiforme) and its bacterial endosymbionts. New Phytologist, 221(3), 1556-1573.

[51] Skinnider, M. A., Johnston, C. W., Gunabalasingam, M., Merwin, N. J., Kieliszek, A. M., MacLellan, R. J., ...&Magarvey, N. A. (2020). Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nature communications, 11(1), 1-9.

[52] Yoshida, S., Kim, S., Wafula, E. K., Tanskanen, J., Kim, Y. M., Honaas, L.,&Shirasu, K. (2019). Genome sequence of StrigaAsiatica provides insight into the evolution of plant parasitism. Current Biology, 29(18), 3041-3052.

[53] Wang, X., Morton, J. A., Pellicer, J., Leitch, I. J., & Leitch, A. R. (2021). Genome downsizing after polyploidy: mechanisms, rates, and selection pressures. The Plant Journal, 107(4), 1003-1015.