Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Leukemia Cancer Gene Identification Through Genetic Optimization Technique and Prediction with LSTM

M.Kalaivani[1] , Dr.K.Abirami[2] and Dr.K.Dharmarajan[3]

[1,2,3]Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India

E-Mail: mkvanimca@gmail.com

Article History:

Received: *12-11-2024*

Revised:*24-12-2024*

Accepted:*09-01-2025*

Abstract:

Cancer disease persists as one of the major causes of mortality worldwide, emphasizing the critical necessity for accurate and early diagnosis. Identifying genes associated with cancer can facilitate effective early-stage treatment. DNA Microarray data plays a vital role in detecting and diagnosing cancer. Microarray analysis allows the examination of gene expression levels in specific cell samples, enabling the simultaneous analysis of thousands of genes. However, microarray gene data is characterized by high dimensionality and contains redundant genes. The gene (feature) selection process is essential in microarray gene expression data analysis for selecting highly relevant genes with low redundancy that may cause improved prediction results. This research article presents a model designed to reduce the high-dimensional gene data to low-dimensional space, eliminate redundant genes, and improve early-stage disease prediction. The proposed model is implemented in the following phases. The initial process involves checking the multicollinearity between genes and identifying the low intercorrelated genes using the Karl Pearson correlation coefficient on the cancer dataset. Next, the Genetic Optimization Algorithm integrated with the Signal to Noise Ratio method is applied to determine an optimal set of genes, focusing on reducing dimensionality while filtering out noisy and redundant genes. In the final phase, the selected genes are classified using a Deep learning classification method namely the Long Short Term Memory classifier is utilized. The model's performance is evaluated with various optimizers, including Adam, Adagrad, RMSProp, and SGD by comparing accuracy and loss values. Also calculate the other classification parameters such as Precision, Recall, and F1-Score value. Finally, comparative analyses of these metrics across the different optimizers are performed**.** The experimental results demonstrate that the research model Cor-GA_SNR-LSTM significantly improves the detection analysis by reducing the number of features in the gene data, enhancing the accuracy value, and minimizing the loss value. The Adam optimizer yielded the optimal performance with an accuracy of 88.45% and a loss value of 0.1612. The proposed method effectively identifies the most relevant genes responsible for detecting and predicting Leukemia cancer disease.

Keywords: Classification, Genetic optimization technique, Long Short Term Memory, Microarray Analysis, Signal-to-Noise Ratio.

1. INTRODUCTION

A key challenge in the medical industry is accurately diagnosing cancer patients at an early stage. Conventional diagnostic techniques, including Non-invasive methods, Invasive procedures, and Fine needle aspiration cytology may sometimes misinterpret the distinction between normal and cancerous genes, leading to adverse effects. Leukemia is a form of cancer originating in the bone marrow, often caused by genetic abnormalities [1]. The primary cause of Leukemia is the mutation of cells within the bone marrow which can sometimes obstruct the production of healthy cells. Detecting DNA mutations for the early identification of genetic disorders is a challenging process. Before the development of next-generation sequencing, microarray technology had a significant impact on the domain of cancer biology[2]. The

technology enabled thousands of genes to be monitored simultaneously to detect complex gene expression patterns [3]. DNA microarray data, containing gene expression profiles, facilitate precise cancer diagnoses in the healthcare sector. However, the nature of the microarray gene dataset is high-dimensional, containing a large number of genes and a small number of patient sample instances. This leads to the curse of dimensionality that affects the prediction rate of the cancer disease [4]. Also, the majority of genes are non-informative and redundant. The dataset contains a 1:10 ratio of relevant to noisy data, which negatively impacts the model's performance when conventional methods are directly implemented [5]. To enhance the effectiveness of the microarray data, gene/attribute selection techniques perform an important role in processing the data and improving the accuracy of classifier models [6]. The main goals of the research are to identify significant biomarker genes in gene expression data using gene selection techniques and to classify the data through deep learning techniques. Initially, the model checks the multi-collinearity between genes using correlation and identifies the low inter-correlated features. An optimization technique with the rank filter method is utilized to identify key genes from cancer data, aiming to enhance prediction accuracy. The designed model utilized the Genetic Optimization technique with Signal To Noise Ratio method to effectively evaluate the gene patterns and identify an optimal set of prominent biomarker genes. The selected genes are applied in Deep learning classification methods such as the Long short-term memory model to accurately recognize the cancer subtype genes.

The research paper is structured as follows: Section 2 presents a review of related literature, while Section 3 outlines the research resources and methods, including details of the dataset, an overview of the Genetic optimization technique with Signal To Noise Ratio method, and Long Short Term Memory classification model. Section 4 explores the details of the proposed model (Cor-GA_SNR-LSTM). Section 5 discusses the simulation analysis of the research work and comparative analysis with existing techniques. Finally, the research findings are presented in the conclusion part.

2. RELATED WORK

This section briefly overviews existing attribute/gene selection methods for identifying biomarker genes in Leukemia cancer and evaluating their prediction performance through different classifiers.

Waleed Ali et al. [7] presented a hybrid attribute selection model that integrates filter and optimization techniques. In the proposed method, Mutual Information, Chi-square, and IGR filter methods are used to eliminate irrelevant features. After selecting the relevant attributes, a genetic algorithm is employed to select optimal features, enhancing the model's performance for cancer classification. Mehrdad Rostami et al. [8] proposed an innovative biomarker gene selection model based on a community detection cluster method with a genetic algorithm. Initially, the similarities of the features are calculated and the features are grouped using community detection cluster methods. A genetic algorithm with a repair operation was applied to each cluster to select the significant set of features and perform the comparison analysis with other optimization techniques. The proposed method produced higher accuracy in predicting

the disease using the genetic process. Sabah Sayed et al. [9] developed an ensemble-based attribute selection approach that combines a t-test and a nested genetic algorithm to eliminate redundant features. An incremental FS method was then utilized to identify a significant subset of genes. Finally, the biological relevance of the selected genes was validated through Enrichment Analysis, achieving a classification accuracy of 98.4%.

Lawrence et al. [10] designed a framework that combines GA and DBN for the effective selection of attributes and classification. The research study utilized GA to remove the noisy attributes and select the most prominent set of attributes. Finally, the DBN classifier was applied to predict the disease. The presented method accurately identifies the cancer types. Pragadeesh et al. [11] suggested a compounded feature selection approach for classifying cancer gene data. In this approach, the Information Gain filtering technique eliminates redundant attributes, and a micro GA evolutionary computing technique is applied to identify the prominent set of features. Finally, the classification accuracy is calculated using the SVM method.Yuvaraj et al[12] developed an efficient attribute selection model using optimization techniques such as Whale Optimization to select biomarker attributes. Then, a modified LSTM network was utilized to classify the cancer types. The suggested model produced a significant accuracy rate, F-measure, and recall values.

Sergii Babichev et al.[13] explored gene data classification, focusing on the relationships within genetic information, gene activity patterns, and biological processes. The proposed work developed a model that optimizes the architecture and hyperparameter values of Recurrent NN specifically GRU and LSTM, based on the classification accuracy, loss function value, training time, and the F1-Score index using the Harrington desirability method. Finally, the best hyperparameters for each network model are identified to improve the accuracy of disease prediction.Zixuan Wang et al [14] introduced a feature selection framework that combines Isomap and Genetic search techniques. Isomap is used to map high-dimensional nonlinear data into a lower-dimensional linear space, while the genetic search selects feature subsets and enhances the fitness function. The final set of features is selected based on a threshold derived from the binomial distribution. The selected features are expected to improve classification accuracy for cancer gene data.Ebtisam Abdullah Alabdulqader et al [15] presented a diagnostic method utilizing 22,283 Leukemia gene microarray data. The proposed model used a Chi-squared filter method to select the top 300 genes, and the SMOTE-Tomek technique was applied to generate synthetic data. A weighted CNN is then proposed for disease prediction and achieving an accuracy of 99.9%.

Pradeep Kumar Mallick et al. [16] introduced a method designed to enhance the convergence of DNN. The model used Leukemia cancer gene data and implemented a five-layer DNN to classify ALL and AML gene expression data. The proposed approach achieved an accuracy rate of 98.2% and a specificity value of 97.9%. Bibhuprasad Sahu et al [17] developed an ensemble approach for classifying Leukemia cancer data. The research method integrates various filter techniques, including Mutual information, chi-square, Correlation-based FS, and Gain Ratio, to select the relevant features. The selected features are further optimized using the Grey Wolf Optimizer to generate an optimal subset of genes. Recurrent NN and LSTM

classifiers are applied for disease prediction, achieving high accuracy. Abdul Karim et al [18] defined four distinct types of Leukemia, including AML, CML, ALL, and CLL. The authors proposed LDSVM, an ensemble ML algorithm to enhance performance and prediction accuracy. The research emphasizes the need for a distinctive methodology that utilizes optimization techniques to identify leukemia cancer genes and incorporates deep learning classification methods with various optimizers for the prediction of the disease. The current study aims to achieve the expected results.

3. MATERIALS AND METHODS

This section provides an overview of the dataset and the essential methods required to obtain the desired accuracy in predicting cancer disease.

3.1. Dataset

This section explores the outline of the dataset utilized in the research work. The proposed system utilizes the repositories from the Kent Ridge Biomedical Data Set Repository [19]. The microarray cancer dataset is structured as a matrix, with columns representing the attributes or features and rows expressed as the patient's sample instance The dataset enables binary classification models since it has only two output labels called ALL and AML genes. Table-1 shows the properties of the datasets utilized in the research work.

Table-1 Description of Leukemia Cancer Gene Dataset

Dataset	Leukemia Cancer
Gene Count	7182
Class Distribution	2(ALL/AML)
Sample Size	72 49 samples-ALL 23 samples-AML

3.2. Genetic optimization Algorithm

Genetic optimization Algorithm is a well-known heuristic method designed to solve optimization problems through evolutionary techniques. The optimization technique simultaneously explores information from multiple search points[20] and effectively prevents the iterative process from converging on local optimal solutions and producing a global optimal solution[21]. The technique operates based on a population of chromosomes. The process begins with initializing the population, with a set of randomly generated chromosomes. Each chromosome, composed of multiple genes, functions as a candidate solution to the problem. The GA execution process includes several key steps, including Selection, Crossover, and Mutation [22]. In the selection operation, two parent chromosomes are selected using various selection methods like Roulette Wheel, Tournament Selection, and random selection. After the selection process, the selected parent chromosomes are applied to the crossover operator, where parts of the genetic material are combined according to a specified crossover rate to produce

the offspring. The next process executes the mutation operator where certain genes within the offspring are randomly changed. The fitness of the resulting offspring is evaluated, and if it demonstrates better fitness than the parent, the offspring replaces the parent in the population. This execution process continues until a predetermined termination condition is satisfied and the most optimal chromosome is selected as the final solution. Algorithm-I outlines the steps for determining the optimal number of attributes in the dataset.

Algorithm-I # Attribute selection using Genetic Optimization Technique

Step 1: Initialize i=0

Step 2: Generate the population of individuals p(i)

Step 3: Determine the fitness of each individual in the population.

Step 4: While (termination condition is not satisfied)

Step 5: Do i=i+1

Step 6: Apply the Selection operation of two parent chromosomes based on fitness.

Step 7: Perform Crossover with the specified crossover rate to generate offspring

Step 8: Apply the Mutation function to the offspring

Step 9: End While

Step 10: Return the fittest individuals in the population.

3.3. Long Short Term Method (LSTM)Network

LSTM is a type of Recurrent NN and is effective in processing sequential data. The network incorporates a memory cell capable of holding information for a long period and enabling the model to retain the recently computed values [23]. The network consists of three kinds of gates such as input, output, and forget gates, and maintains two state units such as cell and hidden state. The input gate regulates the flow of data into the cell state, while the forget gate is responsible for determining the information from the previous cell state, and the output gate controls the flow of data out of the cell state. The sigmoid activation function is utilized to produce the output values within the range of 0 to 1. The gates are trained using a backpropagation algorithm through the network. The training process for the gates depends on the input and the previous hidden state, allowing the LSTM to selectively retain or forget information, and effectively manage long-term dependencies [24]. The LSTM model was employed to identify patterns and relationships in genetic mutations over sequential time intervals.

4. RESEARCH METHODOLOGY

This section explores the proposed system for effectively identifying biomarker genes with high predictive value. Figure-1 presents the design of the proposed research model. The process includes the following phases: Exploratory Data Analysis, Data preprocessing, Optimal Gene Selection, Classification, and Evaluation of the model. Algorithm-II presents the key steps of

the proposed methodology for identifying the prominent set of genes associated with Leukemia and classifying the subtypes in the dataset.

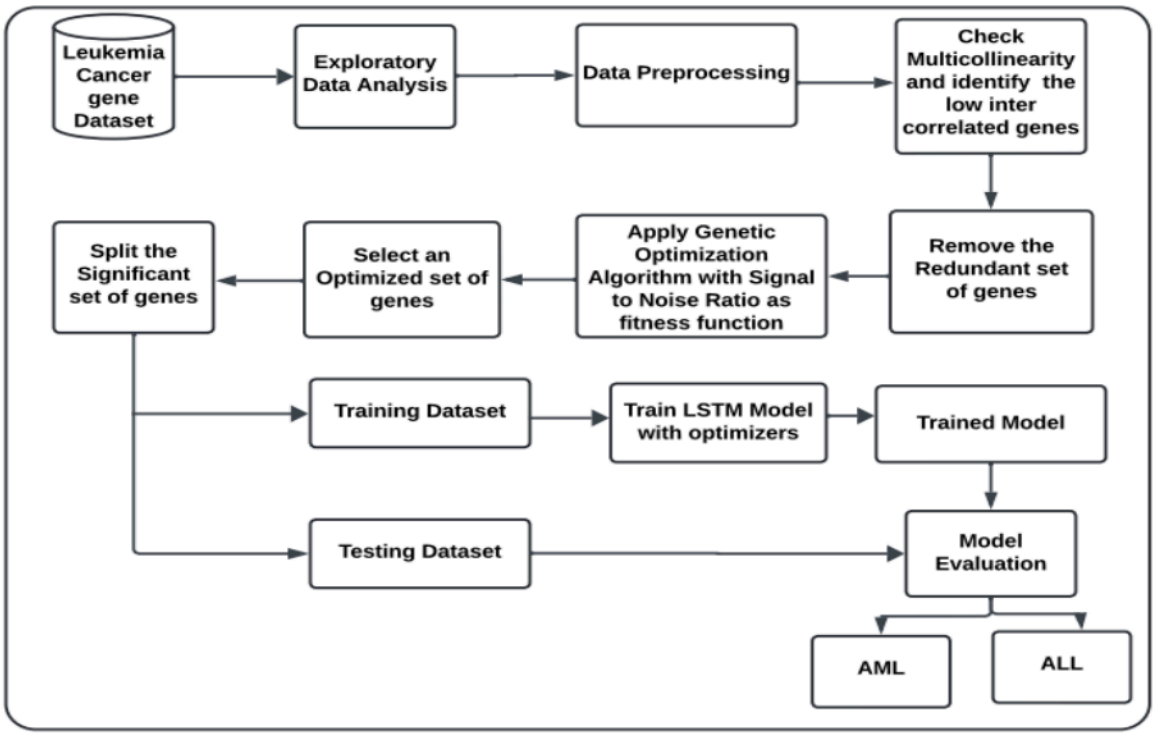

Figure-1 Framework of the research model

Algorithm-II(Cor-GA_SNR-LSTM)

1. Load the Leukemia Cancer Gene Data Set.
2. Perform Exploratory Data analysis:
   2.1 Identify the characteristics and patterns of the gene expression data.
   2.2 Calculate the central tendency and dispersion of the gene values.
3. Perform the Preprocessing Stage:
   3.1 Handle missing and Null values within the dataset by imputation method.
   3.2 Transform categorical gene information into Numerical values using Label Encoding.
   3.3 Apply Feature Scaling using Min-Max Normalization to normalize the data in the dataset.
   3.4 Evaluate Multicollinearity among genes using a Correlation matrix to identify and measure low intercorrelated genes using the equation (1) and (2)
   3.5 Remove redundant genes and store the relevant genes.
4. Perform Optimal Gene Selection:
   4.1 Apply a Genetic Optimization technique with Signal-to-Noise Ratio as the fitness function to identify an optimized subset of significant genes.
   4.2 Store the optimized set of features in a .csv file.
5. Perform the classification process:

5.1 Train the LSTM Model on the optimized gene set to classify the data.

6. Performance Evaluation:

6.1 To evaluate the model performance, different optimizers are applied to measure and analyze accuracy and loss values during the training phase and testing phase, along with precision rate, recall measure, and F1-Score value.

In the first phase, read the Leukemia cancer gene dataset from the KRBR repository. During the Exploratory data analysis, the shape and size of the genes/features, and samples are examined. Identify each gene datatype and calculate the mean, median, min, max, standard deviation, and percentile value of all genes in the dataset. This stage acts as the initial step in investigating the dataset to understand its characteristics, patterns, or anomalies for later stages of analysis. The data preprocessing is an important phase for cleaning the data to achieve suitability for developing the deep learning model, thereby enhancing the accuracy of the model. This phase includes several key processes. The initial process involves handling missing data and null values present in the given data. Replace each missing value with the mean of the sample values for the respective gene. In the next process, the Label Encoding technique is applied to convert the levels of the categorical features into numeric values. In this study, the class labels ALL and AML are encoded as 0 and 1 respectively. Finally, the Feature scaling process is performed, through the min-max normalization technique, which maps the values to a range [0,1]. The highest feature value in the dataset is transformed to 1, the lowest value is changed to 0, and the remaining values are rescaled to fall within the range of 0 to 1. The next process is to evaluate the multicollinearity between features. The correlation matrix evaluates the dependency between features, identifies the low inter-correlated features, and eliminates redundant genes [25]. The correlation matrix is calculated using the formula

$$R_{X,Y} = \frac{CO\_Variance(X,Y)}{\sigma_X \sigma_Y} \quad \text{-------- (1)}$$

$$Co\_Variance(X,Y) = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y) \quad \text{--------(2)}$$

where $R_{X,Y}$ represents the Pearson correlation coefficient and σx, σy represent the standard deviation of x and y respectively.μ_X and μ_Y denote the means of X and Y, and n indicates the number of data points. The next phase involves selecting an optimal gene subset for predicting the disease. In this approach, a Genetic optimization algorithm, integrated with the SNR method is utilized to select an important subset of biomarker genes. The fitness function used in this process is calculated using the formula

$$SNR(f) = \frac{\left|\frac{1}{n_1}\sum_{i \in C_1} x_i - \frac{1}{n_2}\sum_{i \in C_2} x_i\right|}{\sqrt{\frac{1}{n_1}\sum_{i \in C_1}(x_i - \mu_1)^2 + \frac{1}{n_2}\sum_{i \in C_2}(x_i - \mu_2)^2}} \quad \text{-------(3)}$$

where μ_1 and $\mu_2$are the means of the class ALL and AML respectively. In this process, the population P is initialized by generating N chromosomes with random values between 0 and 1. SNR rank method is applied to evaluate the fitness value of individual chromosome and enable the classification of microarray data [26]. The Roulette Wheel method is applied as a selection

operation to select two parent chromosomes with higher fitness values. After selection, the selected parent chromosomes are applied to the crossover operation, with a specified crossover rate to produce the offspring. The next process executes the mutation operator where specific genes within the offspring are randomly changed. The fitness of the resulting offspring is evaluated, and represents better fitness than the parent, the offspring replaces the parent. The procedure continues until a specified termination condition is satisfied, resulting in the selection of the most optimal chromosome as the final solution. This technique reduces the number of genes by eliminating irrelevant and redundant genes from the dataset. This stage helps identify the optimal set of genes connected to cancer disease in the dataset.

To predict the presence of ALL and AML genes, the selected 341 optimal genes are divided into training and testing sets. The training data is applied to train the Long Short Term Memory (LSTM) classifier model. In this process, the network is structured with six hidden layers, each containing 50 neurons. The Sparse categorical cross-entropy loss function is applied to determine the loss value. In this model, different optimizers including ADAM, Adagrad, RMSProp and SGD are tested to evaluate the efficiency of the model. Finally, an evaluation of the metrics for different optimizers is performed.

5. EXPERIMENTAL ANALYSIS

The research methodology was developed and implemented in Python using Google Colob NoteBook, with the Deap package version 1.4.1 configured for the genetic optimization technique. Table-2 presents the parameters used to identify an optimal set of features. The Keras and TensorFlow API Library functions were utilized to develop the LSTM model. The research methodology was evaluated using the Leukemia cancer gene dataset, which consists of 7129 genes and 72 samples. This section presents multiple experiments, and the network was trained on 70 percent of the data and validated on the remaining 30 percent of the data.

Table-2 Parameter Settings for Genetic Optimization Technique

Genetic Optimization Technique Parameter	Values
Selection Method	Roulette Wheel Selection
Crossover Type	Single point Crossover
Mutation Rate	0.5
Number of Chromosomes	50
Number of Populations	50

Table-3 presents the performance of the research methodology using the LSTM classifier, evaluating the accuracy and loss values of various optimizers, including Adam, Adagrad, RMSProp, and SGD. The results highlight the model's learning efficiency, and convergence stability in accurately predicting the disease at an early stage.

Table-3 Performance Metrics of Accuracy and Loss value for various optimizers

Optimizers	Training_ Accuracy(%)	Testing_ Accuracy(%)	Training_ Loss(%)	Testing_ Loss(%)
Adam	94.48	88.45	7.26	16.12
Adagrad	92.56	81.82	20.19	66.38
RMSProp	95.45	86.36	26.40	37.99
SGD	95.36	84.45	7.85	19.05

Figure- 2 indicates the accuracy metrics obtained through different optimizers during the training and testing phases. The ADAM optimizer yielded an increased training accuracy rate of the classification by 1.92% in comparison to the Adagrad optimizer, and reduced training accuracy by 0.97 % than the RMSProp and the accuracy by 0.88 % with the SGD optimizer. Comparing, the testing accuracy of the model indicates that the ADAM optimizer achieved a higher testing accuracy rate of the classification by 6.63% in comparison to the Adagrad optimizer, achieved an increased accuracy rate by 2.09% than the RMSProp optimizer, and 4.9% than the SGD optimizer. The outcomes demonstrate that the LSTM model using the Adam optimizer predicted a strong overall performance with a testing accuracy rate of 88.45% and a training accuracy of 94.48% for predicting Leukemia Cancer.

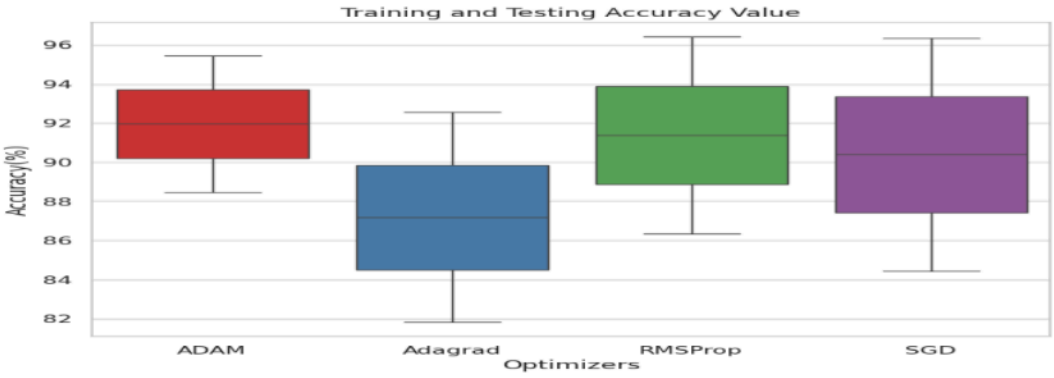

Figure-2 Accuracy Analysis of Various Optimizers

Figure-3 indicates the loss value obtained through different optimizers during the training and testing phases. The ADAM optimizer achieved a testing loss of 0.1612, indicating to generalize effectively to unseen data, and a training loss of 0.0726 reflecting efficient convergence. The Adagrad optimizer recorded a testing loss of 0.6638 and a training loss of 0.2019 indicating increased loss value during the training and testing phase. The RMSProp optimizer showed a testing loss of 0.3799 which was higher than the Adam optimizer and lower than the Adagrad optimizer. The training loss of 0.0785 and a testing loss of 0.1905 obtained for using SGD optimizer indicate the stability in convergence.The outcomes demonstrate that the LSTM model using the Adam optimizer reduced the loss rate to 0.5026 compared with the Adagrad optimizer, 0.2187 with RMSProp, and 0.0293 with the SGD optimizer.

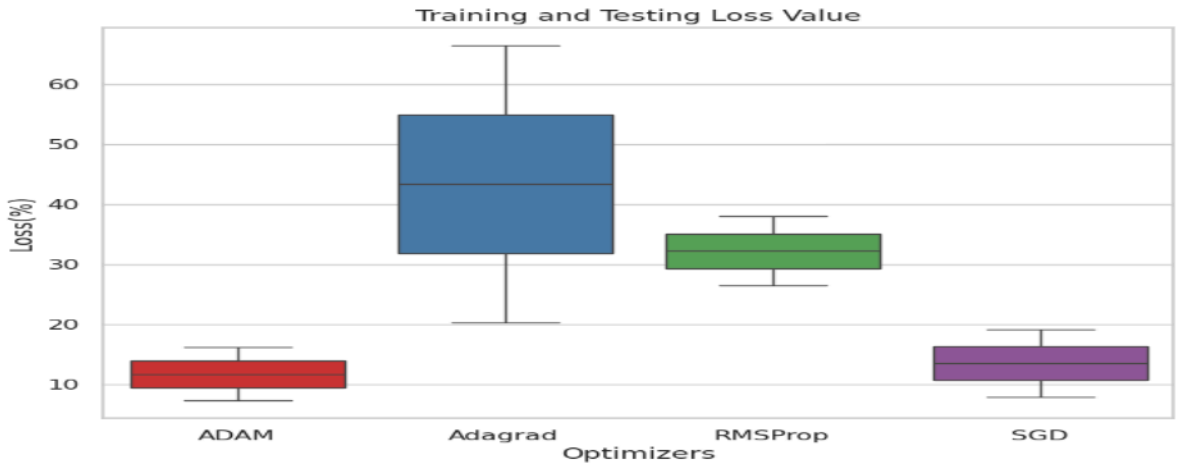

Figure-3 Loss Analysis of Various Optimizers

Table-4 depicts evaluation metrics, including Precision, Recall, and F1-Score for Class 0 (ALL) and Class1(AML) using various optimizers such as ADAM, Adagrad, RMSProp, and SGD applied to the LSTM Classifier.

Table-4 Classification metrics using various optimizers with LSTM

	ADAM			**Adagrad**			**RMSProp**			**SGD**		
Metrics (%)	Preci_ value	Recall_value	F1-Score_value	Preci-_value	Recall_value	F1-Score_value	Preci-_value	Recall_value	F1-Score_value	Preci_value	Recall_value	F1-Score_value
Class 0 ALL	0.93	0.87	0.90	0.80	0.92	0.86	0.81	1.00	0.90	0.82	0.93	0.87
Class 1 AML	0.82	0.90	0.86	0.86	0.67	0.75	1.00	0.67	0.80	0.88	0.70	0.78
Macro Average	0.87	0.88	0.88	0.83	0.79	0.80	0.91	0.83	0.85	0.85	0.82	0.83
Weighted Average	0.89	0.88	0.89	0.82	0.82	0.81	0.89	0.86	0.86	0.85	0.85	0.84

Figure-4 illustrates the Precision metric performance, showing that the proposed model achieved a Macro average Precision of 87**%** and a weighted average precision value of 88% while applying ADAM Optimizer. For the Adagrad optimizer, the precision rates for class 0 (ALL)and class 1 (AML) are 80% and 86% respectively. The optimizer obtained a macro average precision rate of 83% to predict the disease. The RMSProp optimizer received the precision rate in terms of Weighted average is 89%. The SGD optimizer recorded a Precision rate of 85% for both Macro average and Weighted average values.

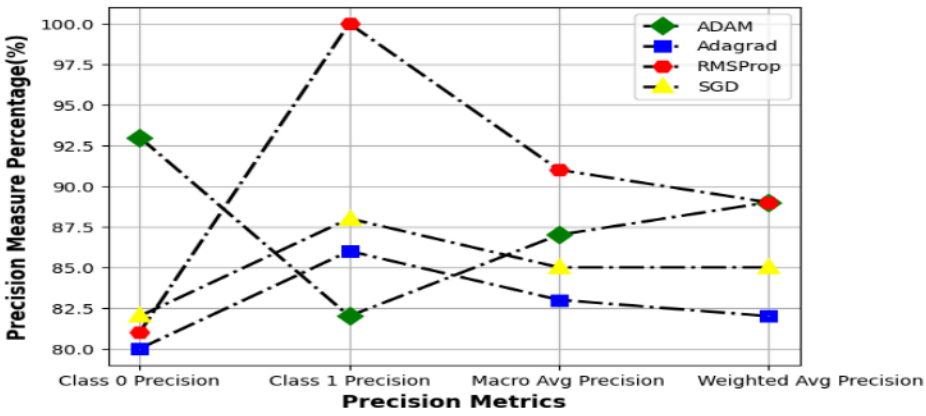

Figure-4 Estimation of Precision value using various optimizers

Figure-5 presents the performance evaluation based on Recall metrics. The model achieved a weighted Average Recall rate of 82% with the Adagrad optimizer, 86% with the RMSProp optimizer and 85% with the SGD Optimizer. The ADAM optimizer attained a higher Recall rate of 88% for data classification.

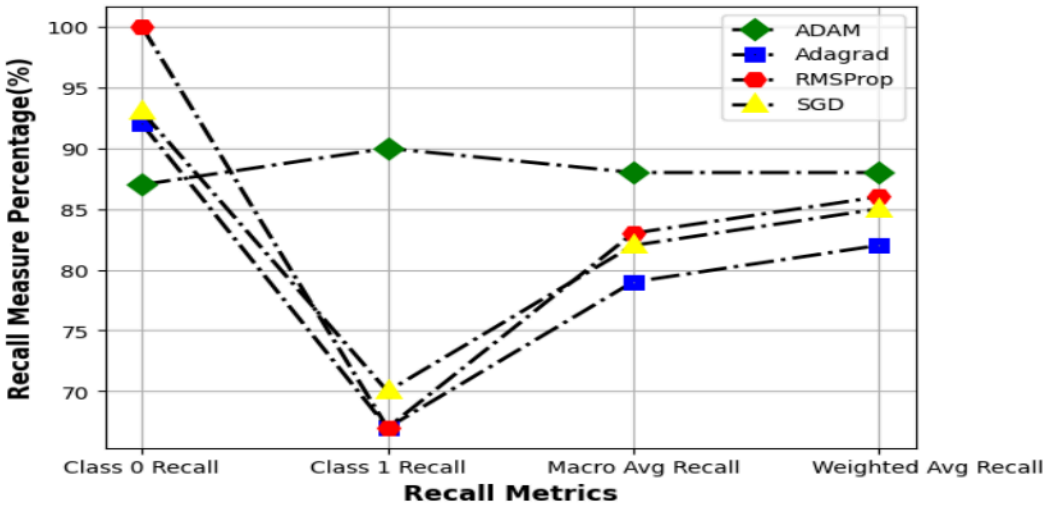

Figure-5 Estimation of Recall value using various Optimizers

Figure-6 highlights the model's performance based on F1-Score metrics. The Adam optimizer achieved an F1-Score of 88%, while the Adagrad optimizer recorded an F1-Score of 80% compared to the other optimizers.

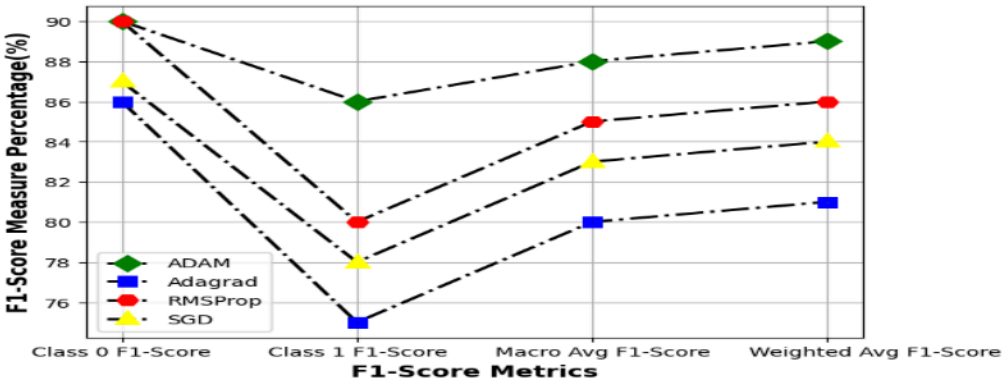

Figure-6 Estimation of F1-Score Value using various optimizers

Researchers have explored different techniques for identifying Leukemia cancer disease. Table-5 presents a comparison of different methodologies used for cancer classification along with their accuracy values. Figure-7 depicts the computational efficiency of the research model compared to existing techniques. The proposed method achieves the highest accuracy of 88.45% for detecting the disease.

Table-5 Comparison between Proposed Approach and Existing Techniques

Reference	**Methodology**	**Accuracy(%)**
[27]	BSS/WSS-based CART	84.43
[28]	CART	85.29
[29]	SNR-SVM	75.6
[30]	PCA-QDA	82.4
[31]	MLP+MI	67.6
[32]	LNNDP	88.24
[33]	Chi2_DT	74
[34]	DLFCC	87
Proposed Method	Cor-GA_SNR-LSTM	88.45

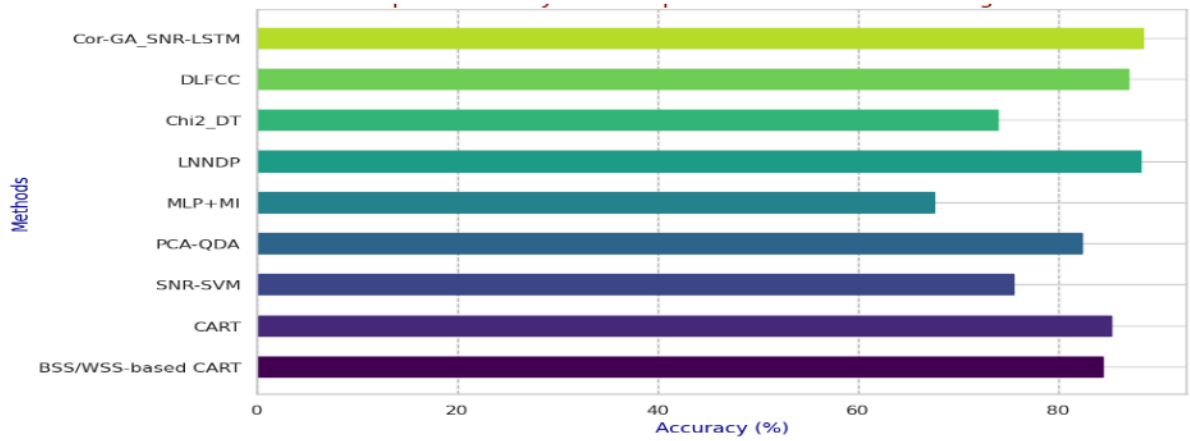

Figure-7 Comparative Analysis of Existing methods Versus Proposed method

6. CONCLUSIONS

Leukemia cancer is one of the most threatening and devastating diseases around the world. Early and accurate detection of Leukemia at its initial stages remains a significant challenge in the medical field. The research study proposes a method that utilizes microarray clinical gene data to identify Leukemia cancer. In the proposed method, the gene data is preprocessed initially, which includes handling missing and null values present in the data, transforming categorical variables into numerical format through the label encoding method, and applying Min Max Normalization for scaling the feature transformation. After preprocessing, the multicollinearity is checked, and genes with low intercorrelation are identified using the Karl Pearson correlation coefficient. After that, the Genetic Optimization Algorithm combined with the Signal to Noise Raio method is applied to select an optimal set of genes in the dataset while eliminating noisy and redundant genes, thereby enhancing classification accuracy. In the final phase, the selected genes are classified using a Long Short Term Memory classifier. The model's performance is evaluated using various optimizers, including Adam, Adagrad, RMSProp, and SGD by comparing accuracy and loss values. Additionally, other classification metrics such as precision, Recall, and F1-Score value are calculated. Finally, comparative analyses of these metrics across the optimizers are performed**.** The experimental results demonstrated that the Cor-GA_SNR-LSTM model effectively identifies a prominent subset of biomarker genes, enabling the LSTM model to accurately classify the gene expressions. The research model obtained an accuracy of 88.45%, a precision score of 87%, and a loss rate of 0.1612 using the ADAM Optimizer. In future research, different optimization techniques can be implemented to further enhance the classification outcomes.

REFERENCES

[1] Kubota, Yasuo, Misam Zawit, Jibran Durrani, Wenyi Shen, Waled Bahaj, Tariq Kewan, Ben Ponvilawan et al. "Significance of hereditary gene alterations for the pathogenesis of adult bone marrow failure versus myeloid neoplasia." Leukemia 36, no. 12 (2022): 2827-2834.

[2] Liu, Li, Alex Yick-Lun So, and Jian-Bing Fan. "Analysis of cancer genomes through microarrays and next-generation sequencing." Translational Cancer Research 4, no. 3 (2015).

[3] Ehigie, Adeola Folashade, Fiyinfoluwa Demilade Ojeniyi, Adetayo Aborisade, James Busayo Agboola, and Leonard Ehigie. "Transcriptomic analysis of differential gene expression in staphylococcus aureus-induced pneumonia in pediatrics based on microarray analysis." (2022).

[4] Hambali, Moshood A., Tinuke O. Oladele, and Kayode S. Adewole. "Microarray cancer feature selection: Review, challenges and research directions." International Journal of Cognitive Computing in Engineering 1 (2020): 78-97.

[5] Alshareef, Abdulrhman M., Raed Alsini, Mohammed Alsieni, Fadwa Alrowais, Radwa Marzouk, Ibrahim Abunadi, and Nadhem Nemri. "Optimal deep learning enabled prostate cancer detection using microarray gene expression." Journal of Healthcare Engineering 2022, no. 1 (2022): 7364704.

[6] Ogunbiyi, Temitope, Michael Adegoke, Adebisi Oluwatosin, Bamidele Aremo, Olufemi Adekunle, Emmanuel Ayoariyo, and Austin Udemba. "Improving Acute Leukemia Classification through Recursive Feature Elimination and Multilayer Perceptron Analysis of Gene Expression Data." International Journal of Data Science 5, no. 1 (2024): 33-49.

[7] Ali, Waleed, and Faisal Saeed. "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data." Processes 11, no. 2 (2023): 562.

[8] Rostami, Mehrdad, Kamal Berahmand, and Saman Forouzandeh. "A novel community detection based genetic algorithm for feature selection." Journal of Big Data 8, no. 1 (2021): 2.

[9] Sayed, Sabah, Mohammad Nassef, Amr Badr, and Ibrahim Farag. "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets." Expert Systems with Applications 121 (2019): 233-243.

[10] Lawrence, Morolake Oladayo, Rasheed Gbenga Jimoh, and Waheed Babatunde Yahya. "An efficient feature selection and classification system for microarray cancer data using genetic algorithm and deep belief networks." Multimedia Tools and Applications (2024): 1-42.

[11] Pragadeesh, C., Rohana Jeyaraj, K. Siranjeevi, R. Abishek, and G. Jeyakumar. "Hybrid feature selection using micro genetic algorithm on microarray gene expression data." Journal of Intelligent & Fuzzy Systems 36, no. 3 (2019): 2241-2246.

[12] Yuvaraj, V., G. Pandiyan, and G. Purusothaman. "Gene selection and modified long short term memory network based lung cancer classification using gene expression data." ICTACT Journal on Soft Computing 12, no. 2 (2022).

[13] Babichev, Sergii, Igor Liakh, and Irina Kalinina. "Applying a recurrent neural network-based deep learning model for gene expression data classification." Applied Sciences 13, no. 21 (2023): 11823.

[14] Wang, Zixuan, Yi Zhou, Tatsuya Takagi, Jiangning Song, Yu-Shi Tian, and Tetsuo Shibuya. "Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data." BMC bioinformatics 24, no. 1 (2023): 139.

[15] Alabdulqader, Ebtisam Abdullah, Aisha Ahmed Alarfaj, Muhammad Umer, Ala'Abdulmajid Eshmawi, Shtwai Alsubai, Tai-hoon Kim, and Imran Ashraf. "Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted convolutional neural network." Scientific Reports 14, no. 1 (2024): 15625.

[16] Mallick, Pradeep Kumar, Saumendra Kumar Mohapatra, Gyoo-Soo Chae, and Mihir Narayan Mohanty. "Convergent learning–based model for leukemia classification from gene expression." Personal and Ubiquitous Computing 27, no. 3 (2023): 1103-1110.

[17] Sahu, Bibhuprasad, and Sujata Dash. "Hybrid multifilter ensemble based feature selection model from microarray cancer datasets using GWO with deep learning." In 2023 3rd International Conference on Intelligent Technologies (CONIT), pp. 1-6. IEEE, 2023.

[18] Karim, Abdul, Azhari Azhari, Mobeen Shahroz, Samir Brahim Belhaouri, and Khabib Mustofa. "LDSVM: Leukemia cancer classification using machine learning." (2022).

[19] Ma'ruf, Firda Aminy, and Untari Novia Wisesty. "Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier." In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012011. IOP Publishing, 2019.

[20] Tabassum, Mujahid, and Kuruvilla Mathew. "A genetic algorithm analysis towards optimization solutions." International Journal of Digital Information and Wireless Communications (IJDIWC) 4, no. 1 (2014): 124-142.

[21] Deng, Xiongshi, Min Li, Shaobo Deng, and Lei Wang. "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification." Medical & Biological Engineering & Computing 60, no. 3 (2022): 663-681.

[22] Shukla, Alok Kumar, Pradeep Singh, and Manu Vardhan. "A new hybrid feature subset selection framework based on binary genetic algorithm and information theory." International Journal of Computational Intelligence and Applications 18, no. 03 (2019): 1950020.

[23] Gupta, Surbhi, Manoj K. Gupta, Mohammad Shabaz, and Ashutosh Sharma. "Deep learning techniques for cancer classification using microarray gene expression data." Frontiers in Physiology 13 (2022): 952709.

[24] Aburass, Sanad, Osama Dorgham, and Jamil Al Shaqsi. "A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe." Systems and Soft Computing 6 (2024): 200110.

[25] Hasan, Abid, and Md Akhtaruzzaman Adnan. "High dimensional microarray data classification using correlation based feature selection." In 2012 International conference on biomedical engineering (ICoBE), pp. 319-321. IEEE, 2012.

[26] Rezaee, Khosro, Gwanggil Jeon, Mohammad R. Khosravi, Hani H. Attar, and Alireza Sabzevari. "Deep learning-based microarray cancer classification and ensemble gene selection approach." IET Systems Biology 16, no. 3-4 (2022): 120-131.

[27] Du, Wen, Ting Gu, Li-Juan Tang, Jian-Hui Jiang, Hai-Long Wu, Guo-Li Shen, and Ru-Qin Yu. "Unimodal transform of variables selected by interval segmentation purity for classification tree modeling of high-dimensional microarray data." Talanta 85, no. 3 (2011): 1689-1694.

[28] Karimi, Sadegh, and Maryam Farrokhnia. "Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique." Chemometrics and Intelligent Laboratory Systems 139 (2014): 6-14.

[29] Furey, Terrence S., Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, and David Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16, no. 10 (2000): 906-914.

[30] Nguyen, Danh V., and David M. Rocke. "Tumor classification by partial least squares using microarray gene expression data." Bioinformatics 18, no. 1 (2002): 39-50.

[31] Cho, Sung-Bae. "Exploring features and classifiers to classify gene expression profiles of acute leukemia." International Journal of Pattern Recognition and Artificial Intelligence 16, no. 07 (2002): 831-844.

[32] Ocampo-Vega, Ricardo, Gildardo Sanchez-Ante, Marco A. de Luna, Roberto Vega, Luis E. Falcón-Morales, and Humberto Sossa. "Improving pattern classification of DNA microarray data by using PCA and logistic regression." Intelligent Data Analysis 20, no. s1 (2016): S53-S67.

[33] Rupapara, Vaibhav, Furqan Rustam, Wajdi Aljedaani, Hina Fatima Shahzad, Ernesto Lee, and Imran Ashraf. "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model." Scientific reports 12, no. 1 (2022): 1000.

[34] Sharma, Aman, and Rinkle Rani. "An optimized framework for cancer classification using deep learning and genetic algorithm." Journal of medical imaging and health informatics 7, no. 8 (2017): 1851-1856.