ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

ARTIFICIAL INTELEGENCE FOR DNA ANALYSIS TO PREDICT GENETIC DISEASES

GREESHMA SAI PRIYA, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

M. MANI, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

R. VIJAYA MARTINA, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

D. CHANAKYA, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

B. ESHWARARAO, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India eshwar.world@gmail.com

Abstract:

For the goal of classifying samples, gene expression analysis is used to determine the relative relevance of each gene. There have been a number of important findings and advancements in clinical care based on microarray data relating to gene expression profiles. Microarray data often have a limited sample size and a high dimension. Using an all-purpose categorization system would be problematic in this situation. There are some genes, however, that may not be useful in identifying the type of sample. A good feature (gene) selection strategy and an efficient gene extraction method are both required for accurate analysis of gene expression profiles in order to reduce the classification error rate. The AI and the correlation-based feature selection (CFS) were integrated in this paper into a hybrid technique. It was used as a classifier for ten gene expression profiles that were analyzed using AI with LOOCV (leave-one-out cross-validation).

Keywords: DNA, genetic diseases, AI, Microarray data.

I. Introduction

In recent years, the epidemic in breast cancer, diabetes, liver disorder, prostate cancer, colon tumor, obesity and many other heart diseases has become a challenge to global health. The dreadful diseases like cancer often proves to be life-altering, life threatening and fatal. Most often their symptoms stem from a genetic basis and a host of challenges demand for the prevention, diagnosis, treatment and cure of these diseases. As medicine plays a great role in saving human life, medical data classification has remained as one of the leading research areas in the domain of biomedical informatics, machine learning and pattern classification. Medical data classification as one of the key areas of datamining tasks. Medical data is often found to be heterogeneous, unorganized, high dimensional, noisy and associated with outliers. It includes both clinical data and genomic data, which is the basis of biomedical informatics. The domain of biomedical informatics has emerged due to the cross fertilization of Bioinformatics, Medical Informatics and Clinical Genomics.

The findings of an experiment suggest that this hybrid strategy reduces the number of characteristics needed for feature selection. For the 10 gene expression data set challenges investigated, the proposed method's classification error rate was the lowest. There were zero categorization errors in six of the gene expression profiles. In terms of categorization error rate, the new method outscored five other existing approaches. As a result, it has the potential to be an important tool for future studies analyzing gene expression.

During the last few decades, the global epidemic of breast cancer, diabetes and liver disease has posed a major threat to global health. Cancer is one of the most devastating, life-altering, and life-threatening diseases. There are a slew of hurdles that must be overcome in order to prevent, diagnose, treat, and cure these diseases. Medical data classification has remained a significant study subject in the realm of biomedical informatics, machine learning, and pattern classification since medicine plays such an important role in saving lives. One of the most important facets of data mining is the classification of medical records. Data from the medical field is typically found to be asymmetrical and prone to outliers. Biomedical informatics is built on the foundation of clinical and genomic data. Cross-fertilization among bioinformatics, medical informatics, and clinical genomics has produced the field of biomedical informatics as we know it today.

II. Literature Survey

Medical data mining is one of the most important applications of data mining in biomedical informatics, and it is the topic of this paper. Microarray medical datasets have hundreds of characteristics, whereas small medical datasets have fewer features. Machine intelligence and evolutionary computing techniques have been used by researchers for many years. The classifier, however, still has a lot of room for improvement utilizing machine intelligence and evolutionary computing.

III. Methodology:

Data mining:

In the past, the concept of identifying patterns in data has gone by many names, including data mining, knowledge extraction, information discovery, and processing of data patterns. [22] Data mining is the use of specific algorithms to extract patterns from data. Additional phases in the KDD process are necessary to ensure that relevant information is produced from the data, such as data selection, cleaning, incorporation of appropriate previous knowledge, and proper interpretation of the results.

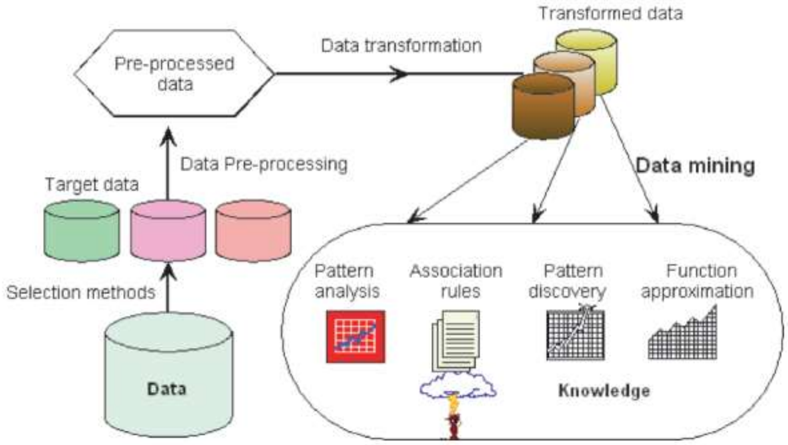


Figure 1. Steps of the knowledge discovery process

Steps of Knowledge Discovery

The data mining process is depicted in Fig1 as a series of sequential processes.

1. Identifying the purpose of the KDD process and developing an understanding of the application domain and applicable prior knowledge.
2. Defining the target data set. 1 Data mining using Swarm Intelligence 11
3. Cleaning and preprocessing: removing noise, addressing missing data fields.
4. Data slicing and dicing: selecting the most useful attributes to describe the data, depending on the task Reducing the amount of variables to be considered or finding an invariant representation of data using dimensionality reduction or transformation methods
5. Although the boundary between prediction and description isn't particularly crisp, it's there to help comprehend the general discovery aim of the KDD process.

The following data mining techniques are used to meet the goals of knowledge discovery:

- Clustering is the process of classifying data into a small number of distinct categories or clusters.
- For example, summarizing an association of rules and using multivariate visualization techniques to find a concise description of a subset of data.
- Identification of substantial interdependencies among variables by means of dependency modelling.

- Finding functional connections between variables through the study of regression functions that transfer each piece of information in the data to a meaningful prediction variable with a real-valued prediction.
- To learn a classification function, you must first learn how to classify a data point into one of several predefined categories.
- Discovering the most significant changes in data from previously measured or normative values
- Change and Deviation Detection.

A real time DNA analysis to predict genetic diseases and classification using AI techniques

When using a DNA microarray analysis, researchers can quickly collect massive amounts of data by measuring the expression of thousands of genes at once. Gene expression profiles are more objective, accurate, and dependable than traditional illness diagnostic methods. However, when it comes to data mining, there are a number of difficulties that need to be addressed. The dimensionality of these databases is a big issue. Additionally, there aren't enough samples to train and test the models that were built. Furthermore, only a small number of the many gene sets examined by microarrays are found to be relevant, while the majority of the other gene sets are redundant, noisy, or of considerably lower significance. The classifier is skewed by the presence of this gene collection, which reduces prediction accuracy and raises the analysis's cost. Microarray data analysis has several computational intelligence models to deal with these difficulties.

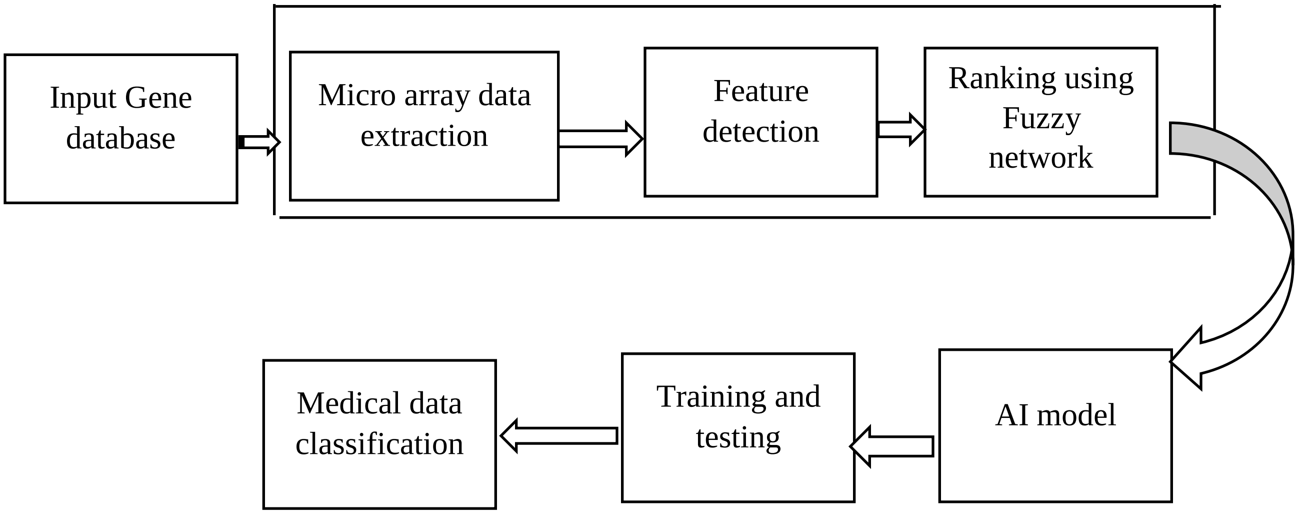

Input Gene database
Micro array data extraction
Feature detection
Ranking using Fuzzy network
AI model
Training and testing
Medical data classification

Figure 2. Block Diagram for Microarray using AI

In addition to deleting unnecessary and redundant features, feature selection plays an essential role in decreasing computing complexity. Gene selection in microarray data analysis can be accomplished in two ways: through the use of filters and wrappers. Employing mutual information (MI), as well as a genetic algorithm (GA) to integrate both filters and wrappers, a hybrid technique is created (GA). Next, various classifiers are trained on a subset of the features. Classifiers such as SVM, KNN, and Naive Bayes.

Contribution

- Further, to deal with the microarray medical data, introduced swarm intelligence techniques a modified African Buffalo optimization (MABO) and AI techniques have been applied to select most optimal features from micro array medical datasets like Prostate Cancer, Colon Tumor, ALL AML Leukemia, Leukemia1, Leukemia2, Breast Cancer, BrainTumor1 and SRBCT.
- Finally, to get most suitable classifiers for the above said datasets, models like Support Vector machine, AI, Naive Bayes, MLP, random forest, Ridge Regression (RR) are explored and results are compared.
- The various performance evaluation Measures Considered for comparison of results are accuracy, sensitivity, specificity, and F-score.
- In this study focuses on medical data mining which is regarded as one the core applications of data mining in the domain of biomedical informatics.
- Small medical datasets related to any disease have less number of features and in microarray medical data contains thousands of features. A lot of research has already been done by the researchers using machine intelligence and evolutionary computing techniques over the years.
- But still there is a lot of scope to improve the performance of classifier using machine intelligent and evolutionary computing techniques.

AI

Since 1990, a number of algorithms inspired by collective behavior (such as social insects or flocks of birds) have been put forth. NP-hard optimization issues, network routing, clustering, data mining, job scheduling, etc. AI and Ant Colonies Optimization (ACO) are now the most prominent algorithms in the swarm intelligence arena.

Iterative Optimization based on Particle Swarms (PSO) as the name implies, PSO utilizes a population of random solutions referred to as "particles" as its starting point. Unlike other methods of evolutionary computation, PSO assigns a velocity to each individual particle. Dynamic velocities of the search space's particles change depending on their past behavior. Since the search process progresses, the particles prefer to fly toward the best possible search area. A 'cornfield vector,' as it is known, is a flock of birds on the prowl for food. The PSO was originally developed to mimic that behavior.

What if a bunch of birds are roaming about an area looking for food? Only one piece of food has been found in the area being searched. The birds are clueless as to where the food is located. But they know where the food is and where their classmates are in relation to them. What's the greatest way to find food in this situation? Following the bird closest to the food is a successful technique.

To tackle optimization difficulties, PSO learns from the situation. As a particle in PSO, each solution is like a "bird" in the search space. There are fitness values assigned to all particles, and these values are assessed using the fitness function in order to optimize the particles' flight. (The particles follow the particles with the best answers so far as they fly through the issue space.) PSO begins with a sprinkling of random particles (solutions), and it then iteratively seeks for optimum values.

In the D-dimensional search space, each person is viewed as a single, volume-less particle (or "point"). Xi = is used to represent the ith particle (xi1, xi2, xiD). The following two 'best' values are applied to each particle at the end of each generation. To date, this is the particle's best former location (the place with the highest fitness value). The name of this parameter is pBest. It is symbolized by Pi = pi = pi = pi = pi = pi (pi1, pi2. . . piD). In order to change the particle's velocity in each dimension, the P vectors of the particle with the best fitness in the neighborhood, labelled l or g, and the current particle's P vector are mixed at each iteration. Individual's best position (P) influences the cognition component of velocity adjustment; the best in neighborhood influences the social component of velocity adjustment. The inertia factor, introduced by Shi and Eberhard [59] (to balance the global and local search), is now included in these equations:

$$\mathrm{v}_{id} = \omega^{*}\mathrm{v}_{id} + \eta_{\mathrm{x}}^{*} rand()^{*}(p_{id} - x_{id}) + \eta_{2}^{*} Rand()^{*}(p_{gd} - x_{id}) \tag{1}$$

$$\mathrm{x}_{id} = \mathrm{x}_{id} + v_{id} \tag{2}$$

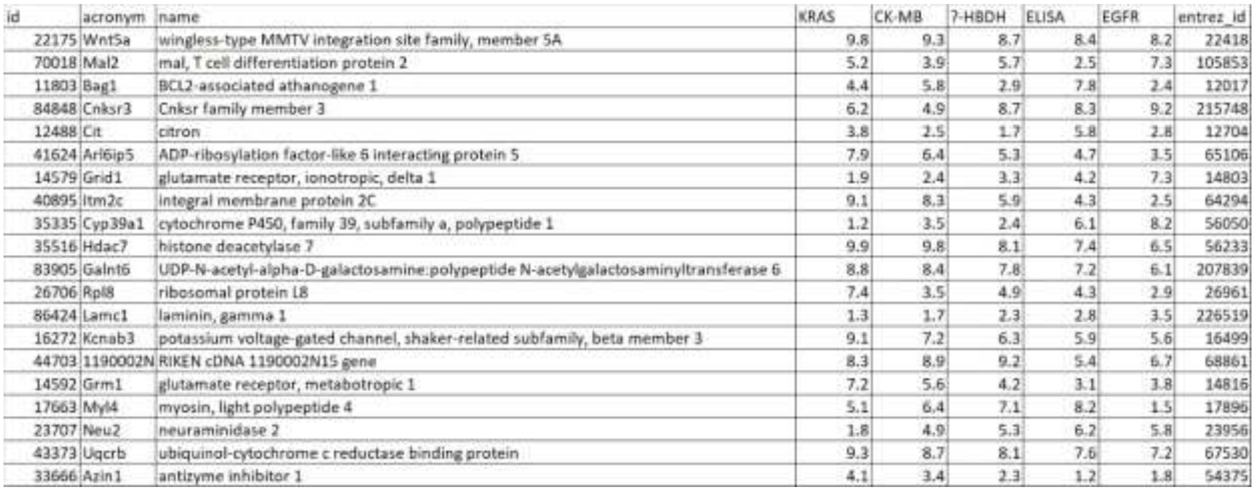
Where rand () and Rand () are two independent random numbers, and 1 and 2 are two learning parameters that control the influence of social and cognitive components. If the right-side total is greater than a predetermined value in (1.1), the velocity in that dimension is set to Vmax. A constraint is placed on the particle swarm's global exploring ability by limiting the particle velocities to the range [-Vmax, Vmax]. As a result, the probability of particles escaping the search area is decreased. Note that this does not limit the values of Xi inside the range [-Vmax, Vmax]; it merely restricts the maximum distance that a particle can move during one iteration of the algorithm ([19], [20], [21]). A summary of Pomeroy's core PSO algorithm can be found.

IV. Result

Finally, a modified African Buffalo optimization (MABO) is combined with Rough Set to select the most optimal features from microarray medical datasets like prostate cancer, colon tumor, leukemia, and breast cancer in order to improve the performance of the microarray data analysis swarm intelligence technique. This is followed by the use of AI, MSVM, EKNN, and INB as multi-classifiers for the above-mentioned datasets, and the results are compared to see which works best. Performance evaluation measures like as sensitivity, discriminative power, and F-score are all used to compare outcomes.

Dataset

Table 1. Data set



id	acronym	name	KRAS	CK-MB	7-HBDH	ELISA	EGFR	entrez_id
22175	Wnt5a	wingless-type MMTV integration site family, member 5A	9.8	9.3	8.7	8.4	8.2	22418
70018	Mal2	mal, T cell differentiation protein 2	5.2	3.9	5.7	2.5	7.3	105853
11803	Bag1	BCL2-associated athanogene 1	4.4	5.8	2.9	7.8	2.4	12017
84848	Cnksr3	Cnksr family member 3	6.2	4.9	8.7	8.3	9.2	215748
12488	Cit	citron	3.8	2.5	1.7	5.8	2.8	12704
41624	Arl6ip5	ADP-ribosylation factor-like 6 interacting protein 5	7.9	6.4	5.3	4.7	3.5	65106
14579	Grid1	glutamate receptor, ionotropic, delta 1	1.9	2.4	3.3	4.2	7.3	14803
40895	Itm2c	integral membrane protein 2C	9.1	8.3	5.9	4.3	2.5	64294
35335	Cyp39a1	cytochrome P450, family 39, subfamily a, polypeptide 1	1.2	3.5	2.4	6.1	8.2	56050
35516	Hdac7	histone deacetylase 7	9.9	9.8	8.1	7.4	6.5	56233
83905	Galnt6	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6	8.8	8.4	7.8	7.2	6.1	207839
26706	Rpl8	ribosomal protein L8	7.4	3.5	4.9	4.3	2.9	26961
86424	Lamc1	laminin, gamma 1	1.3	1.7	2.3	2.8	3.5	226519
16272	Kcnab3	potassium voltage-gated channel, shaker-related subfamily, beta member 3	9.1	7.2	6.3	5.9	5.6	16499
44703	1190002N	RIKEN cDNA 1190002N15 gene	8.3	8.9	9.2	5.4	6.7	68861
14592	Grm1	glutamate receptor, metabotropic 1	7.2	5.6	4.2	3.1	3.8	14816
17663	Myl4	myosin, light polypeptide 4	5.1	6.4	7.1	8.2	1.5	17896
23707	Neu2	neuraminidase 2	1.8	4.9	5.3	6.2	5.8	23956
43373	Uqcrb	ubiquinol-cytochrome c reductase binding protein	9.3	8.7	8.1	7.6	7.2	67530
33666	Azin1	antizyme inhibitor 1	4.1	3.4	2.3	1.2	1.8	54375

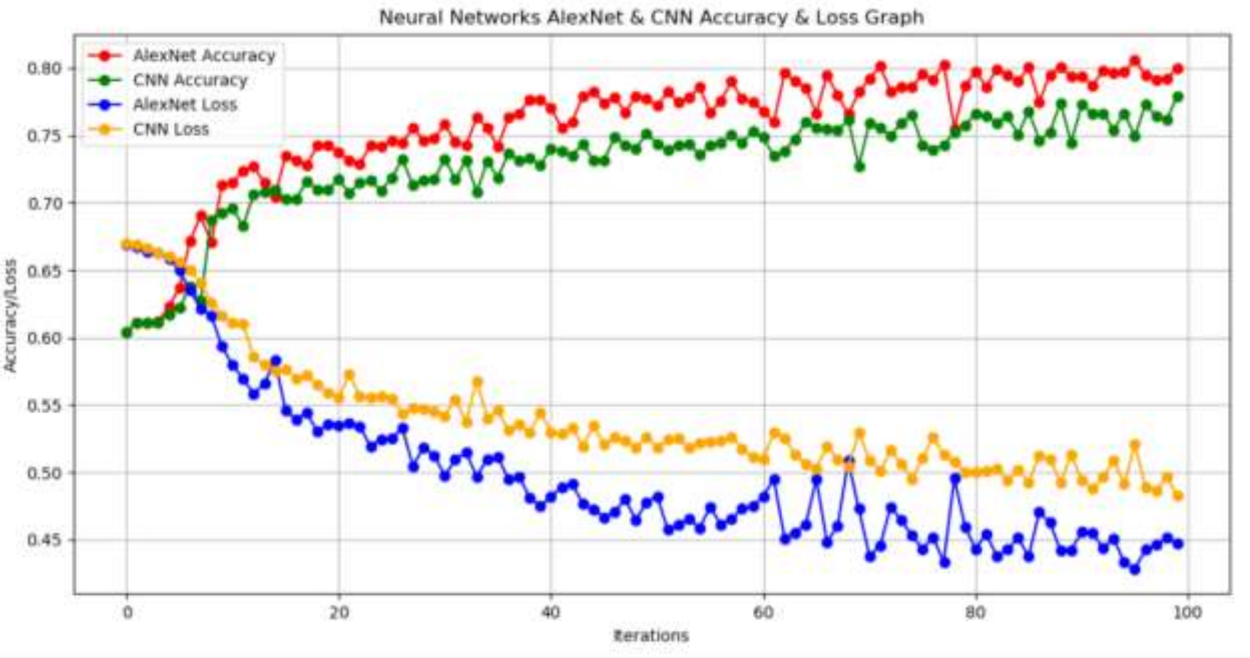


Figure 3. Simulation result of Neural Network Alex Net and CNN accuracy and loss

Contribution:

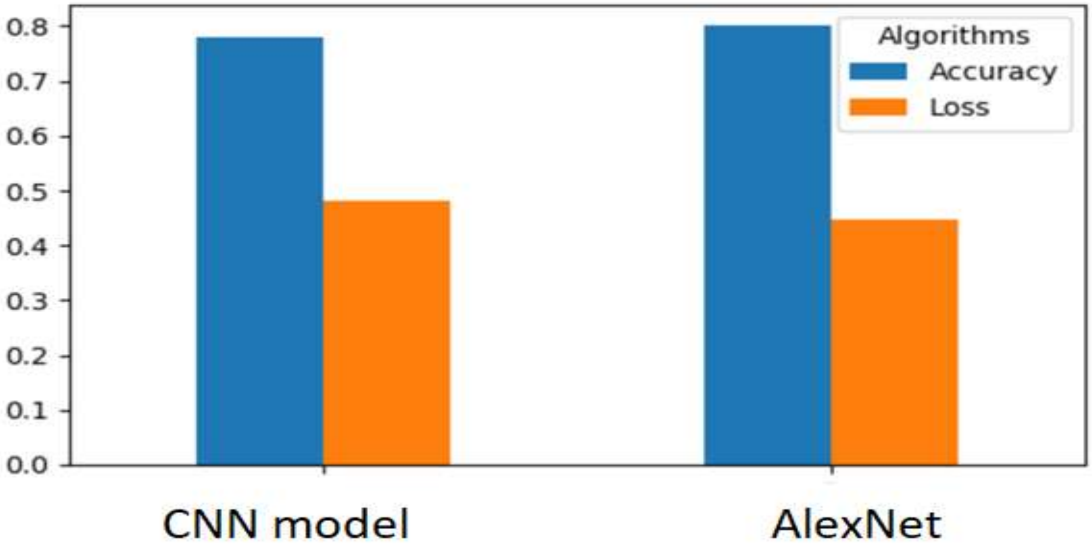
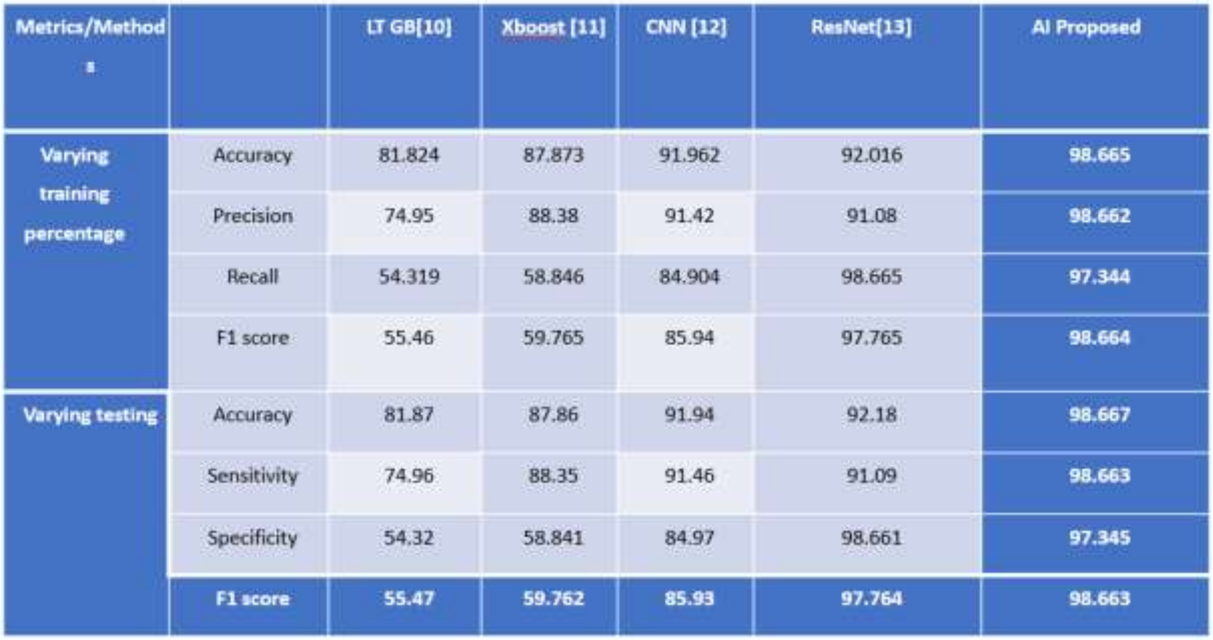


Figure 4. Simulation result of Algorithm CNN Model Alex Net

Table 2 Comparison of AI Accuracy and loss



Metrics/Method s		LT GB[10]	Xboost [11]	CNN [12]	ResNet[13]	AI Proposed
Varying training percentage	Accuracy	81.824	87.873	91.962	92.016	98.665
	Precision	74.95	88.38	91.42	91.08	98.662
	Recall	54.319	58.846	84.904	98.665	97.344
	F1 score	55.46	59.765	85.94	97.765	98.664
Varying testing	Accuracy	81.87	87.86	91.94	92.18	98.667
	Sensitivity	74.96	88.35	91.46	91.09	98.663
	Specificity	54.32	58.841	84.97	98.661	97.345
	F1 score	55.47	59.762	85.93	97.764	98.663

V. Conclusion:

Finally, to improve the performance of the microarray data analysis AI is combined with Rough Set to select most optimal features from micro array medical datasets like Prostate Cancer, Colon Tumor, Leukemia, and Breast Cancer. Finally, multi-classifiers for the above said datasets is applied using Modified Support vector machine (MSVM) , Extended K-nearest neighbor (EKNN), and Improved Naive Bayes(INB) are explored and results are compared. The various performance evaluation measures considered for comparison of results are accuracy, sensitivity, specificity, and F-score.

References

1. Kim, M., Fernwood, F., & Milinkovic, O. (2015). Hydra: gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics*, *31*(7), 1034-1043.

2. J.P. Gonzales, S.C. Madeira, A.L. Oliveira, Biggest: integrated environment for clustering analysis of time series gene expression data, BMC Res. Notes 2 (1) (2009) 124.

3. Jha, M., Guzzi, P., & Roy, S. (2019). Qualitative assessment of functional module detectors on microarray and RNASeq data. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *8*(1), 1-22.

4. Othman, M. S., Kumaran, S. R., & Yusuf, L. M. (2020). Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data. *IEEE Access*, *8*, 186348-186361.

5. Sudha, M. N., & Selvarajan, S. (2016). Feature selection based on enhanced cuckoo search for breast cancer classification in mammogram image. *Circuits and Systems*, *7*(04), 327.

6. Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, *6*, 29637-29647.

7. Hameed, S. S., Petinrin, O. O., Hashi, A. O., & Saeed, F. (2018). Filter-wrapper combination and embedded feature selection for gene expression data. *Int. J. Advance Soft Compu. Appl*, *10*(1), 90-105.

8. Ng, W. S., Neoh, S. C., Htike, K. K., & Wang, S. L. (2017). Particle Swarm Feature Selection for Microarray Leukemia Classification. *Progress in Energy and Environment*, *2*, 1-8.

9. Anger, Z., Bolat, B., & Diri, B. (2019). A probabilistic multi-objective artificial bee colony algorithm for gene selection. *Journal of Universal Computer Science*, *25*.

10. Jahwar, A., & Ahmed, N. (2021). Swarm intelligence algorithms in gene selection profile based on classification of microarray data: a review. *Journal of Applied Science and Technology Trends*, *2*(01), 01-09.

11. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., & Soboleva, A. (2010). NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic acids research*, *39*(suppl_1), D1005-D1010.

12. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, *96*(6), 2907-2912.

13. Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Seinfeld, I., Sharan, R., & Elkon, R. (2005). EXPANDER–an integrative program suite for microarray data analysis. *BMC bioinformatics*, *6*(1), 1-12.

14. Sharan, R., & Shamir, R. (2000, August). CLICK: a clustering algorithm with applications to gene expression analysis. In *Proc Int Conf Intell Syst Mol Biol* (Vol. 8, No. 307, p. 16).

15. Bhattacharya, A., & De, R. K. (2008). Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles. *Bioinformatics*, *24*(11), 1359-1366.

16. Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, *18*(suppl_1), S136-S14.

17. Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, *18*(suppl_1), S136-S144