Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Diverse Varieties of Cancer Forecasting System using Microarray Genes with the support of Improved Stacked Auto-Encoder

Santhosh Kumar C[1], Arivukkodi R[2], Muninathan N[3], Tarandeep Singh Walia[4], Deepender[5], Sridevi Sangeetha K S[6], R. Naveenkumar[7]

[1]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Email: cjsksag@gmail.com, https://orcid.org/0000-0003-4973-2352

[2]Assistant Professor, Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, tamilnadu, India. Email: arivukodir@maher.ac.in

[3]Scientist, Central Research Laboratory, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu 631552, Email: muninathan@maher.ac.in

[4]Associate Professor, School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India, Email : taran_walia2k@yahoo.com. Orcid id:- https://orcid.org/0000-0001-8127-3112

[5]Research Scholar, School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India, Email: deependerduhan6@gmail.com, Orcid : https://orcid.org/my-orcid?orcid=0000-0002-0529-4007

[6]Professor, Meenakshi College of Allied Health Sciences, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Chennai, tamilnadu, India, Email: ssks@maher.ac.in,000-0002-8175-1484

[7]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India, Email: drnk1983@gmail.com, 0000-0001-9033-9400

Abstract

Microarray gene expression data is one of the most commonly used gene expression datasets applied for cancer sample prediction. It consists of thousand expression levels of the genes in a single experiment. Cancer subtypes in microarray gene expression data are vague, indiscernible, imprecise and overlapping in nature which often lowers down the cancer prediction accuracy of the traditional classification models in general. Cancer prediction from gene expression data is an important and challenging area of research in the field of computational biology and bioinformatics. This research presents a deep learning approach to cancer detection, and to the identification of genes critical for the diagnosis of leukaemia,lung cancer and prostate cancer. The dimensionality of the data is handled using improved stacked auto encoder (ISAE) algorithm, whereby the features are classified using neural network.The proposed approach is enhanced by adding the regularization and reconstruction loss. Classification accuracy of ISAE is 97% that shows the effectiveness of the ISAE and this approach outperforms the existing state-of-art techniques.

Keyword: Gene expression, cancer prediction, optimal feature, error rate, auto-encoder, and regularisation.

1. Introduction

All the cells in an organism contain the same set of genes. Thus, there would be the same set of proteins present in all the cells of an organism. If the distribution of proteins in all the cells were the same, functional behavior of the cells would be similar. In practice, this does not happen, rather the distribution of the proteins vary with tissue types [1]. Due to this fact, different organs function in different ways. Thus the basic question is why this difference. In order to answer to this question, central dogma of molecular biology have to be studied. Although all the cells in an organism contain the same set (genome) of genes, these genes may not be equally active to get

the information coded by them copied into proteins. Due to variation of this activity level of genes, the distribution of proteins in the cells vary with tissue types. This variation of distribution of proteins leads to variation in functional behavior of cells [2]. The amount of mRNA produced from a gene may be considered as a measure of activity level of gene. This activity level is called gene expression. Current technologies allows us to measure expression levels of genes, which forms a very important resource, called gene expression data [3].

One of the most useful experimental data for the investigation of genetic analysis by utilising various computerized modelling approaches that falls within the bioinformatics category. As previously stated, a criterion of gene expression is the amount of mRNA that a gene produces. Expression levels results from higher mRNA levels, and vice versa. Various experimental approaches are used to evaluate gene expression. Real-time PCR [5] and Northern blot [4] are methods for determining the expression levels of genes. Microarray technology [6], a more advanced method of measuring gene expression, enables the simultaneous assessment of the expression of thousands of genes in a cell in a particular tissue [7].

A DNA microarray, sometimes referred to as a gene chip, DNA chip, or biochip, is a group of minute DNA patches joined to a solid surface [8]. Concurrent evaluation of the levels of several distinct genes' expression is done using DNA microarrays. Picomoles (10−12 moles), or "probes," of a certain DNA sequence are present in each DNA location (or reporters). They may be short gene segments or other DNA fragments that are used in accordance with rigorous protocols to hybridise a target cDNA sample [9]. A microarray experiment may run several genetic tests concurrently since each array can hold tens of thousands of probes. Therefore, many different sorts of research have been significantly expedited using arrays [10].

Monitoring the amount of gene expression at the genome size is one of the goals of a microarray experiment. By examining how the genes' expression changes, patterns may be deduced and new biological understandings can be achieved [11]. The expression levels for all genes under an experimental condition are collectively referred to as the sample expression profile, while the expression levels for a given gene are collectively referred to as the gene expression profile across several experimental circumstances. Additional layers of annotation can be added to the gene or the sample after we have the gene expression matrix [12].

Although the high dimension of the genes and the small number of samples are efficient at overcoming the obstacles of gene expression and classification, microarray data formats are essential for identifying and classifying many types of malignant tissues and illnesses [13]. Since just a few genes adequately characterise the illness physiologically, it is challenging and time-consuming to interpret disease-causing genes [14]. Early illness detection allows for the development of potent treatments for conditions like cancer. Furthermore, it is challenging, time-consuming, and expensive to find effective genes in a lab setting. It is feasible to decrease classification mistakes as well as the amount of time needed to complete the processing to the appropriate level by automating the gene separation from microarray data [15].

The dimensionality and the feature selection is attained using ISAE, which is enhanced to handle the effective feature selection process. Using an updated SAE approach, the selection process and the incidence of classification mistake are addressed. The gradient removal and overfitting caused by deep learning networks learned on small datasets can be substantially mitigated by it. Typically, an unique reconstruction loss is introduced that can improve selection accuracy. The significant features are retrieved and considered for classification that is done by neural network.

The remainder of the article is organized as follows: the gene expression-based disease prediction approaches are reviewed in Section 2, the proposed feature selection as well as classification using improved stacked auto encoder (ISAE) with neural network is elucidated in Section 3, the results are detailed with graphical illustration in Section 4, and the article is concluded in Section 5.

2. Related works

Vidaki, A., et al [3] used the artificial neural network to predict the age from the DNA methylation and it is used in the field of forensic. Garro, B. A., et al [4] classified the DNA microarrays with the help of ANN and Artificial Bee Colony optimization algorithm. This algorithm predicts, approximates, and classifies the genomic data. Atkov, O. Y., et al [5] included the clinical parameters and genetic polymorphism for the identification of heart

disease. ANN is applied into the system and the disease is identified. Coppedè, F., et al [33] developed the ANN approach to identify the risk of down syndrome where the algorithm identifies the chromosome damage and polymorphism. Khan, J., et al [6] used the ANN approach to gene expression profiling, which predicts and classifies the gene. The gene prediction helps in the identification of cancer in humans. Koçer, S., & Canal, M. R. [7]incorporated the genetic algorithm and ANN approach which classifies the epilepsy disease. The process prediction of disease with ANN is a complicated and identification disease from the genomic data is tedious when the genes are complicated to assign any biomarkers.

Ramirez, R., et al [8] proposed a graph convolution neural network (GCNN). Finding gene markers for cancer in its early diagnosis and treatment methods has become difficult as a result. To forecast the cancer survival rate, a variety of techniques have been developed, including regression-based Cox-PH, artificial neural networks, and more recently, deep learning techniques.Jia, P., et al[9] proposed a deep generative neural network (DGNN). Utilizing a deep variational autoencoder (VAE) model, latent vectors in a low-dimensional space may be created from thousands of genes. Then, these encoded vectors may properly predict drug response, perform better than conventional signature-gene based methods, and effectively manage the overfitting issue. Khan, M. A., et al[10] discussed adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN) to guarantee the correctness of the imputed drug response in both cell lines and cancer samples, we use stringent quality assessment and validation techniques, such as evaluating the influence of cell lineage, cross-validation, cross-panel evaluation, and application in independent clinical data sets.

The existing approaches has no effective feature selection technique where the optimization approaches lack in exploration and exploitation ability. The drawback is rectified in enhanced MFO technique. The process of training issue and the occurrence of error in classification is handled using improved SAE technique.It can somewhat compensate for the gradient removal and overfitting brought on by deep learning networks trained on limited datasets. A novel reconstruction loss that can increase the detection performance was created as a regular term.

3. Proposed Methodology

The proposed framework formulates cancer classification from microarray gene expression data as a high-dimensional nonlinear optimization problem, where the objective is to learn a compact latent representation while preserving discriminative information. Let the gene expression dataset be represented as $X = \{x_1, x_2, \dots, x_n\}$, where each sample $x_i \in \mathbb{R}^d$corresponds to d-dimensional gene features with $d \gg n$. Due to scale heterogeneity and noise in microarray data, normalization is first performed using min–max scaling to constrain the feature space within a bounded interval, expressed as

$$x_i^{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

where $x_{min}$and x_{max}denote the minimum and maximum values across each feature dimension. This normalization ensures numerical stability during gradient-based optimization.

The feature learning mechanism is modeled using an improved stacked autoencoder architecture, where each autoencoder learns a nonlinear mapping from input space to latent space. The encoder function is defined as

$$h^{(1)} = f(W^{(1)}x + b^{(1)})$$

where $W^{(1)} \in \mathbb{R}^{k \times d}$is the weight matrix, $b^{(1)}$is the bias vector, and $f(\cdot)$is a nonlinear activation function such as sigmoid or ReLU. The decoder reconstructs the input using

$$\hat{x} = g(W^{(1)T}h^{(1)} + b'^{(1)})$$

where $g(\cdot)$is typically symmetric to $f(\cdot)$. For deeper representation learning, multiple encoders are stacked such that

$$h^{(l)} = f(W^{(l)}h^{(l-1)} + b^{(l)})$$

for $l = 2,3, \ldots, L$, enabling hierarchical abstraction of gene interactions. The reconstruction process at each layer is similarly defined as

$$\hat{h}^{(l-1)} = g(W^{(l)T} h^{(l)} + b'^{(l)})$$

To address overfitting and enhance generalization in small-sample scenarios, dropout regularization is incorporated into the encoding process. A Bernoulli mask $r \sim \text{Bernoulli}(p)$is applied to the hidden units, yielding

$$\tilde{h}^{(l)} = r \odot h^{(l)}$$

where $\odot$denotes element-wise multiplication and $p$is the retention probability. This stochastic neuron deactivation prevents co-adaptation of features and improves robustness.

The core objective of the improved stacked autoencoder is to minimize a hybrid loss function that combines reconstruction fidelity and regularization. The reconstruction loss is formulated as

$$\mathcal{L}_{rec} = \frac{1}{2n} \sum_{i=1}^{n} \| x_i - \hat{x}_i \|^2$$

To control model complexity and avoid overfitting, an L_2regularization term is introduced on the weights:

$$\mathcal{L}_{reg} = \lambda \sum_{l=1}^{L} \| W^{(l)} \|_F^2$$

where $\lambda$is the regularization coefficient and $\|\cdot\|_F$denotes the Frobenius norm. Additionally, a sparsity constraint is imposed on hidden activations to encourage selective feature learning:

$$\mathcal{L}_{sp} = \sum_{j=1}^{k} KL(\rho \| \hat{\rho}_j)$$

where $KL(\cdot)$represents Kullback–Leibler divergence, $\rho$is the desired sparsity level, and $\hat{\rho}_j$is the average activation of hidden neuron j.

The overall optimization objective becomes

$$\mathcal{L}_{total} = \beta \mathcal{L}_{rec} + \mathcal{L}_{reg} + \gamma \mathcal{L}_{sp}$$

where $\beta$and γbalance reconstruction accuracy and sparsity constraints. The parameters $\theta = \{W, b\}$are optimized using gradient descent with backpropagation:

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta \mathcal{L}_{total}$$

where ηdenotes the learning rate.

After feature extraction, the learned latent representation $h^{(L)}$is forwarded to a deep neural network classifier. The classification layer computes the probability distribution over classes using softmax activation:

$$P(y = c \mid h^{(L)}) = \frac{\exp\,(w_c^T h^{(L)} + b_c)}{\sum_{j=1}^{C} \exp\,(w_j^T h^{(L)} + b_j)}$$

where $C$is the number of classes. The classification loss is defined using cross-entropy:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{n} \sum_{c=1}^{C} y_{ic} \log\, P(y = c \mid h_i^{(L)})$$

The final joint optimization integrates representation learning and classification as

$$\mathcal{L} = \mathcal{L}_{total} + \alpha \mathcal{L}_{cls}$$

where αcontrols the contribution of classification loss. The gradient flow across the stacked architecture ensures end-to-end learning of discriminative gene features.

To further enhance robustness, batch-wise updates are employed:

$$\theta = \theta - \eta \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \mathcal{L}^{(i)}$$

where $m$is the batch size. The final decision rule assigns each sample to the class with maximum posterior probability:

$$\hat{y} = \arg\max_c P(y = c \mid h^{(L)})$$

This formulation ensures that the proposed ISAE-based framework effectively reduces dimensionality, preserves biologically relevant gene interactions, and achieves high classification accuracy by jointly optimizing reconstruction and discriminative objectives in a unified deep learning paradigm.

Figure 1 illustrates the end-to-end architecture of the proposed cancer prediction framework based on an Improved Stacked Autoencoder (ISAE) integrated with a Deep Neural Network (DNN). The workflow begins with the acquisition of microarray gene expression datasets comprising leukemia, lymphoma, and prostate cancer samples, which are inherently high-dimensional and noisy. These datasets are first subjected to a pre-processing stage, where normalization and data shuffling are performed to ensure uniform feature scaling and to mitigate overfitting caused by biased sample distributions.

Figure 1 conveys a hybrid learning paradigm where unsupervised feature extraction via stacked autoencoders is seamlessly integrated with supervised classification using a deep neural network. The architecture effectively addresses the curse of dimensionality, enhances feature discrimination, and improves classification accuracy by leveraging deep hierarchical representations of gene expression data.

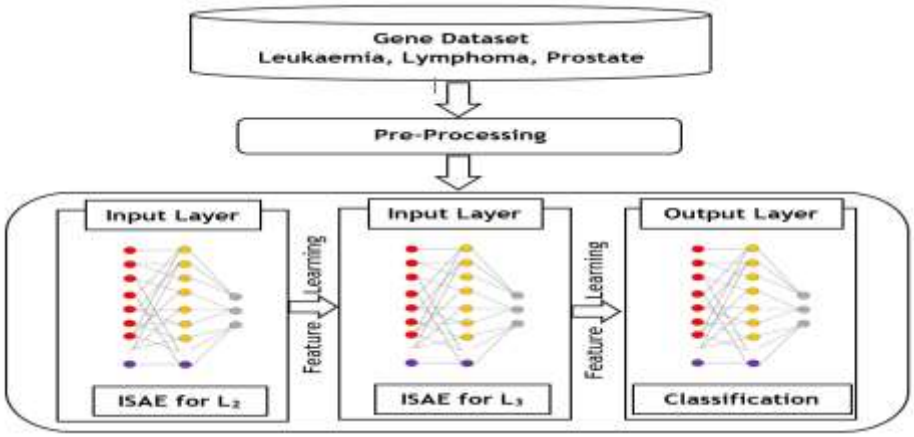

Figure 1: Improved Stack Auto Encoder with Deep Neural Network

4. Result and Discussion

This section illustrates the numerical outcomes of the existing and proposed methodology.The existing approaches namely GCNN [8], DGNN [9], ANN [10], and ANFIS [10] are compared with the proposed ISAE. These approaches are investigated using the accuracy, precision and error rate.

4.1. Dataset Details

Three microarray datasets, including those for leukaemia, lymphoma, and prostate cancer, were used in the trial. These datasets are accessible to public online. 72 occurrences, 3572 features, and 2 classes make up the leukaemia dataset; 77 instances, 2647 features, and 2 classes make up the lymphoma dataset. There are 102 instances, 2135 characteristics, and 2 classes in the prostate dataset. The performance test and training dataset for this model are used to evaluate it. The testing and training datasets are used to evaluate the performance of this model. Of the data gathered, 40% are thought to be for learning, while 60% are thought to be for testing. The dataset description for the research project is shown in Table 1.

Table 1: Dataset Description

Dataset	Instances	Features	Classes	Source
Leukemia	72	3572	2	[11]
Lymphoma	77	2647	2	[12]
Prostate	102	2135	2	[41]

Accuracy: Accuracy refers to how close the determined value from the classified occurrences is to the true value. The representation of quantitative bias and persistent flaws is known as accuracy. It is also the recognition (both TP and TN values) amongst number of the assessed classes, as well as the estimation's similarity to the true value. Variation between the outcome and genuine resulting values occurs when the lowest accuracy occurs. It's the proportion of successful fall detection to the number of information examined. The rate of accuracy is given n Figure 3 and Table 2. It's calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{---------(9)}$$

Table 2. Comparison of Accuracy

Dataset Name	Iteration Count	Existing				Proposed
		GCNN	DGNN	ANN	ANFIS	ISAE
Leukemia	50	81	85	83	87	93
	100	83	87	86	89	95
	150	84	88	88	90	96
	200	85	89	89	91	97
Lymphoma	50	81.5	84	84	86	94
	100	82	86	85	88	95
	150	83	87	88.5	89	96
	200	86	88	89.5	90.5	96.5
Prostate	50	82	84	81	86.5	93
	100	82.5	85	86	89	94
	150	84	87	87	89.5	95
	200	85	88.5	89.5	91	97

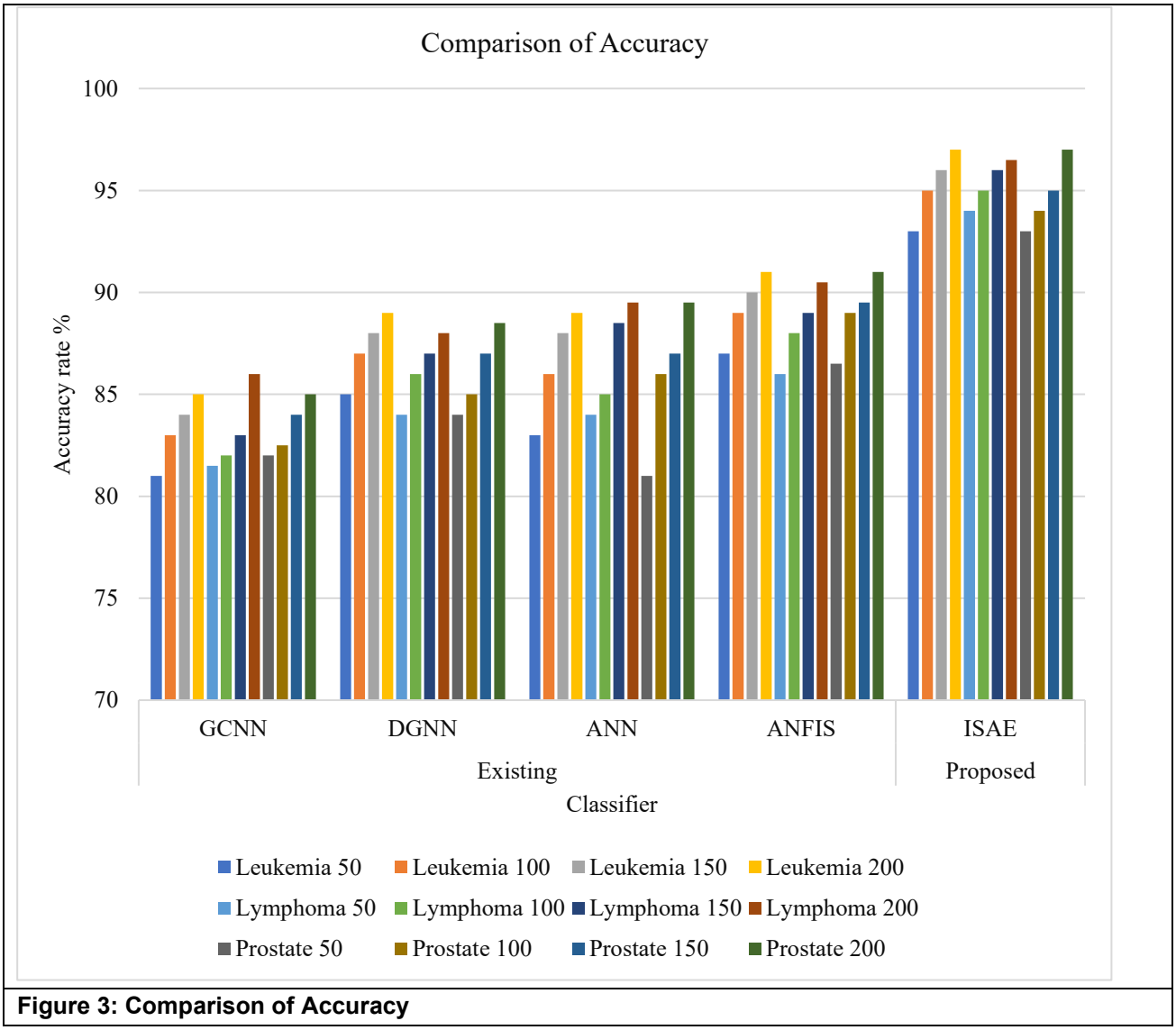

Figure 3: Comparison of Accuracy

Figure 3 indicates comparison of Accuracy for three different datasets namely Leukemia, Lymphoma, and Prostate. The proposed approach is compared with existing techniques namely GCNN, DGNN, ANN, and ANFIS. The accuracy of ISAE is{12%, 8%, 10%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 50 iterations. The accuracy of ISAE is {12%, 8%, 9%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 100 iterations. The accuracy of ISAE is {12%, 8%, 8%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 150 iterations. The accuracy of ISAE is {12%, 8%, 8%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 200 iterations. The accuracy of ISAE is {12.5%, 10%, 10%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 50 iterations. The accuracy of ISAE is {13%, 9%, 10%, and 7%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 100 iterations. The accuracy of ISAE is {13%, 9%, 7.5%, and 7%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 150 iterations. The accuracy of ISAE is {10.5%, 8.5%, 7%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 200 iterations. The accuracy of ISAE is {11%, 9%, 12%, and 6.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 50 iterations. The accuracy of ISAE is {11.5%, 9%, 8%, and 5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 100 iterations. The accuracy of ISAE is {11%, 8%, 8%, and 5.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 150 iterations. The accuracy of ISAE is {12%, 8.5%, 7.5%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 200 iterations. The accuracy of ISAE is higher and it outperforms the existing state of art algorithms.

Precision: The positive quantitative value or precision indicates the proximity of the observation and the relevance among the values discovered. The percentage rate of p The terms precision and accuracy are synonymous. It is calculated using True Positive (TP) and False Positive (FP) rates. The fraction of positive values in the entire population is related to precision. The accuracy estimate for a given issue in the classification stage is the number of real positive qualities (i.e. the count of the item correctly labelled as positive classes). The rate of precision is given in Table 3 and Figure 4. As a result, the algorithm with high precision creates more required data than unnecessary data.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \text{---------(10)}$$

Table 3. Comparison of Precision

Dataset Name	Iteration Count	Existing				Proposed
		GCNN	DGNN	ANN	ANFIS	ISAE
Leukemia	50	83	85	83	86	93
	100	83.5	86	86	89	95
	150	84	88	88	90	96
	200	85	89	89	93	97.5
Lymphoma	50	83.5	84	84	86	94
	100	86	86	85	88	95
	150	87	86.5	88.5	89	95.5
	200	88	88	89.5	90.5	96
Prostate	50	86	84	83	86.5	93
	100	86.5	85	86	89	94
	150	87	86	87	89.5	95
	200	88	88.5	89.5	93	96

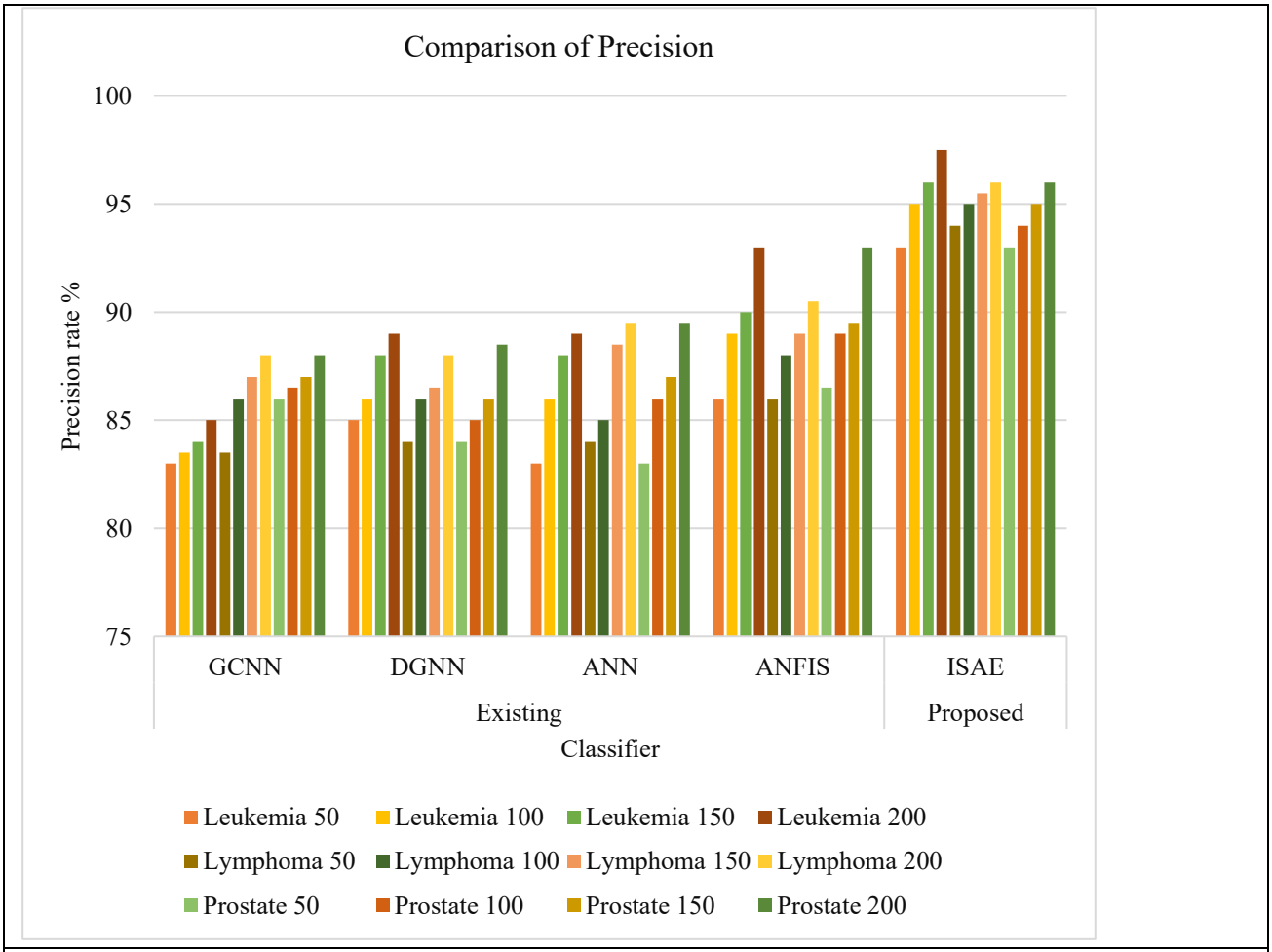

Figure 4: Comparison of Precision

Figure 4 indicates comparison of Precision for three different datasets namely Leukemia, Lymphoma, and Prostate. The proposed approach is compared with existing techniques namely GCNN, DGNN, ANN, and ANFIS. The precision of ISAE is {10%, 8%, 10%, and 7%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 50 iterations. The precision of ISAE is {11.5%, 9%, 9%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 100 iterations. The precision of ISAE is {12%, 8%, 8%, and 6%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 150 iterations. The precision of ISAE is {12.5%, 8.5%, 8.5%, and 4.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Leukemia with 200 iterations. The precision of ISAE is {10.5%, 10%, 10%, and 8%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 50 iterations. The precision of ISAE is {9%, 9%, 10%, and 7%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 100 iterations. The precision of ISAE is {8.5%, 9%, 7%, and 6.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 150 iterations. The precision of ISAE is {8%, 8%, 6.5%, and 5.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Lymphoma with 200 iterations. The precision of ISAE is {7%, 9%, 10%, and 6.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 50 iterations. The precision of ISAE is {7.5%, 9%, 8%, and 5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 100 iterations. The precision of ISAE is {8%, 9%, 8%, and 5.5%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 150 iterations. The precision of ISAE is {8%, 7.5%, 6.5%, and 3%} higher than {GCNN, DGNN, ANN, and ANFIS} for Prostate with 200 iterations. The precision of ISAE is higher and it outperforms the existing state of art algorithms.

Error Rate: Errors in digitalisation are caused by a variety of factors, including noise, distortion, and data processing disturbances. It's a ratio of competency rates. The mistake rate is the percentage of patterns that the decision-making framework incorrectly classifies. By multiplying the total of the FP and FN values by the sum of the TP,TN, FP, and FN values, the error rate is computed. The rate of error is given in Figure 5 and Table 4.

Table 4. Comparison of Error Rate

Dataset Name	Iteration Count	Existing				Proposed
		GCNN	DGNN	ANN	ANFIS	ISAE
Leukemia	50	0.578	0.561	0.567	0.551	0.498
	100	0.598	0.569	0.578	0.567	0.512
	150	0.612	0.571	0.591	0.571	0.534
	200	0.623	0.574	0.623	0.581	0.547
Lymphoma	50	0.561	0.541	0.557	0.561	0.488
	100	0.588	0.549	0.558	0.577	0.492
	150	0.592	0.553	0.561	0.581	0.504
	200	0.613	0.564	0.573	0.591	0.517
Prostate	50	0.571	0.551	0.563	0.561	0.471
	100	0.577	0.564	0.574	0.577	0.479
	150	0.581	0.572	0.578	0.581	0.481
	200	0.601	0.589	0.573	0.595	0.499

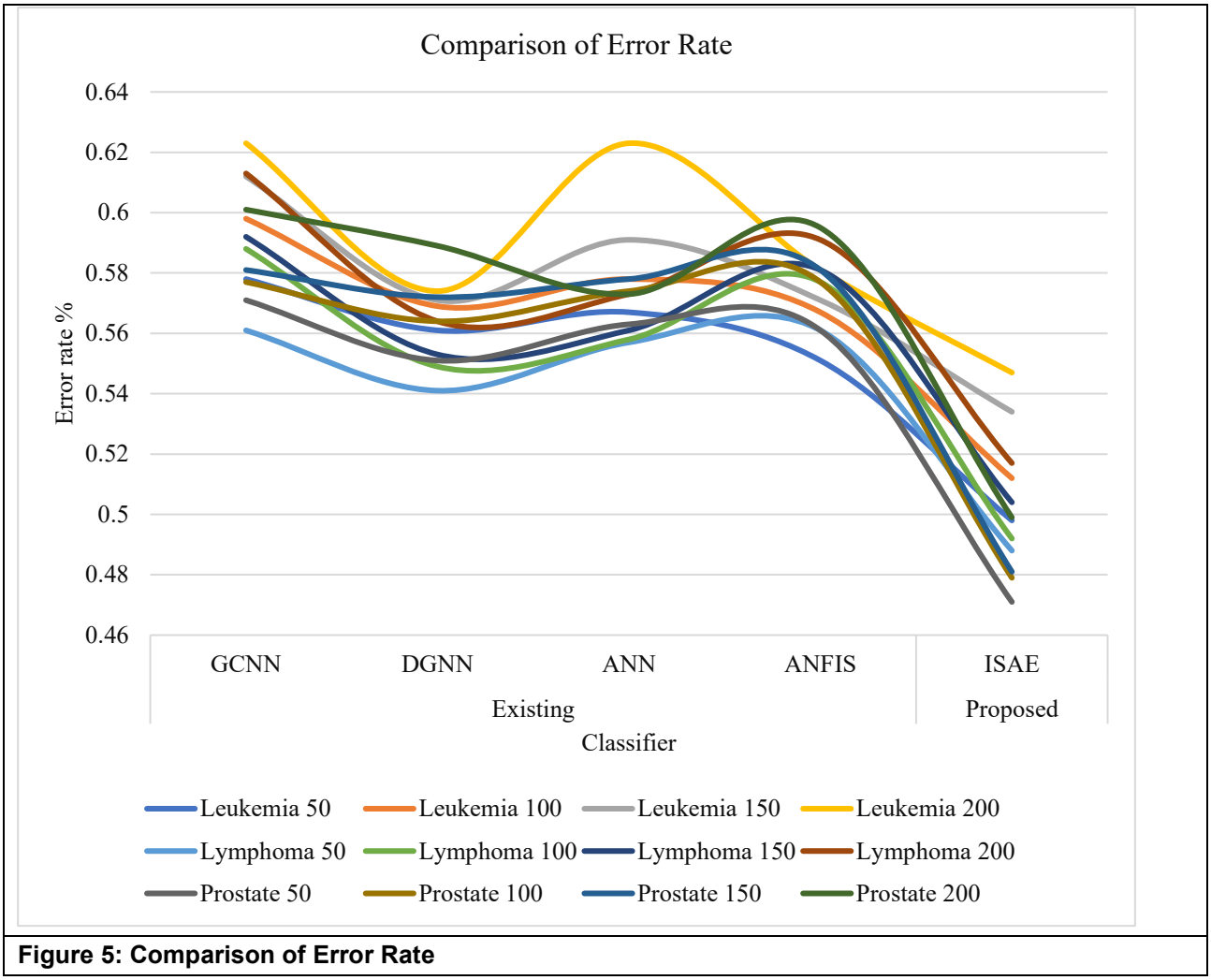

Figure 5: Comparison of Error Rate

Figure 5 indicates comparison of error rate for three different datasets namely Leukemia, Lymphoma, and Prostate. The proposed approach is compared with existing techniques namely GCNN, DGNN, ANN, and ANFIS. The ISAE attains minimal error rate over other existing approaches that indicates the effectiveness of the proposed approach.

5. Conclusion

One of the most often utilised gene expression datasets for cancer sample prediction is microarray gene expression data. An important and difficult topic of research in the fields of computational biology and bioinformatics is the prediction of cancer using gene expression data. In this research, a deep learning method to detection of disease and the discovery of leukaemia, lung cancer, and prostate cancer diagnostic genes is presented. improved stacked auto encoder (ISAE) is used to manage the data's dimensionality as well as for the retrieval of features, and deep learning technique is used to classify the features. The inclusion of the regularisation and reconstruction loss improves the suggested strategy. The classification accuracy of ISAE based DL is 97%, demonstrating its usefulness and outperforms current state-of-the-art methods.

Reference

1. Vidaki, A., Ballard, D., Aliferi, A., Miller, T. H., Barron, L. P., & Court, D. S. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Science International: Genetics*, *28*, 225-236.
2. Garro, B. A., Rodríguez, K., & Vázquez, R. A. (2016). Classification of DNA microarrays using artificial neural networks and ABC algorithm. *Applied Soft Computing*, *38*, 548-560.

3. Atkov, O. Y., Gorokhova, S. G., Sboev, A. G., Generozov, E. V., Muraseyeva, E. V., Moroshkina, S. Y., &Cherniy, N. N. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of cardiology*, *59*(2), 190-194.
4. Coppedè, F., Grossi, E., Migheli, F., & Migliore, L. (2010). Polymorphisms in folate-metabolizing genes, chromosome damage, and risk of Down syndrome in Italian women: identification of key factors using artificial neural networks. *BMC medical genomics*, *3*(1), 42.
5. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ... & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, *7*(6), 673-679.
6. Koçer, S., & Canal, M. R. (2011). Classifying epilepsy diseases using artificial neural networks and genetic algorithm. *Journal of medical systems*, *35*(4), 489-498.
7. Ramirez, R., Chiu, Y. C., Zhang, S., Ramirez, J., Chen, Y., Huang, Y., & Jin, Y. F. (2021). Prediction and interpretation of cancer survival using graph convolution neural networks. *Methods*, *192*, 120-130.
8. Sumit Ramswami Punam. (2025). Electromagnetic–Thermal Co-Optimization Models for Hybrid Solar Collectors under Spectral Variability. *IAECES Journal of Electronics and Communication Engineering*, 34–41.
9. S.Poornimadarshini. (2025). Automated Trust Lifecycle Management Using Smart-Contract-Driven Feedback Control. *Journal of Scalable Data Engineering and Intelligent Computing*, 37-44.
10. K P Uvarajan(2025). Predicting Pneumonia Progression in Children Using Spatio-Temporal Graph Neural Networks (GNN-PulmNet). *Journal of Computational Medicine and Informatics* , *1*(1), 42-51.
11. Jia, P., Hu, R., Pei, G., Dai, Y., Wang, Y. Y., & Zhao, Z. (2021). Deep generative neural network for accurate drug response imputation. *Nature communications*, *12*(1), 1-16.
12. Khan, M. A., Zafar, A., Farooq, F., Javed, M. F., Alyousef, R., Alabduljabbar, H., & Khan, M. I. (2021). Geopolymer concrete compressive strength via artificial neural network, adaptive neuro fuzzy interface system, and gene expression programming with K-fold cross validation. *Frontiers in Materials*, *8*, 621163.
13. https://web.stanford.edu/~hastie/CASI_files/DATA/Leukemia.html
14. https://ico2s.org/datasets/microarray.html
15. https://ico2s.org/datasets/microarray.html