A Learning Model for Acute Myeloid Leukemia Prediction Using Dense Polynomial Dimensionality-Based Predictor

K. Venkatesh[*1], S. Pasupathy[2], S. P. Raja[3]
[1,2]Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[3]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
Emails:venkiur01@gmail.com; pathyannamalai@gmail.com; avemariaraja@gmail.com

Abstract

Analysis of microarray data is extremely complex and considered as a hot topic in recent research. Acute Myeloid Leukemia (AML) prediction based on machine learning shows huge impact on prediction which automatically diagnoses the disease severity and any malfunctions. It is important to design the relevant classifier that processes the large data volume with large data size. Deep learning is an updated machine learning approach for mitigating these issues. It is easy to handle the huge volume of data because of the large number of hidden layers. The proposed classification methodology is used for understanding the training of the proposed Dense Polynomial Dimensionality based Predictor Model (DPDPM). The hidden neuron numbers are large in a sufficient way where the proposed DPDPM is elaborated to predict AML. AML and ALL samples are classified using five layers in the deep network model. The data is partitioned as 20% data and 80% data testing and training in the network. Compared with other classifiers, the satisfying outcome from the proposed DPDPM is higher and fulfilling. The validation is done in three datasets: Kaggle, Gene expression and Bio GPS and it gives 96% accuracy, 94% precision, 96% recall, 96% F1-score, and 98% AUROC while executing with Kaggle; then, 95.50% accuracy, 94% precision, 95% recall, 96% F1-score, and 96% AUROC is achieved while executing with Gene expression and finally 98% accuracy, 94.5% precision, 98.5% recall, 96% F1-score, and 94% AUROC is achieved while executing with Bio GPS. Based on this analysis, it is proven that the model works well with the proposed DPDPM and establishes a better trade-off.

Keywords: Acute leukaemia, prediction; deep learning; validation; dense network model

1. Introduction

These Human beings have the information about heredity commonly saved in the gene. This kind of disease generates a disorder in the human body's expressions of the gene because of the mutation in one gene, a combination of gene mutation, different gene mutations, different factors related to environments and damage to chromosomes. Few mutations or genetic disorders are the reason for developing human body cancer [1]. In addition, the gene mutation turns into a different new virus generation. This kind of change is mentioned in hereditary cancer syndromes, which are transferred to the child from the parents. Due to the genetic disorder, this kind of bone marrow cancer is called leukemia [2]. The leukemia risk is increased by the factors of the environment and the human being's lifestyle. The stem cell leads to the cancerous cell and slowly multiples it uncontrolled at the primary phase of leukemia. It has no factors which are aware. Cancer cells are not working in a relevant way and use the healthy cells in the bloodstream and the bone marrow. Leukaemia is developed in the human body, is the fundamental factor is the cell mutations in the bone marrow. The healthy cells are developed using the bone marrow; this is thwarted. DNA mutation as deoxyribonucleic acid is detected, which is difficult for different genetic diseases for the earlier diagnosis due to the many genes that have multiple areas from where the mutation occurs [3]. One upgraded methodology is the DNA microarray that measures the expression level for the huge volume of genes. The ability to evaluate if individual DNA has the gene of mutation or not using technology. Various kinds of leukemia are used in microarray prediction and analysis technology [4].

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

The difficult task is the earlier prediction in today's world for physicians and adopting automatic computer oriented for disease diagnosis systems in the clinical phases. Various machine learning (ML) approaches are adopted to model the intelligent diagnosis system for clinical datasets. A large volume of data is created from the clinical sectors because of the upgrade and digital revolution in information technology. The large data is analyzed, and different approaches are used to diagnose diseases called machine learning approaches [5] - [6]. Various data for the diagnosis are used in medical records forms at different hospitals for the successful run of machine learning algorithms. Microarray technology generates a large volume of DNA expression in hospitals. The data classification and the automated analysis are important for decision-making and earlier diagnosis of genetic disease [7]. One of the many famous regions is gene expression data in studies in the analysis related to biomedical data based on machine learning. The data is analyzed for various algorithms of machine learning [8]. The proposed DPDPM approach is employed for leukemia prediction. There are two kinds of leukemia with the assistance of deep neural networks: AML as acute myelocytic and ALL as acute lymphocyte [9] - [10]. The suggested methods have a classification performance that is satisfactory when compared with previous works described in the conclusion section. The major research gap occurs due as the earlier prediction of leukemia is more challenging for experts and it can be achieved effectually by the automated disease diagnosis system. Thus, there is a need of huge gene expression dataset. Since, the provided AML dataset shows less number of labelled samples, there is a need of suitable features for the smaller dataset which potentially learn better feature representation from data for AML prediction. However, layer-level stacking for feature extraction generally provides superior representation of learning models which motivates to design DPDPM to learn the higher level features. Also, it is proven that the multi-modal data works well compared to single model data. Thus, it makes better prediction/classification outcomes.

The proposed system has the below organizations. Section 2 presents the related literature. Suggested work is offered in section 3. Section 4 presents the outcomes that are achieved via the suggested method. The conclusion work is presented in work 5.

2. Related Works

Various approaches use authors' gene expression data analysis because of the latest machine learning and enhancement in AI artificial intelligence. The important aim of the machine learning approaches is to enhance the model's performance using the decrease in error and increase in accuracy. The researchers in [11] suit three various algorithms for supervised machine learning to classify cancer from gene expression data. Because of the latest enhancement in AI artificial intelligence and the authors uses various approaches and ML for the analysis of the data in the gene expression, there are seven kinds of data by researchers. The machine learning approaches aim to enhance the models' performance with increased accuracy and decreased error. There are three supervised machine learning algorithms based on a tree used in [12] by authors from gene expression data to classify cancer. Seven cancer data are classified with the help of boosted and bagged decision trees and the decision tree C4.5. Compared with the other two classifiers, this is observed as the bagging decision tree classifier performs better. The original data has the relevant feature selection, which is difficult work in the technique of machine learning, and this is essential in biomedical data.

The researchers utilized the feature selection methodology based on null space in [13] to enhance the performance. The null space-relevant feature selection method discards the redundant gene expressions in [14] via the scatter matrices. There are three classifiers: Support Vector Machine (SVM), Naïve Bayes (NB) and Linear discriminant analysis (LDA) after successfully reducing feature dimension to perform the classification. The use of one class of SVM classifier in the gene expression data in [9] is to find the samples of AML. The performance in the proposed classifier is checked to have various kernel functions, and this is observed that better accuracy is achieved with the linear kernel. The outcome is compared with a few early conventional classification models, and a better outcome is obtained than another classifier. The higher dimension issue is obtained in [15] by using the approach of chi-square feature selection. The anticipated model adopts ML approaches in the SAGE data. ELM algorithm as Extreme Learning Machine has the performances determined by having the gene expression dataset in [16]. The ELM algorithm is used to enhance the training time and eliminate the problems of over-fitting and local minima. Various kinds of microarray data sets are performed in the multi-category classification. The authors select the gene patterns with the help of a random forest algorithm [17]. The suggested method aims to identify the considerable genes the accuracy is increased. The authors are concerned with various kinds of gene expression to validate the ML model's performance.

The suggested hybrid model achieves an effective outcome when compared with conventional models. The gene patterns suggested in [18] by using another hybrid method. The gene patterns are classified using the rough set theory technique and hierarchical clustering. There are three-step structures used in the suggested technique. Primarily, the hierarchical method is used to form the gene clusters. The primary clusters are classified into many

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

clusters with the help of upper and lower approximation properties for the rough set algorithm. Hence, cluster ranking and gene ranking are used for ranking the clusters to select the considerable genes. The SVM classifier performs the classification after completing the pre-processing phases. The authors use the SSA as a salp swarm algorithm and MOSHO as a multi-objective spotted hyena optimizer [19] to design the proposed system's optimized classification model of gene data. The suggested approach of SSA handles the training information and diversity compared with the real-time optimizing algorithm. Lower time for computation is considered by using the MOSHO approach to maintain the essential data. There are four kinds of classifiers used in the proposed system for classification. Then, the gene expression classifications are performed using the SVM classifier. The authors develop the algorithm of gray wolf optimization in [20] for the classification of gene expression data for the optimized data mining model.

The approaches of information gain are used in the proposed system to select the features. The colon and breast cancer data of genes are utilized for classification. The selection of correlation-oriented features of the gene in [21] is done for the classification. The TLBO algorithm as teaching-learning-based optimization is used in the proposed system to choose the SVM classifier's optimized parameters. The new encoding approach is chosen in the proposed system to convert the contiguous optimized parameter to the binary version. The other famous data mining method is KNN, as the K-nearest neighbor is adopted for the classification of gene expression. The proposed combination of the SVM classifier and KNN is used to classify cancer colon and leukemia data. K-NN and PSO as particle swarm optimization are used in the proposed system [22] by researchers to classify the gene expressions of ALL and AML.

The genetic algorithm obtains the various gene subsets classified using the k-NN. Researchers provide gene expression analysis to deploy the neural network and its variations. The gene expression is classified using the ANN as an artificial neural network. The authors use feature selection approaches and normalization to pre-process the gene data [23]. The particular gene clusters are created by using the K-means clustering approach. The SVM is used in the proposed system to lower the size of the feature. The ANN is used to classify the reduced selective features. There are two kinds of tumor datasets to validate the suggested model. The AML and ALL data of genes are classified by suggesting the technique of supervised machine learning that is suggested. A better outcome is obtained by the classification using the ANN classifier when the comparison is done with other classifiers as it is claimed. The RBFN, as a radial basis function network, classifies the gene sequences. A theoretical approach to feature selection based on a rough set is given in [24] – [25].

The method is chosen with the merit of eliminating the clustering of gene expression. The roughest approach of feature selection and RBFN is used in the proposed system to predict lung cancer, prostate cancer, and leukemia. The outcome in the proposed system is validated by having two other kinds of classifiers from the outcome. It is noticed that RBFN is given better outcomes when compared with naïve Bayes and LSVM. The prediction accuracy is improved in classifying a gene by using the generalized RBFN. The BIRS, FCBF, and BARS approaches are used to choose the effective features. The RBFN and FLDA as Functional logistic discriminant analyses are utilized for the gene expression analysis data presented in [26]. Another kind of neural network is called PNN, a probabilistic neural network utilized for the classification and analysis of gene expression [27]. The classification is performed with the help of PNN after choosing the efficient features of the gene. Compared with KNN, the performance of PNN is better in classifying the three kinds of information. The upgraded approach of the neural network is used in the various applications of biomedical called deep learning. In addition, the authors use various deep-learning techniques in gene expression data analysis. The convolutional neural network (CNN) extracts the essential gene features [28]. SVM is employed to categorize the extracted convolutional features. The image-processing approaches rely on DL and are used to classify all kinds of leukemia [29].

After gathering the microarray image, the segmentation step is performed to enhance the performance. The seven layers of the convolutional network are utilized after completing the segmentation phase for classification. The hybrid technique uses the combination of SVM and CNN to classify the bone marrow expression images by researchers. Leukaemia is classified using the deep convolutional neural network [30]. The pre-trained AlexNet model is used in the proposed system to train the model successfully. CNN is used to classify the four kinds of leukemia. Researchers in another research use the hybrid deep learning model to classify leukemia cells. The suggested hybrid technique can extract effective features from the image of input. The global average pooling technique is used in the proposed system to enhance performance. CNN is used to classify blood cell images. Researchers use the similar seven-layer structure of CNN to classify leukemia. The authors develop the model of the CNN classifier based on DCT as a discrete cosine transform. The DCT is used in microscopic images to extract the features, and the performance in classification is obtained with the help of CNN. However, the issues in the microscopy inspection of blood films are considered to be crucial for predicting the acute leukaemia predication. The model consumes huge resource settings which is a time consuming process and inconsistency. The throughput of the model is also reduced due to the lack of resource sufficiency and increases the cognitive human error [31]

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

– [33]. Much work is presented from the study to classify the gene expression data. Various algorithms of machine learning are utilized to obtain satisfactory accuracy. The authors utilized the technique based on deep learning in gene expression data to enhance accuracy and reduce the time in computation. Many works based on deep learning are performed having the images of microarray. The attempts are taken in the proposed system to classify previous records of gene features extracted and obtain a better outcome with the information when the comparison is done with another classification model.

3. Methodology

The suggested classification model of leukemia based on deep learning has three stages in Fig 1. The pre-processing is taken place after the collection of data. Due to the hampering of the classification outcome, the complete dataset is verified in the pre-processing step to detect the missing value. The value obtained in the proposed dataset is present. The complete dataset is classified into two subsets after verifying the missing value. The proposed DPDPM is used for prediction.

3.1 Data acquisition and pre-processing

The public data on microarray genes is available to gather leukemia datasets. The dataset has 72 samples of bone marrow expression having the genes of 7128. The dataset has two kinds of leukemia classes. ALL class has forty-seven samples, and the AML class has 25 having zero missing values. After verifying the missing values, the complete data set is classified for testing and training. The training purposes have the separated data with nearly 58 samples, 80%, and the remaining 18 samples, 20%, are utilized for validation and testing purposes. Similarly, the pre-processing is also done for successive datasets like Kaggle and Bio GPS.65. Some pre-processing steps like augmentation are employed with sample rotation, changing the orientation and angle. Also, the noise over the input samples is eliminated to enhance the model performance and quality. This step is substantially essential for all the input samples to make better prediction. After pre-processing, the sample features are analyzed using the proposed dense polynomial dimensionality based prediction model. The features are extracted by the intermediate layers to perform better classification.

3.1.1 Kaggle dataset

The cells are segmented from microscopic images and the real-time image representation as the images contain illumination errors and staining noise. These errors are fixed in course acquisition. The task intends to predict the leukemic cell. There are 15, 135 images gathered from 118 patients.

3.1.2 BIO GPS

It is a collection of dataset where thousands of samples are available which are related to chronic myeloid, myeloid leukemia, genome, cell, bone marrow, cancer, class and so on. These are tags related to the dataset and the species are included.

3.1.3 Gene expression

It includes the proof-of-concept which includes the new cancer cases which is classified by gene expression and provides general approach for predicting the cancer classes and allocating tumours to known classes. The data is classified to classify patients with acute myeloid leukemia and acute lymphoblastic leukemia.

3.2 Dense Polynomial Dimensionality based Predictor models network Units

In this work, the proposed DPDPM algorithm is anticipated and applied for multi-modal gene expression analysis and diagnosis. The proposed model can effectually fuse and learning feature representation from multi-modal data. The model is developed as a supervised algorithm to evaluate the quadratic and linear function and the learned predictors are determined as the polynomial functions. The architecture of the deep network helps the proposed algorithm for learning the polynomial predictors, which gives a better bias of approximation using the polynomials for the obtained values across the training samples. Some existing investigator needs to be referred for further information, and a brief introduction is to be given in the proposed system regarding the algorithm of DPDPM. The DPDPM architecture with four network layers is shown in Fig 1 as the instance. The input is given to the DPDM where the samples $(n^1, n^2, n^3 \ldots, n^t)$ are considered and the outputs are extracted from every individual block and finally the outcomes are accumulated to give the final outcome. Once the outcomes are accumulated, the features are provided to the classifier where the final prediction outcomes are extracted.

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

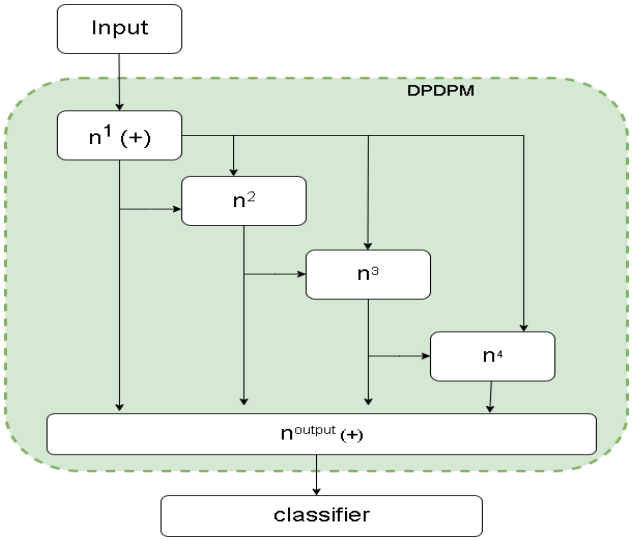

Figure 1: DPDPM

The m training samples are $\{(x_i, y_i)\}_{i=1}^{m}$, the label's relevant value is y_i, and the d −dimension sample is x_i. The definition of multivariate polynomials on R^d is provided below:

$$p(x) = \sum_{i=0}^{\Delta} \sum_{\alpha^i} W_{\alpha^i} \prod_{l=1}^{d} x_l^{\alpha_l^i} \quad (1)$$

Here, the degree of the polynomial is represented as the degree index is represented as i, the d −dimensional index with $\sum_{l=1}^{d} \alpha_l^i = i$ is presented as α, and the index weight of α^i is presented as W_{α^i}. Then, $n_j^i(.)$ is used in the proposed system in DPDPM for describing the j^{th} node in the layer of i^{th} inputs functions. The function computed for simplifying every node is either linear or a product of two weighted inputs. The linear polynomial as degree-1 polynomials function has the set of values across the training samples when establishing the primary network layer in DPDPM that is provided below.

$$\{(< w, [1, x_1] >, \ldots., < w, [1, x_m] >): w \in R^{d+1} \quad (2)$$

Here the dimensional linear subspace of R_m is presented as $(d + 1)$, Where the inner product is presented as $\langle \cdot, \cdot \rangle$. The singular value decomposition performs the basis for the linear dependent set. The W matrix is presented in the proposed system to simplify the model that maps the [1 X] to the basis of construction. Here, all one's vectors are presented as 1, and the samples' matrix is presented as X. The first layer is formed in DPDPM using the columns of W, which means the $d + 1$ linear functions. The j^{th} node of the primary layer is the function for all $j = 1, \ldots, d + 1$.

$$n_j^1(x) = < W_j[1\ X] > \quad (3)$$

The fundamentals for all the obtained values are presented as $\left\{\left(n_j^1(x_1), \ldots, n_j^1(x_m)\right)\right\}_{j=1}^{d+1}$ over the training samples by the linear polynomials as degree-1 polynomials, the $m \times (d + 1)$ matrix having $F_{i,j}^1 = n_j^1(x_i)$ be F^1. Implementing the one-layer network is done with the pan of all the output values achieved on the training samples using the linear functions. Any degree-2 polynomial is defined below using the polynomials decomposition theorem.

$$\begin{gathered} \sum_i \propto_i^{(g_i)} n_j^1(x) \left(\sum_j \propto_i^{(g_i)} n_j^1(x)\right) + \left(\sum_j \propto_i^{(k)} n_j^1(x)\right) = \\ \sum_{i,j} n_j^1(x) n_s^1(x) \left(\sum_j \propto_i^{(g_i)} \propto_s^{(h_i)}\right) + \left(\sum_j n_j^1(x) \left(\propto_i^{(k)}\right)\right) \end{gathered} \quad (4)$$

Here, a scalar is represented as α, g_i represents the superscripts, and the degree 1 polynomial having $g_i(x) = \sum_j \propto_i^{(g_i)} n_j^1(x)$ and $h_i(x) = \sum_j \propto_s^{(h_i)} n_s^1(x)$ is represented as h_i accordingly. The degree-2 polynomial is used to achieve the vector of values that spans the vector of values using the nodes present in the primary layer. Every two

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

nodes in the primary layer have the products of the outputs. Further, $\tilde{F}^2$ is defined in the proposed system as below.

$$\tilde{F}^2 = [(F_1^l \circ F_1^l) \dots (F_1^l \circ F_{|F_1|}^l) \dots (F_{|F_1|}^l \circ F_{|F_l|}^l)] \quad (5)$$

Here, the Hadamard product is presented as $\circ$, the number of columns of matrix F_1 is denoted as $|F_1|$, and the i^{th} column vector of matrix F is presented as F_i. Thus, $[F\ \tilde{F}^2]$ is the concatenation of a new matrix that occurs in the span of all the essential obtained values using the degree-2 polynomials. The basis is constructed by performing the single value decomposition again. The subset of the columns of $\tilde{F}^2$ is presented as F^2, the basis for the degree-2 polynomial is generated as F^2, and the decomposition method is used to choose the linear independent columns from $\tilde{F}^2$. The F^2 columns are specified as the second layer of the network. Hence, every column is presented the $F_i^1 \circ F_j^1$ related to the node in the second layer, which computes the product of the node as $n_i^1(\cdot)$ and $n_j^1(\cdot)$ in the primary layer. Here, F is defined again as the $[F\ F^t]$ as an augmented matrix for the simpler notation. Thus, the matrix F is maintained at every iteration t with the columns to form the basis for all the obtained values using the polynomials of degree $\leq (t-1)$. The representation of the new matrix is defined below.

$$\tilde{F}^t = [(F_1^{t-1} \circ F_1^l) \dots (F_1^{t-1} \circ F_{|F_1|}^l) \dots (F_{|F_1|}^{t-1} \circ F_{|F_l|}^l)] \quad (6)$$

Here, $[F\ F^t]$ columns create the basis for the $[F\ \tilde{F}^t]$ columns. A network is implemented using the addition of the newly implemented layer with the outputs to create the basis for the achieved values over the training samples using all polynomials of degree $\leq t$. Then, $\tilde{F}^t$ is converted to F^t through the matrix W of size $|F^{t-1}| \times |F^1|$ for maintaining the stability of numbers that are presented below:

$$F_{as}^t : W_{i(s),j(S)} F_{i(s)}^{t-1} \circ F_{j(j)}^1 \text{ where } s = 1,2, \dots., |F^t| \quad (7)$$

Here, i is a function of s. It is represented as $i(s)$, and the matrix projection is presented as W that maps to the basis F^t of degree t polynomials from $\tilde{F}^t$ obtained using the procedure of Gram-Schmidt or more alternate methodologies in the stable form. The process is similar to constructing W in the primary layer. After Δ the -1 iteration, a matrix F is finished with the columns to create the basis for all values across the training samples using the polynomials of degree $\leq (\Delta - 1)$. Fig 2 represents the mentioned processes needed to implement the network in DPDPM. It is identified that DPDPM includes the architecture of feed-forward that creates DPDPM, which is very simple. Here, two DPDPM modules are considered where the input from the datasets and the corresponding samples are provided for simultaneous process. The training and testing can be done in these DPDPM blocks where the accumulated outcomes are provided for further classification. The proposed blocks are made of independent layers where the errors are extracted and the errors are eliminated to make better classification.

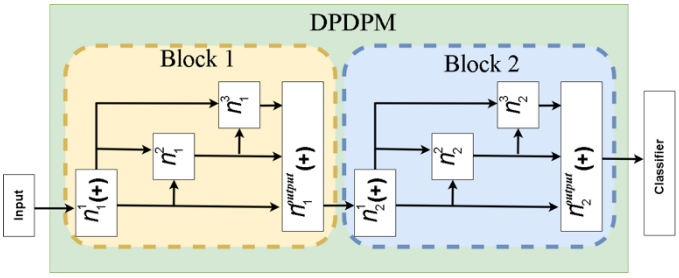

Figure 2: DPDPM Block

Algorithm 1:

1. Set empty matrix with F;
2. Evaluate the single decomposition for predicting W;
3. Generate the initial layer:
4. Construct the linearly independent layers;
5. if the error is small, then
6. Compute $\tilde{F}^t$ and construct bias value;

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

7. Design the linearly independent layers;

8. $F = [F\ F^t]$;

//output layer design;

9. Evaluate w;

10. Evaluate $n^{output}(.)$;

11. Evaluate the error rate;

12. Return $(n^{output}(.), error)$

The feature output layer $n^{output}(+)$ gives the learned features on the network's top. The standard L2-regularized hinge loss depends on the simpler linear classifier, like the output of the final decision resolved on the hinge loss function using the stochastic gradient descent optimization. In addition, other classifiers are also utilized for DPDPM. The mentioned methodology constructs DPDPM. Moreover, the current algorithm is utilized with restricted nodes for small datasets. The width of the network, that is, the number of created nodes in every layer, is also very huge, which results in a huge network having higher complexity in computation when using the DPDPM to the larger training samples. The altered solution is explicitly suggested for the network constraint width in every iteration to mention this problem that utilizes the small partial basis for creating the smaller nodes in the layer. In particular, the proposed system gives up on the spanning of $\tilde{F}$ in DPDPM yet asks for the "approximately span" with the help of a small partial basis for the size bounded r, which gives in the width layer r. Hence, the generation of the network is done by having sparse connections having most of the nodes based on the two nodes' inputs, which creates fast computation in the network. Hence, the new solution is merged having the sparse node's connection which creates the DPDPM works well for the data of large scale with no loss in the merits of efficiency and less complexity in the computation for the smaller data.

The primary network layer calculates the linear transformation, which converts to top k leading singular vectors from augmented data matrix [1 X] for the supervised DPDPM using the PCA as principal component analysis. The next networks layer chooses the standard orthogonal least squares process to pick the $\tilde{F}^t$ columns greedily for the most related to predict. The least-square has the supervised technique that iterates the pick up of $\widetilde{the\ F^t}$ column in the proposed system with the residual after the projection of the previous basis F that is more correlated having the prior residual labels.

3.3 Dense stacking network model (DSNM)

The one RBM has the output to enhance DBN's representation as input to other RBM is discussed by Hinton et al. Posterior RBM provides a more complicated input presentation. The DSNM algorithm as DPDPM is suggested in the proposed system using the inheritance of the stacked technique to handle the problem utilized in deep learning. Different fundamental DPDPM is stacked on one another in DSNM to create the deep hierarchy, which is the output of fundamental DPDPM wired to the input for the next fundamental DPDPM at the consecutive stage. The two-level DSNM approach is the example in Figure 3 extended to the m level. The original input of the feature vector to the primary stage fundamental block of DPN having the n layers of networks for creating the output of learned feature as $n_1^{output}(+)$ utilized as the input to the successive stage of fundamental DPDPM for the m level DSNM. The $n_1^{output}(+)$ output is given to the next level of fundamental DPDPM to train once the present fundamental DPDPM completes the training. Every fundamental block of DPDPM has a simpler and easier for learning the module, which is stacked to create the complete deep networks. Every fundamental DPN in the block-wise fashion is trained with no requirement for the backpropagation that varies from another famous architecture of deep learning. Hence, DSNM is much simpler, having less computation complexity than other deep learning algorithms with the backpropagation technique. It is essential to learn the primary layer of a network of every fundamental DPDPM is computed by the linear transformation, which converts to top k leading singular vectors from augmented data matrix [1 X] using the PCA.

3.4 Modality Analysis

An easy solution is concatenating all the vectors of various modalities to combine and learn the representation of features using DSNM from the multi-modal data. Moreover, the easy concatenation technique eliminates the different modalities' diversity to a particular extent, which needs to be explored better. The complementary nature and the high non-linear correlations are presented between many modalities. Hence, the modality algorithm is used in the proposed system, which depends on the two-phase DSNM presented in Fig 3. Each data of AML is given to

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

the relevant module of DSNM in the primary phase for learning the higher-level representation of the feature. Every particular modality has a higher level of features that reflects its attribute with no correlation data between various modalities. Complete learned features are given to the new module of DSNM in the second phase for the association of complete modalities. The last learned greater level of features has both the properties of intrinsic of every modality and the correlations between the modalities as the outcome. Hence, DSNM learns the features which are robust and more discriminative.

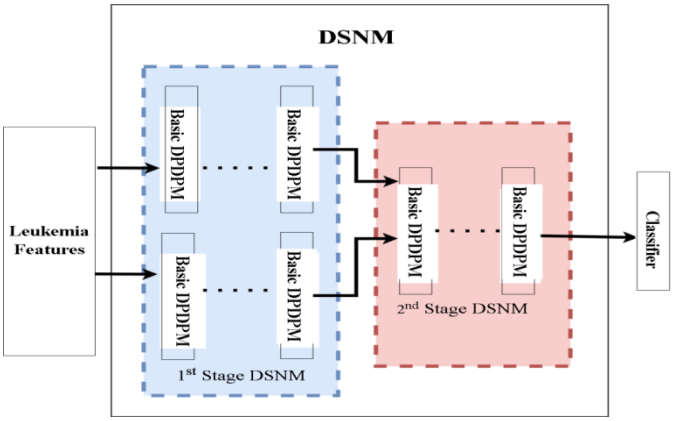

Figure 3: DSNM Block Diagram

The proposed system chooses the inputs as the multi-modality generally relies on AML prediction. The fusion technique differs in modality from the multi-modal encoder-related algorithm for identifying the synergy. It is important to note that the pre-training approach is chosen with the proportionate error inputs that include one presented modality alone. In particular, the inputs are set for the random hide of one modality for the inputs as 0 and provide the remaining training samples having both modalities. The first level encoder has the hidden layer, trained to reconstruct all the original inputs with hidden modalities from the mixed input. The corrupted inputs and the original inputs are transferred to the greater layers of the network to obtain a clear presentation independently. Since DPDPM is performed the feed-forward supervised learning in every layer of the network with no fine-tuning during the proposed algorithm, this is complex for performing a similar learning technique for inferring the correlations. Hence, the secondary level of DSNM is trained to concatenate features at the primary level to learn the shared representation. It is the same as the simpler method of fusion.

4. Result and Analysis

This section provides the numerical outcomes of the anticipated DPDPM model, and the simulation is done with MATLAB 2020a. Some metrics include accuracy, precision, recall, F1-score, AUROC and time (min). The representation of the input data set is done with the $x - by - y$ matrix. Here the total number of samples is presented. The proposed model uses the separated data for the training of nearly 56 samples as 80% of data among 72 samples, and the remaining 16 samples as 20% is used for validation. The training data is used to train the DPDPM network. The validation is done with the testing data after completing training. The loss and accuracy of the suggested model having the testing data are shown. The suggested outcome is presented in Fig 4, and the proposed outcome is compared with the other three kinds of classifiers. Fig 4 to Fig 6 and Table 1 present the outcome. Compared with hybrid and standard CNN, the proposed model performance is better. The suggested model has observed from the outcome has superior performance when evaluated with the other three classifiers. The achieved outcome is compared with a few previous works in Table 1, and the observation is done with the proposed deep learning model better predicting AML.

Table 1: Overall Comparison of training samples

Dataset	Approaches	Accuracy (%)	Precision (%)	Recall (%)	F1-score	Time (min)	AUROC
Kaggle dataset for	Hybrid CNN	92.5	75	72	88	0.10	93.8
	Standard CNN	87.8	78	70	68	0.10	93.36

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

Leukemia classification	CNN with boosting model	88.9	72	72	71	0.78	94.86
	D-CNN with data augmentation	86.7	75	73	78	6.80	92.40
	CNN with Genetic optimization	87.8	86	90	93	1.100	93.32
	Standard CNN	93.1	91	93	91	0.06	96.89
	R-CNN	95.8	93	95	95	0.04	97.78
	DPDPM	96	94	96	96	0.03	98
Gene Expression dataset	Hybrid CNN	90.1	74	70	89	0.05	92.50
	Standard CNN	91.1	76	69	69	0.08	92.14
	CNN with boosting model	90.4	77	70	72	0.78	91.14
	D-CNN with data augmentation	91.9	75	68	79	6.06	90.38
	CNN with Genetic optimization	90.15	71	73	92	1.62	90.97
	Standard CNN	93.25	72	93	90	0.04	93.56
	R-CNN	95.50	93	94	95.8	0.04	94.58
	DPDPM	97	94	95	96	0.03	96
Bio GPS Dataset	Hybrid CNN	95	80	78	89.25	1.25	87
	Standard CNN	96	62	77	67	2.658	86.5
	CNN with boosting model	93	84	78	72	4.56	85
	D-CNN with data augmentation	94	82	77	77.9	3.56	82
	CNN with Genetic optimization	94.50	92	96	94.24	4.57	80
	Standard CNN	90.10	90	88	90	2.89	90
	R-CNN	97.30	94	98	95.65	0.04	92

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

	DPDPM	98	94.5	98.5	96	0.03	94

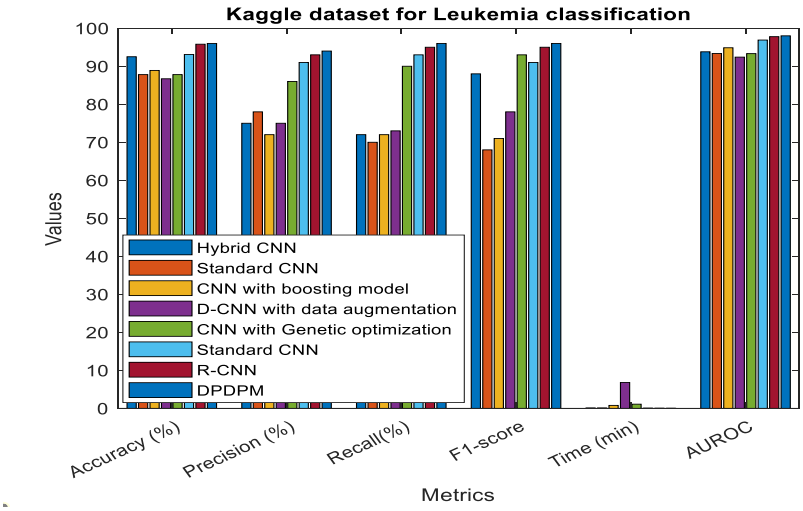

Figure 4: Performance Evaluation With Kaggle Dataset

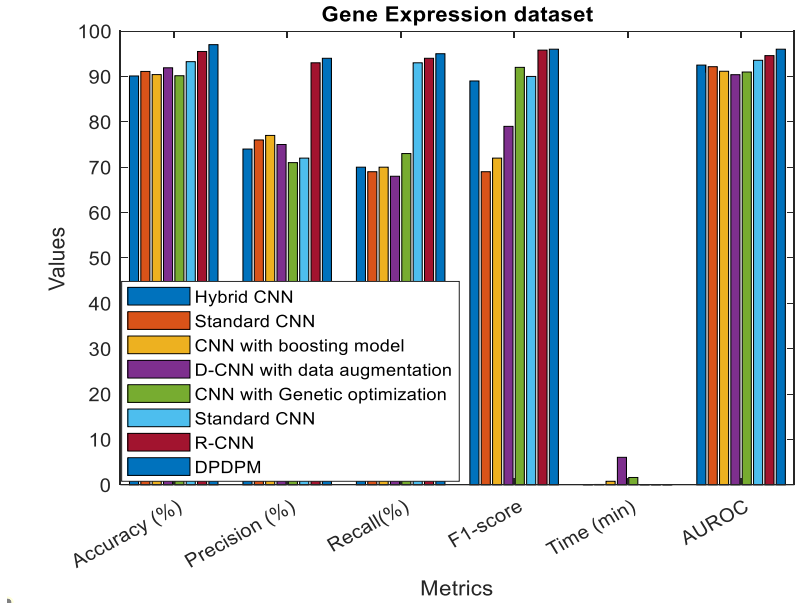

Figure 5: Performance evaluations with Gene Expression Dataset

Table 2: Overall Comparison of testing samples

Dataset	Approaches	Accuracy (%)	Precision (%)	Recall (%)	F1-score	Time (min)	AUROC
Kaggle dataset for Leukemia classification	Hybrid CNN	92	75.1	72.5	88.2	0.10	93
	Standard CNN	87	78.5	70.5	68.5	0.10	93
	CNN with boosting model	88	72.5	72.8	71.8	0.78	94.6

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

	D-CNN with data augmentation	86	75.2	73.5	78.2	6.80	92.6
	CNN with Genetic optimization	87	86.1	90.5	93.6	1.100	93.6
	Standard CNN	93	91.2	93.5	91.2	0.06	96.5
	R-CNN	95	93.8	94	94	0.04	97.2
	DPDPM	96.2	94.5	96.5	96.5	0.03	98.5
Gene Expression dataset	Hybrid CNN	90	75	72	90	0.05	92
	Standard CNN	92	75	70	70	0.08	92
	CNN with boosting model	9	77.5	70.5	72.5	0.78	91
	D-CNN with data augmentation	91	75.4	68.5	79.5	6.06	90
	CNN with Genetic optimization	90	71.5	73.2	92.1	1.62	90
	Standard CNN	93	72.3	93.2	90.2	0.04	93
	R-CNN	95	93	94.1	95	0.04	94
	DPDPM	97.5	94.6	95.8	96.5	0.03	96.5
Bio GPS Dataset	Hybrid CNN	95.5	80.5	78.5	89.30	1.25	87.5
	Standard CNN	96.5	62.8	78	68	2.658	87
	CNN with boosting model	94	85	79	73	4.56	86
	D-CNN with data augmentation	95	83	78	78	3.56	83
	CNN with Genetic optimization	94	93	97	94	4.57	81
	Standard CNN	90	91	89	92	2.89	91
	R-CNN	97.2	95.2	99	95.65	0.04	92
	DPDPM	98	94.5	98.5	96	0.03	94

Table 1 depicts the overall comparison of various existing approaches with the proposed model. While analyzing the Kaggle dataset, the proposed model gives 96% accuracy, 94% precision, 96% recall, 96% F1-score, and 98% AUROC. The accuracy of the anticipated model is 96% which is 3.5%, 8.2%, 7.1%, 9.3%, 8.2%, 2.9% and 0.2%, superior to other approaches. The precision of the anticipated model is 94% which is 19%, 16%, 22%, 19%, 8%,

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

3% and 1%, superior to other approaches. The recall is 96% which is 24%, 26%, 24%, 23%, 6%, 3% and 1%, superior to other approaches. The F1-score is 96% which is 8%, 28%, 25%, 16%, 3%, 6% and 1% higher than other approaches. The AUROC is 98% which is 4.2%, 4.64%, 5.6%, 4.68%, 1.11% and 0.22%, higher than other approaches. Also, Table 2 shows the testing data results.

5. Conclusion

The proposed DPDPM model attempts to perform the classification of leukaemia type with the help of a deep learning technique. The deep network layer is developed to classify two kinds of leukaemia. The proposed model gives better classification accuracy with less computational complexity compared to other classifiers. Based on the experimental analysis, it is better than hybrid and standard CNN, CNN with boosting, D-CNN, CNN with GA, and R-CNN classifiers. The proposed DPDPM model is tested using three datasets: Kaggle, Gene expression and Bio GPS. The model gives 96% accuracy, 94% precision, 96% recall, 96% F1-score, and 98% AUROC while executing with Kaggle; then, 95.50% accuracy, 94% precision, 95% recall, 96% F1-score, and 96% AUROC is achieved while executing with Gene expression and finally 98% accuracy, 94.5% precision, 98.5% recall, 96% F1-score, and 94% AUROC is achieved while executing with Bio GPS. The major research constraint is the computational complexity while executing all three datasets. Due to the complex analysis with the three diverse datasets, the model shows some computational complexity and consumes huge processing time. These limitations need to be addressed in the future. The technique is essential for automatic microarray data analysis in the machine learning tool. The network structure is optimized to enhance accuracy. In the future, the research on the gene recognition mutation will support the virology and the genetic authors as the present pneumonia situation is diagnosed and detected earlier.

References

[1] R. J. Leary, M. Sausen, I. Kinde, N. Papadopoulos, J. D. Carpten, D. Craig, D. O'Shaughnessy, K. W. Kinzler, G. Parmigiani, B. Vogelstein, et al., "Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing," Sci. Transl. Med., vol. 4, no. 162, pp. 2012.

[2] H. B. Hsieh, D. Marrinucci, K. Bethel, D. N. Curry, M. Humphrey, R. T. Krivacic, J. Kroener, L. Kroener, R. Ladanyi, N. Lazarus, N., et al., "High-speed detection of circulating tumor cells," Bio-sensors Bioelectron, vol. 21, no. 10, pp. 1893–1899, 2006.

[3] M. Fatma, J. Sharma, "Identification and classification of acute leukemia using neural network," In 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), pp. 142–145, 2014.

[4] L. Pan, G. Liu, F. Lin, S. Zhong, H. Xia, X. Sun, H. Liang, "Machine learning applications for predicting relapse in childhood acute lymphoblastic leukemia," Sci. Reports, vol. 7, no. 1, pp. 1–9, 2017.

[5] V. Kumar, S. Ailawadhi, L. Bojanini, A. Mehta, S. Biswas, T. Sher, V. Roy, P. Vishnu, J. Marin-Acevedo, V. R. Alegria, "Trends in the risk of second primary malignancies among survivors of chronic lymphocytic leukemia," Blood Cancer J. vol. 9, no. 10, pp. 1–10, 2019.

[6] Y. N. Chung, H. N. Kim, S. R. Lee, H. J. Sung, M. H. Nam. "Usefulness of chromosomal microarray in hematologic malignancies: a case of aggressive NK-cell leukemia with 1Q abnormality," Lab Med., vol. 9, no. 3, pp.189–93, 2019.

[7] Y. Ochi, K. Yoshida, Y. J. Huang, M. C. Kuo, K. Sasaki, N. Hosoya, N. Hiramoto, R. Bera, Y. Nannya, Y. Shiozawa, "Prognostic relevance of genetic abnormalities in the blastic transformation of chronic myeloid leukemia," Blood, vol. 136, pp. 3–4, 2020.

[8] M. Houshmand, G. Simonetti, P. Circosta, V. Gaidano, A. Cignetti, G. Martinelli, G. Saglio, R. P. Gale, "Chronic myeloid leukemia stem cells. Leukemia," vol. 33, no. 7, pp. 1543–56, 2019.

[9] S. Kollmann, E. Grundschober, B. Maurer, W. Warsch, R. Grausenburger, L. Edlinger, J. Huuhtanen, S. Lagger, L. Hennighausen, P. Valent, "Twins with different personalities: STAT5B—but not STAT5A—has a key role in BCR/ ABL-induced leukemia," leukemia, vol. 33, no. 7, pp. 1583–97, 2019.

[10] O. Taiwo, F. Kasali, I. Akinyemi, S. Kuyoro, O. Awodele, D. Ogbaro, T. Olaniyan, "Stratification of chronic myeloid leukemia cancer dataset into risk groups using four machine learning algorithms with minimal loss function," Afr J Manag Inf Syst, vol. 1, pp. 1–18, 2019.

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

[11] L. Yu, X. Huang, R. P. Gale, H. Wang, Q. Jiang "Variables associated with patient-reported symptoms in persons with chronic phase chronic myeloid leukemia receiving tyrosine kinase inhibitor therapy," medicine, vol. 98, no. 48, pp. e18079, 2021.

[12] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgemann, V. H. van Berkel, H. B. Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," Int J Med Inform, vol. 108, pp. 1–8, 2017.

[13] A. Mosquera Orgueira, A. Peleteiro Raíndo, M. Cid López, J. A. Díaz Arias, M. S. González Pérez, B. Antelo Rodríguez, N. Alonso Vence, L. Bao Pérez, R. Ferreiro Ferro, M. Albors Ferreiro, "Personalized survival prediction of patients with acute myeloblastic leukemia using gene expression profiling," Front Oncol, vol. 11, pp. 1018, 2021.

[14] K. Sasaki, E. J. Jabbour, F. Ravandi, M. Konopleva, G. Borthakur, W. G. Wierda, N. Daver, K. Takahashi, K. Naqvi, C. DiNardo, "The LEukemia Artificial Intelligence Program (LEAP) in chronic myeloid leukemia in chronic phase: a model to improve patient outcomes," Am J Hematol, vol. 96, no. 2, pp. 241–50, 2021.

[15] O. Koteluk, A. Wartecki, S. Mazurek, I. Kołodziejczak, A. Mackiewicz, "How do machines learn? Artificial intelligence as a new era in medicine," J Personal Med, vol. 11, no.1, pp. 32, 2021.

[16] A. G. Singal, A. Mukherjee, B. J. Elmunzer, P. D. Higgins, A. S. Lok, J. Zhu, J. A. Marrero, A. K. Waljee, "Machine learning algorithms outperform conventional regression models in predicting the development of hepatocellular carcinoma," Am J Gastroenterol, vol. 108, no. 11, pp. 1723, 2013.

[17] Y. Feng, X. Wang, J. Zhang, "A heterogeneous ensemble learning method for neuroblastoma survival prediction," IEEE J Biomed Health Inform, vol. 26, pp. 1472–83, 2021.

[18] A. Jamshidi, J. P. Pelletier, J. Martel-Pelletier, "Machine-learning-based patient-specific prediction models for knee osteoarthritis," Nat Rev Rheumatol, vol. 15, no. 1, pp. 49–60, 2019.

[19] R. Jayashanka, C. Wijesinghe, A. Weerasinghe, D. Pieris, "Machine learning approach to predict the survival time of childhood acute lymphoblastic leukemia patients," In Proc. 18th international conference on Advances in ICT for emerging regions (ICTer), pp. 426–432, 2018.

[20] J. N. Eckardt, C. Rollig, M. Kramer, S. Stasik, J. A. Georgi, P. Heisig, F. P. Kroschinsky, J. Scheteli, U. Platzbecker, C. Müller-Tidow, "Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning," Blood, vol. 138, pp. 108, 2021.

[21] K. Karami, M. Akbari, M. T. Moradi, B. Soleymani, H. Fallahi, "Survival prognostic factors in patients with acute myeloid leukemia using machine learning techniques," PLoS ONE, vol. 16, no. 7, pp. e0254976, 2021.

[22] K. N. Neeraj, V. Maurya, "A review on machine learning (feature selection, classification and clustering) approaches of big data mining in a different area of research," J Crit Rev, vol. 7, no. 19, pp. 2610–26, 2020.

[23] A. M. Alqudah, "Ovarian cancer classification using serum proteomic profiling and wavelet features a comparison of machine learning and features selection algorithms," J Clin Eng, vol. 44, no. 4, pp. 165–73, 2019.

[24] X. Gu, J. Guo, L. Xiao, T. Ming, C. Li, "A feature selection algorithm based on equal interval division and minimal-redundancy–maximal-relevance," Neural Process Lett, vol. 51, no. 2, pp. 1237–63, 2020.

[25] D. Chen, G. Goyal, R. Go, S. Parikh, C. Ngufor, "Predicting time to first treatment in chronic lymphocytic leukemia using machine learning survival and classification methods," In Proc. IEEE international conference on healthcare informatics (ICHI), pp. 407–408, 2018.

[26] A. Kashefzadeh, L. Ohadi, M. Golmohammadi, F. Araghi, S. Dadkhahfar, A. Kiani, A. Abedini, A. Fadaii, A. Ghojoghi, M. Nouraie, et al., "Clinical features and short-term outcomes covid-19 in Tehran, Iran: an analysis of mortality and hospital stay," Acta Biomed, vol. 91, no. 4, pp. 1–10, 2020.

[27] X. Hu, B. Wang, Q. Chen, A. Huang, W. Fu, L. Liu, Y. Zhang, G. Tang, H. Cheng, X. Ni, "A clinical prediction model identifies a subgroup with inferior survival within intermediate-risk acute myeloid leukemia," J Cancer, vol. 12, no. 16, pp. 4912–23, 2021.

[28] S. Rinesh, K. Maheshwari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, Y. A. Waji, "Investigations on brain tumor classification using hybrid machine learning algorithms", Hindawi, Journal of Healthcare Engineering, 2022.

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023

[29] S. Dasariraju, M. Huo, S. McCalla, "Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random Forest algorithm," Bioengineering (Basel), vol. 7, pp. 120, 2020.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun ACM, vol. 60, pp. 84–90, 2017.

[31] Nada M. Sallam, "An efficient EGWO algorithm as feature selection for B-ALL diagnoses and its subtypes classification using peripheral blood smear images" Alexandria Engineering Journal, Vol. 68, pp. 39-66, 2023.

[32] Fallah H Najjar1, Kifah T Khudhair2, Zaid Nidhal Khudhair3,4, Haneen H Alwan2 and Ameer Al-khaykan, "Acute lymphoblastic leukemia image segmentation based on modified HSV model", J. Phys.: Conf. Ser. 2432 012020.

[33] Petru Manescu, Priya Narayanan, Christopher Bendkowski, Muna Elmi, Remy Claveau, Vijay Pawar, Biobele J Brown, Mike Shaw, Anupama Rao, Delmiro Fernandez-Reyes, "Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning", Sci Rep, Vol.3. 13(1), pp.2562-70, 2019.

[34] S. Hemamalini ,V. D. Ambeth Kumar ,R. Venkatesan,S. Malathi. (2023). Relevance Mapping based CNN model with OSR-FCA Technique for Multi-label DR Classification. Journal of Fusion: Practice and Applications, 11 (2), 90-110.

[35] C. S. Manigandaa,V. D. Ambeth Kumar,G. Ragunath,R. Venkatesan,N. Senthil Kumar. (2023). De-Noising and Segmentation of Medical Images using Neutrophilic Sets. Journal of Fusion: Practice and Applications, 11 (2), 111-123

[36] Sathya Preiya, V., and V. D. Ambeth Kumar. 2023. "Deep Learning-Based Classification and Feature Extraction for Predicting Pathogenesis of Foot Ulcers in Patients with Diabetes" Diagnostics 13, no. 12: 1983. https://doi.org/10.3390/diagnostics13121983

[37] Balakrishnan, Chitra, and V. D. Ambeth Kumar. 2023. "IoT-Enabled Classification of Echocardiogram Images for Cardiovascular Disease Risk Prediction with Pre-Trained Recurrent Convolutional Neural Networks" Diagnostics 13, no. 4: 775. https://doi.org/10.3390/diagnostics13040775.

[38] V. D. Ambeth Kumar,S. Malathi,Abhishek Kumar,Prakash M and Kalyana C. Veluvolu, "Active Volume Control in Smart Phones Based on User Activity and Ambient Noise" ,Sensors 2020, 20(15), 4117; https://doi.org/10.3390/s20154117

[39] A. R. Alzahrani, A. Alzahrani, and M. A. Alzahrani, "Blockchain-Based E-Voting System Using Face Recognition Technology," in "Journal of Computer Networks and Communications", vol. 2021, Article ID 123456, 2021.

Doi: https://doi.org/10.54216/FPA.120212
Received: January 27, 2023 Revised: April 24, 2023 Accepted: June 18, 2023