IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 12, December 2024| DOI: 10.15680/IJIRCCE.2024.1212057

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Identification of Leukemia using Machine Learning Approaches

Jonnakuti Rakesh Babu, Bakka Rajeev Gandhi

Assistant Professor, Department of Computer Science & Engineering, Chalapathi Institute of Technology, Guntur, A.P, India[1]

Assistant Professor, Department of Computer Science & Engineering, Chalapathi Institute of Technology, Guntur, A.P, India[2]

ABSTRACT: Leukemia is a type of blood cancer which occurs due to abnormal increase in WBCs (white blood cells) in bone marrow of human body. Leukemia can be classified as acute leukemia and chronic leukemia, in which acute leukemia grows very fast whereas chronic leukemia grows slowly. Further both the types have two sub categories lymphocytic and myeloid. In this paper, we are going to analyze different image processing and machine learning technique CNN model used for classification of leukemia detection

KEYWORDS: Keywords:Leukemia , Deep Learning Approaches,Segmentation, Classification,ANN (Artificial Neural Network),SVM (Support Vector Machine),,Deep Learning Techniques

I. INTRODUCTION

Many age groups have common cancers, especially children. Excessive blood cell proliferation and immature growth can damage red blood cells, bone marrow.Image processing techniques are most widely used for detection of various medical diseases. Leukemia is one of the most interesting areas for researchers because it belongs to the category of blood cancer which can affect the persons of all ages starting from children to the old age people. The use of image processing with Computer-Based algorithm makes possible the classification of very easy. Detection of Cancer cell is when done by some expertise of the field then there may be some error present due to lack of knowledge or incorrect information present in the microscopic image. So, Computer-Based algorithm can be very beneficial in such a field to increase the detection accuracy [1]. There are two types of WBCs present in human blood and when the affected cells are monocytes type and granulocytes type, then the leukemia will be classified as myelogenous

1) Acute Myeloid Leukemia (AML): This type of cancer occurs due to under development or some bad effect on bone marrow.When the WBCs rise rapidly, the working of bone marrow effected badly and causes cancer. In most of the cases early detection of such cancer may lead to successful treatment. During this type of cancer, a person may feel problem in breathing, bleeding etc. [1, 3].

2) Acute Lymphocytic Leukemia (ALL): This type of cancer generally founds in kids and the major reason of such cancer is the rapid growth in white blood cells. A number of the factors that are joined to the present sort of cancer include: radiation exposure, viral infections and transmitted diseases like Down's syndrome. Kids have a better rate of remission than older adults who are diagnosed with this kind of leukemia. ALL is further classified as L1, L2 and L3.

3) Chronic Myeloid Leukemia (CML): This form of cancer happens once the myeloid cells endure a genetic modification. When the genetic modification occurs in cells, the tradition cells could not fight properly with the infections. This type of leukemia is common among adults and is a slow growing type of cancer. The CML cancer has further 3 stages which are known as chronic phase, accelerated phase and the blast phase. In first stage, cancer is in developing stage and develops very slowly so curable at this stage. In the second phase, it becomes more effective and starts harming the blood cells and at last stage blast of cells occurs [2, 3].

4) Chronic Lymphocytic Leukemia (CLL): This form of leukemia affects the blood cells as well as the bone marrow.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

With this kind of cancer, the white blood cell count will increase however they are doing not work properly. If someone were to have cancer, this could be the one with the very best survival rate. It is mostly found in the case of adults and very rare in the case of children. CLL is characterized by the clonal expansion and accumulation of leukemic cells with B-lymphocyte characteristics. It may so happen that people diagnose with CLL may lead to the case of ALL [1, 4].

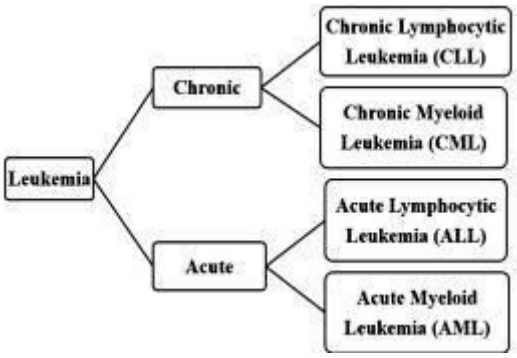

Fig. 1. Types of Leukemia

Many people cannot afford to test or cure it. A system that can recognize cancer cells and produce output based on those results is essential to helping us with this. Thus, machine learning can predict these cells and provide an accurate yield. Rapid cell growth is the earliest sign of malignancy. Cancerous cells can spread from almost anywhere in the body. This is called blood cancer because it can occur in the blood. Leukemia, or blood cancer, is a life-threatening disorder. It starts in the bone marrow and produces many cells. By then, they infect blood cells and kill them. It can kill adults and children if neglected.

II. RELATED WORK

Alexandra Bodzas and colleagues [18] used University Hospital in Ostrava input images in their 2020 study. Preprocessing uses gamma correction and standard arithmetic. Leukocyte localization and region extraction segment leukocytes after pre-processing. Thresholding, filtering, nearby cell identification, and cell separation are the four steps of leukocyte localization. Region extraction separates the nucleus and the cytoplasm. The method involves nucleus localization, extraction, and cytoplasmic extraction. For more useful results, segmented picture data is extracted using feature extraction. Morphological feature extraction measures the nucleus and cytoplasm's area and perimeter, resulting in a nuclear-cytoplasmic ratio. Some morphological extractions occur using this method. Nucleus compactness, shape factor, eccentricity, and solidity. Image blast cells are detected using statistical extraction. This technique extracts statistics. Supervised learning methods like SVM and ANN classify. About 97.52% of classifications are accurate.

SVM was used to identify leukemia by Siddhika Arunachalam et al. [19] (2020). The data comes from an online database. Pre-processing converts pictures to RGB. Borders and objects are often identified using image segmentation. The marker-controlled watershed segmentation technique accurately separates all associated nuclei from the image by detecting cell borders. The HSV color-based segmentation technique removes dust, scratches, and noise. Feature extraction follows segmentation. The effected cells' size, mean, variance, and standard deviation can be calculated using feature extraction. The SVM classifier correctly classifies leukemia. Classification accuracy is 93%.

Italy Joseph Maria et al. (2020) presented a leukemia categorization article. Classifiers were compared in this study to get important results. A blood sample is classified as ALL or AML using SVM, a binary classifier. Lazy learners like the KNN algorithm classify or analyze data by memorizing its training dataset. It gets no classification training. Installing is easy, and adding data is straightforward. Naïve Bayes is a probability model that assumes variable independence. This simple, efficient method takes little data. The classification accuracy is 80.88%. The correct classifications are compared. Saif S. Aljaboriy et al. [20] (2019) used ALL-IBD1. Preprocessing boosts image contrast. Histogram equalization, minimum filter, and linear contrast raise contrast and highlight brighter objects on the dark backdrop. The pre-processing accuracy is 99.517%. Image segmentation isolates specific areas. Global thresholding

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

compares threshold and intensity values. Pixels are set to 0 if the intensity is below the threshold. If the value exceeds the threshold, the pixel value is 1. Quick thresholding and simple computations. The K-means clustering algorithm groups pixels by distance. Large datasets are handled quickly and efficiently. Morphological segmentation segments blast cells by erosion and dilatation. Image-intersecting cells are identified via watershed segmentation. Approximately 96.29% accuracy. Features are extracted from segmented images using feature extraction. Shape, texture, and color extraction.

Shweta Suresh Naik et al. [21] (2019) pre-processed an internet dataset. Nuclei and cytoplasm pictures are segmented. Thresholding boosts visual contrast. Clustering identifies affected cells. Image areas are extracted using region-based segmentation. Feature extraction extracts specific image features.

III. PROPOSED ALGORITHM

The proposed algorithm used to predict Leukemia diseases is shown in Fig 2.

A. Data Source

In this paper, the dataset of leukemia diseases in the form of images. It contains 15,135 images from 118 patients with two labelled classes: Normal; Cancer.

B. Data pre-processing

The goal of pre-processing is to improve the quality of image data by reducing unwanted distortions and highlighting important image features needed for further processing and analysis. Pre-processing encompasses several steps, including image reading, resizing, noise removal using denoising methods, segmentation, and applying morphology operations to refine edges. Feature extraction involves converting the image data into a set of features suitable for pattern recognition. This phase concentrates on isolating and characterizing features from segmented objects within the image or from the entire image itself. The following Fig. 2 shows the methodology used to predict Leukemia diseases.

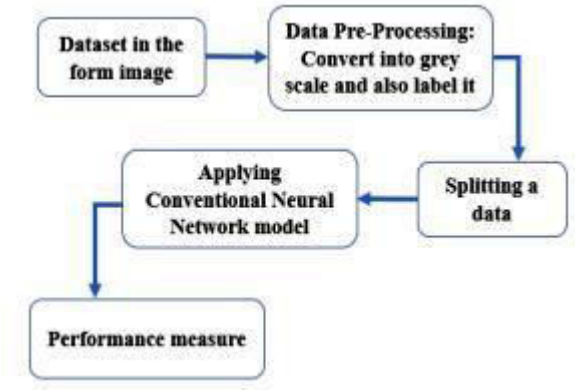

Fig.2. Proposed Methodology for Leukemia Disease Prediction

C. Data Splitting

There are 80% of the data (i.e.) images used for training and rest for testing. The subset of a dataset used to train a model is training set. The data set used to verify the model after initial validation process is termed as test set.

D. Convolutional Neural Network (CNN)

A CNN is specifically utilized to identify complex patterns within datasets. CNNs have diverse applications, ranging from tasks such as image recognition to supporting vision in robotics, interpreting text within images, and facilitating the operation of autonomous vehicles. Structurally, CNNs are composed of layers of neurons optimized for recognizing two dimensional patterns. Typically, CNNs include three main layers: the convolutional layer, pooling layer, and fully connected layer. In the network architecture, total 11 layers are used after exclusion of the input layer. The input layer processes RGB color images, analyzing color of each channel independently to extract significant features.This

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

iterative process is continued, adhering to the same methodology, until it reaches completion of 10 epochs.Ultimately, the convolutional network's performance is gauged using the data from the test set.

E. Performance Measures

Model performance can be evaluated using a confusion matrix, containing True Positive, True Negative, False Positive, and False Negative values, which are used to derive various parameters. Although accuracy is a widely used metric for assessing model performance, it may not always provide a clear indication. Precision quantifies the percentage of correctly predicted positive instances out of the total predicted positive instances, indicating the model's accuracy in identifying positive cases. In contrast, recall measures the percentage of correctly predicted positive instances out of the total actual positive instances, illustrating the proportion of true positives captured by the model. The harmonic mean of precision and recall is termed as F1-score. The performance indicator of model is indicated using the F-1 score. For the better performance of any model, F1-score should be high. The total number of elements in the predicted class is represented as Support that offers the additional context for evaluation of performance of model.

The loss observed in the proposed CNN model for the multiple epochs is depicted in Fig. 6. The correlation matrix highlights the connection between the actual and predicted classes produced from proposed CNN model is shown in Fig. 7. The instances are categorized correctly by means of the matrix that offers the accuracy of the model. The overall performances of CNN model for leukemia cases are tabulated in Table I

PSEUDO CODE

Step 1: Generate all the possible data source
Step 2: Preparing data for pre-processing
Step 3: Extracting data by data splitting technique
Step 4: Apply CNN algorithm for analyzing color of each channel independently to extract significant features.
Step 5: Calculating Performance Measures through Accuracy,Precision,Recall,F1 Score

ACTUAL VALUES	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Step 7:This iterative process is continued, until it reaches completion of epochs
Step 8:End

IV. SIMULATION RESULTS

The dataset of 15,135 images collected from 118 patients that is available on Kaggle have been utilized in this paper. The effectiveness of the proposed model is assessed using the images collected from 118 patients. The CNN model is used in this paper and the results are displayed in Fig. 3

The distribution of samples at training and validation stages is depicted in Fig. 4. A total of 4971 images used for training and 1,249 samples of images have been used for validation process. The accuracy of proposed CNN model over multiple epochs is shown in Fig. 5. The performance trajectory as training progresses for successive epochs is displayed in Fig. 5.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 150, 150, 16)	448
conv2d_2 (Conv2D)	(None, 150, 150, 16)	2320
conv2d_3 (Conv2D)	(None, 150, 150, 32)	4640
conv2d_4 (Conv2D)	(None, 150, 150, 32)	9248
conv2d_5 (Conv2D)	(None, 150, 150, 64)	18496
conv2d_6 (Conv2D)	(None, 150, 150, 64)	36928
max_pooling2d_1 (MaxPooling2	(None, 75, 75, 64)	0
flatten_1 (Flatten)	(None, 360000)	0
dense_1 (Dense)	(None, 64)	23040064
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 2)	130

Total params: 23,112,274
Trainable params: 23,112,274
Non-trainable params: 0

Fig.3. Simulation Result for sequential CNN

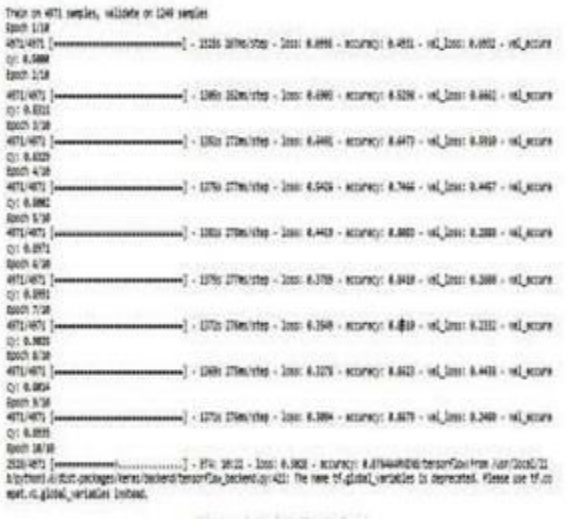

Fig.4. Model Training

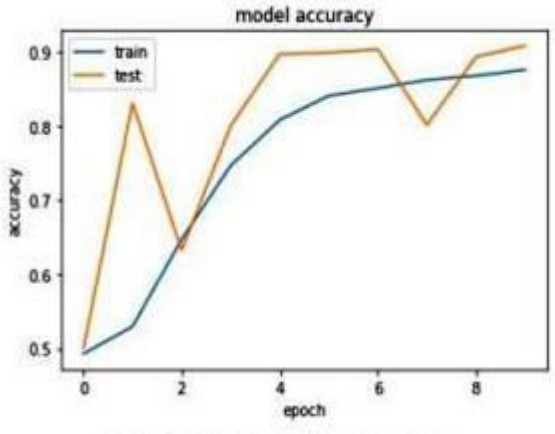

Fig. 5. Model accuracy with respect to epoch

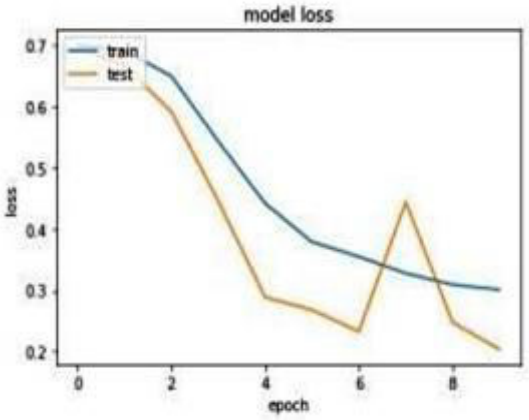

Fig. 6. Model loss with respect to epoch

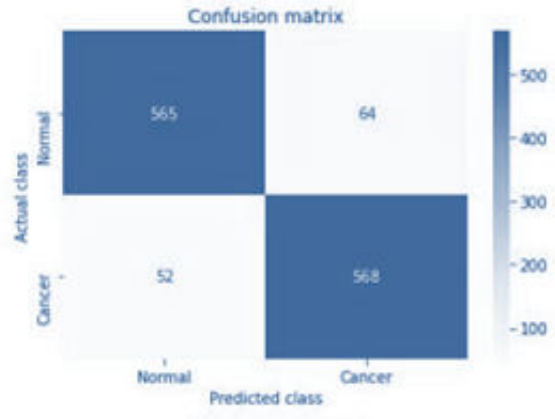

Fig.7. Confusion Matrix

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The performance measures such as accuracy, precision, recall, F-1 Score depicted in Table I provide the predictive accuracy of Leukemia diagnosis.

TABLE I. PERFORMANCE OF CNN MODEL

	Precision	Recall	F1-Score	Support
Not detected	0.92	0.90	0.91	629
Detected	0.90	0.92	0.91	620
Accuracy			0.91	1249
Macro Avg.	0.91	0.91	0.91	1249
Weighted Avg.	0.91	0.91	0.91	1249

V. CONCLUSION AND FUTURE WORK

For many medical professionals and pathologists, detecting leukemia illness is a major challenge. Leukemia has been predicted using a variety of methods. One of the many applications of machine learning (ML) is the examination of various leukemia picture kinds. ML algorithms are also used to identify ALL, a disorder that has received a lot of interest from the hematology and artificial intelligence disciplines.An overview of many studies on machine-learning techniques for leukemia diagnosis is given in this study. A CNN model for leukemia illness detection has been proposed in this research. Machine learning algorithms can speed up the diagnosis of leukemia and improve patient survival rates.

REFERENCES

1. American Cancer Society, “facts spring 2014 ¦Leukemia Lymphoma Society: Fighting Blood Cancer, Revised April 2014.
2. Kalyanmoy Deb, A. Raji Reddy, —Reliable classification of two-class cancer data using evolutionary algorithms¦, Elsevier,BioSystems , Vol.72, pp.111–129, 2003.
3. M. Oostindjer, J. Alexander, G. V. Amdam, G. Andersen, N. S.Bryan, D.Chen, D.E.Corpet ,S.DeSmet, L.O.Dragsted, A.Hauget al., “The role of red and processed meat in colorectal cancer development: a perspective, ”Meatscience, vol.97,no.4,pp.583–596,2014.
4. R.Takiar, D.Nadayil, and A.Nandakumar, “Projections of number of cancer cases in india (2010-2020) by cancer groups, ”Asian Pac J Cancer Prev, vol. 11, no. 4, pp. 1045–1049, 2010.
5. A. Bodzas, P. Kodytek, and J. Zidek, “Automated Detection of Acute Lymphoblastic Leukemia From Microscopic Images Based on Human Visual Perception,” Front. Bioeng. Biotechnol., vol. 8, no. August, pp. 1–13, 2020, doi: 10.3389/fbioe.2020.01005.
6. S. Arunachalam, “Applications of Machine Learning and Image Processing Techniques in the Detection of Leukemia,” Int. J. Sci. Res. Comput. Sci. Eng., vol. 8, no. 2, pp. 77–82, 2020, doi: 10.26438/ijsrcse/v8i2.7782.
7. S. S. Aljaboriy, N. N. A. Sjarif, and S. Chuprat, “Segmentation and Detection of Acute Leukemia Using Image Processing and Machine Learning Techniques: A Review,” Ausrevista, no. September, p. 511, 2019, doi: 10.4206/aus.2019.n26.2.60.
8. J.Rakesh Babu.,” Road sign intimation through voice alert system using deep Learning”, International Journal for Advanced Research in science & Technology, ISSN 2457 – 0362

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

9. Bakka Rajeev Gandhi "Prediction And Classification Of Alzhimer's Disease Using Machine Learning Techniques In 3d Mr Images", at the international conference on sustainable computing and smart systems(ICSCSS 2023)
10. J.Rakesh Babu.," Road Sign Intimation Through Voice Alert System Using Deep Learning ", International Journal for Advanced Research in science & Technology, ISSN 2457 – 0362
11. Jonnakuti Rakesh Babu,Bakka Rajeev Gandhi "Anomaly Attack Identification Security System Using Artificial Intelligence and Deep Learning"International Journal of Innovative Research in Computer and Communication Engineering,e-ISSN: 2320-9801, p-ISSN: 2320-9798
12. J Rakesh Babu,"Neural Network Approach For Prediction Of Crop Yield Using RNN, Feed Forward And LSTM Algorithms"International Journal of Engineering Science and Advanced Technology (IJESAT),ISSN No: 2250-3676,Vol24 Issue 02, 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details