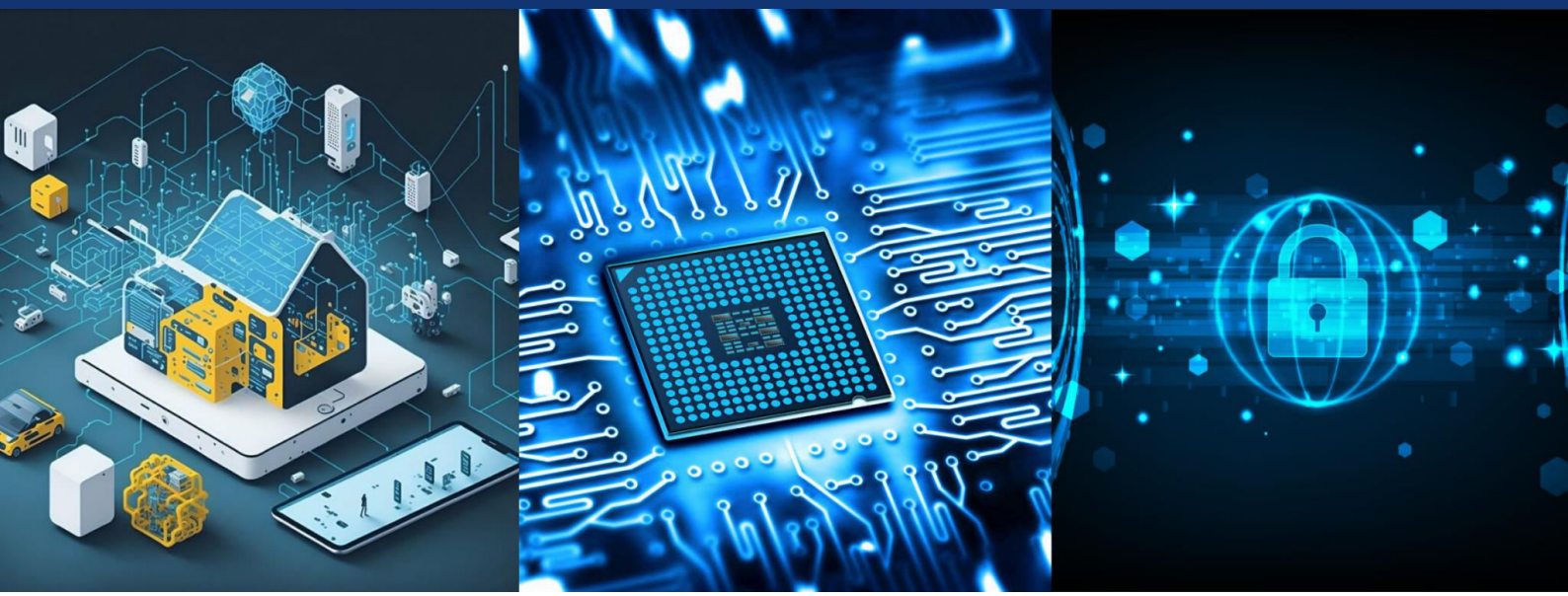
IJIRCCE

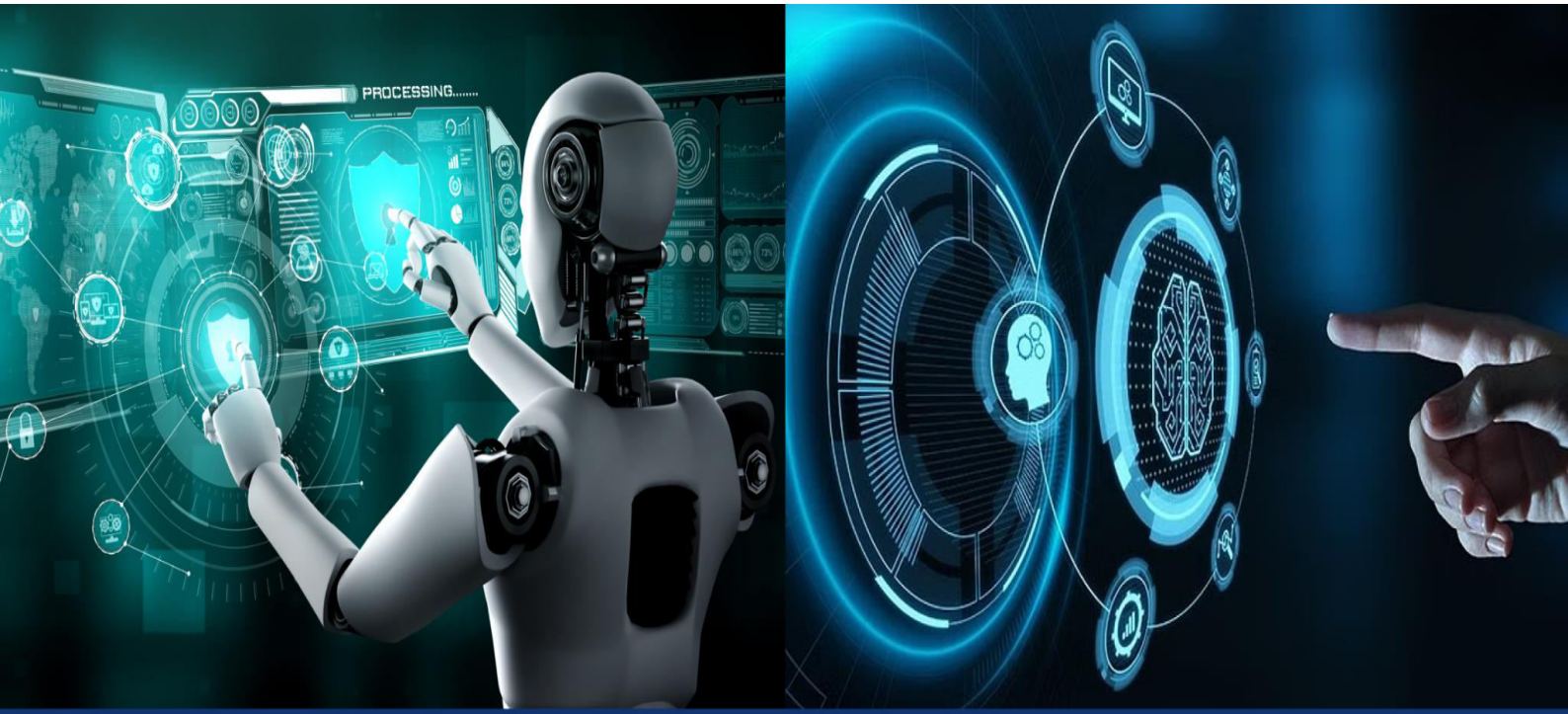
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405020

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images

Gowtham M[1], Dayanand TC[1], Inbarasan S[1], Mr.Raja Monsing R [2]

Department of Artificial Intelligence and Data Science, Christ The King Engineering College Karamadai, Coimbatore, Tamil Nadu, India. [1]

Assistant Professor, Department of Artificial Intelligence and Data Science, Christ The King Engineering College Karamadai, Coimbatore, Tamil Nadu, India. [2]

ABSTRACT: Traditional detection methods are time-intensive, laborious, and depend on skilled manual examination of bone marrow or peripheral blood smears. However, research in automated leukemia detection has significantly advanced with the development of sophisticated image processing techniques using Machine Learning (ML) and Deep Learning (DL) approaches. This literature review analyzes recent studies on automated leukemia detection, utilizing various specimens such as gene expression data, images of bone marrow, and peripheral blood smears. It also provides a list of public repositories offering access to these datasets. This article reviews studies on the automatic detection of leukemia using Peripheral Blood Smear (PBS), Bone Marrow (BM), and gene expression data, and finds that most studies achieve over 90 % accuracy, showcasing the effectiveness of Artificial Intelligence-based techniques. Machine Learning (ML) algorithms, by integrating a comprehensive range of morphological features, offer precise and effective disease diagnosis.

KEYWORDS: Agranulocytes, Granulocytes, Deep learning, Convolutional Neural Networks

I. INTRODUCTION

Stem cells are undifferentiated cells from which different types of body cells are derived. They also help to repair damaged tissues and are seen in various organs, including the brain, blood, bone marrow, muscle, skin, heart, and liver [1]. Embryonic and adult stem cells are the two primary categories of stem cells, distinguished by the stage at which they originate in the human body.

Hematopoietic stem cells are adult stem cells that mostly live in the spongy, red gelatinous tissue called bone marrow inside some bones. Red bone marrow, known as myeloid tissue, and yellow bone marrow, known as fatty tissue, are the two types of bone marrow. The red bone marrow produces blood cells and platelets, while fat and other stem cells that form bone and cartilage are mostly found in the yellow bone marrow. It is in the red bone marrow, that the hematopoietic cells divide into new blood cells, which eventually mature and enter the bloodstream as peripheral blood stem cells [2]. The immature stem cells, called blasts make up fewer than 5 % of all bone marrow cells and are typically absent in the blood of healthy individuals [3]. When blasts overcrowd healthy blood cells, they interfere with the normal production and function of blood cells, leading to various symptoms and complications that characterize leukemia.

Traditional leukemia detection methods are human dependent which is time-consuming and can cause inter-observer and intra-observer variability in examination results. Adept hematologists or pathologists are essential for distinguishing and categorizing the blasts based on their morphology and other features, especially when dealing with cases where the features are not clear-cut. Timely diagnosis of leukemia is crucial for enhancing prognosis and overall well-being. It allows for prompt treatment, early detection before progression, effective complication management, symptom reduction, better quality of life, and essential psychological support for patients and families.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405020

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

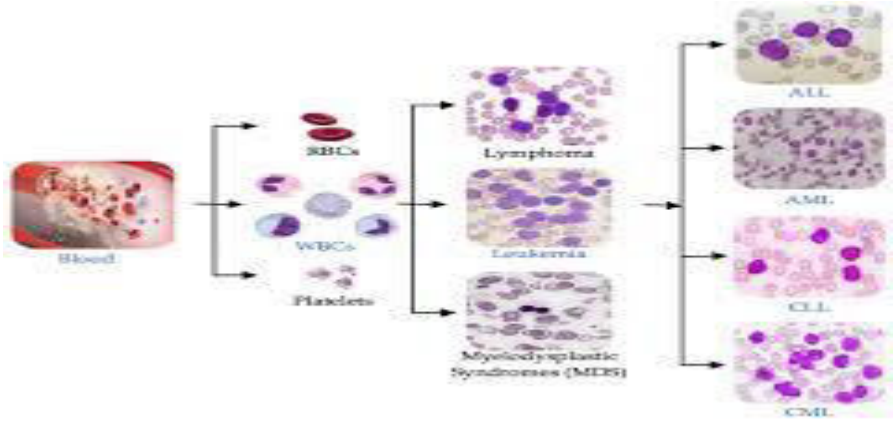


Fig 1: Deep Learning Techniques for Leukemia Cancer Classification

Leukemia is a type of cancer that affects the blood and bone marrow, leading to an abnormal proliferation of immature white blood cells (also known as leukocytes). These abnormal cells, called leukemia cells, disrupt the normal function of healthy blood cells, impairing the body's ability to fight infections and control bleeding.

A total of 474,519 new cases of leukemia were reported in 2020. The global age-standardized incidence rate was 5.4 per 100,000, with an almost five-fold variation worldwide [1]. The American Cancer Society estimates that in the United States, in 2023, there will be over 6500 new cases of Acute Lymphoblastic Leukemia (ALL) and almost 1400 deaths will have occurred. Additionally, projections for the United States in 2024 estimate approximately 62,770 new leukemia cases and about 23,670 related deaths. Sixty percent of all ALL cases occur in children, with a peak incidence at age 2 to 5 years; a second peak occurs after age 50. ALL is the most common cancer in children and represents about 75% of leukemias among children < 15 years of age. The risk declines slowly until the mid-20s and then begins to rise again slowly after age 50. ALL accounts for about 20% of adult acute leukemias. The average lifetime risk of ALL in both sexes is about 0.1% (1 in 1000 Americans) [2].

II. LITERATURE REVIEW

The articles were reviewed and analyzed from the oldest to the most recent publications within the specified period (2019–2023). This approach allows for the systematic exploration of how research in the field has evolved. Each article published within the defined period was systematically examined to understand its contributions, methodologies, findings, and relevance to the research questions. This method ensures thorough coverage of the existing literature. Following a chronological order enables researchers to track the development of ideas, methodologies, and trends in leukemia classification using deep learning techniques. This allows them to identify seminal works, key advancements, and gaps in knowledge as the research progresses.

In 2019, Nizar Ahmed et al. [7] proposed an innovative approach for diagnosing all leukemia subtypes from microscopic blood cell images using Convolutional Neural Networks (CNN). It explores the use of deep learning techniques to classify leukemia subtypes accurately. The proposed method uses CNN architecture for leukemia diagnosis, focusing on all four subtypes. Data augmentation techniques were applied to increase the dataset size, including rotation, height shift, width shift, zoom, horizontal flip, vertical flip, and shearing. Two publicly available leukemia datasets were used: ALL_IDB [8] and ASH [9]. The CNN model achieved 88.25% accuracy in binary classification (leukemia versus healthy) and 81.74% accuracy in multi-class classification of all leukemia subtypes. Comparative analyses were conducted with other machine learning algorithms, highlighting the effectiveness of the CNN model.

Rohit Agrawal et al. [10] proposed a novel method involving preprocessing, segmentation, feature extraction, and classification using a Convolutional Neural Network (CNN) to achieve a correct diagnosis. The proposed system aims to assist medical professionals in diagnosing several types and sub-types of white blood cell cancer diseases, including Acute Myeloid Leukemia (AML), Acute Lymphoblastic Leukemia (ALL), and Myeloma. The dataset comprises 100 microscopic blood sample images [11], with 62 images for training and 38 for testing. Images of blood samples are

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405020

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

converted from RGB to YCbCr color space during input to improve picture quality and make segmentation easier. Utilizing Otsu Adaptive Thresholding in conjunction with Gaussian Distribution to segment images. The K-Means clustering approach is used to identify the segments of the cytoplasm and nucleus and to segment cells. Cells are segmented using K-Means clustering. Texture feature extraction was performed using the Gray Level Co-occurrence Matrix (GLCM). The exact categorization of white blood cell cancer subtypes is achieved using Convolutional Neural Networks (CNNs). The model achieved 97.3% accuracy.

Sara Hosseinzadeh Kassani et al. [12] proposed a hybrid method enriched with different data augmentation techniques to extract high-level features from input images. Features from intermediate layers of two CNN architectures, VGG16 and MobileNet, are fused to improve classification accuracy. Two methods for normalization are mean RGB subtraction divided by standard deviation and ImageNet mean subtraction. Images are resized from 450 × 450 pixels to 380 × 380 pixels using bicubic interpolation. For data augmentation, various techniques are applied, including contrast adjustments, brightness correction, horizontal and vertical flips, and intensity adjustments. Hybrid CNN Architecture combines features from VGG16 and MobileNet architectures to enhance classification accuracy. Features from specific abstraction layers are fused to improve discriminative capability. It utilizes low-level features from intermediate layers to generate high-level discriminative feature maps. The dataset used for experimentation is based on the classification of normal versus malignant cells in B-ALL white blood cancer microscopic images provided by SBILab. The dataset has 76 individual subjects, 47 ALL subjects, and 29 normal subjects, including 7272 ALL cell images and 3389 normal cell images [13]. The proposed method achieves an overall accuracy of 96.17%, sensitivity of 95.17%, and specificity of 98.58%. Comparative analysis shows that the proposed approach outperforms individual CNN architectures (VGG16 and MobileNet) and earlier studies regarding accuracy.

III. METHODS

Data augmentation: In scenarios where the dataset is limited or imbalanced, image data augmentation proves to be a useful technique. It involves generating modified versions of original images through various transformations to increase dataset dimension and heterogeneity. These augmented images, along with the original dataset, are used for training ML models, improving their generalization, and robustness, and reducing the risk of overfitting. Typical augmentation methods include flipping, rotation, scaling, cropping, translation, brightness and contrast adjustments, and introducing Gaussian noise [50]. TensorFlow, Keras, PyTorch, OpenCV, Pillow, and Albumentations are some of the libraries and frameworks that include built-in tools and functions for data augmentation. These packages can be used to create and implement offline or online data augmentation pipelines for the images.

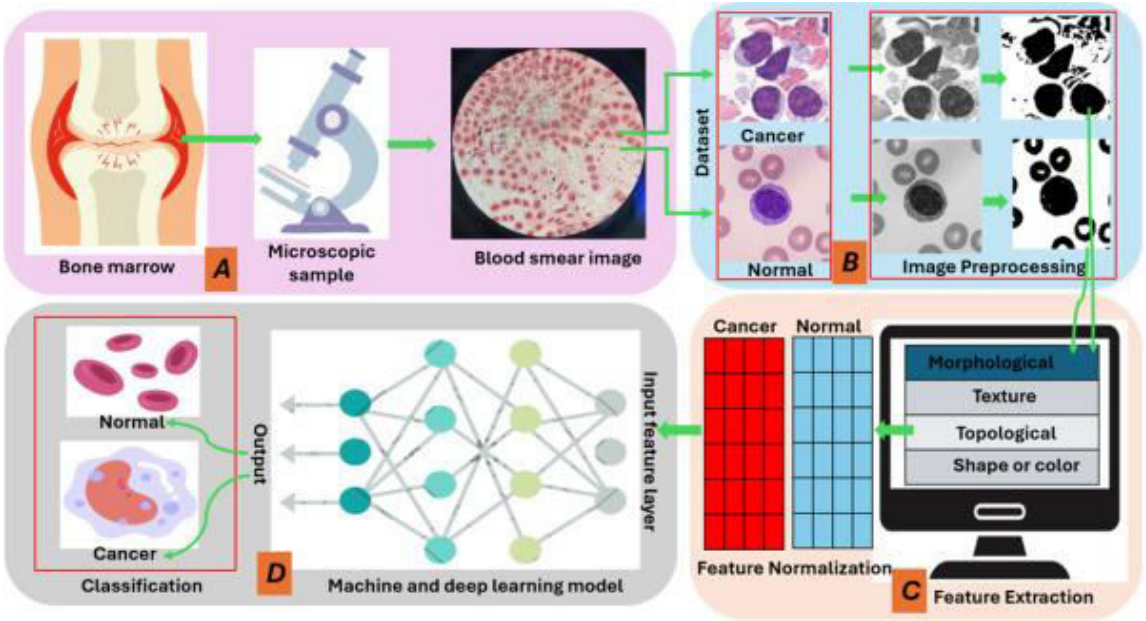


Fig 2: machine and deep learning techniques for acute lymphoblastic leukemia

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

Offline and Online Data Augmentation: Offline data augmentation involves pre-generating and storing augmented images before model training. While it can reduce computation time and memory requirements, it may increase disk space usage and introduce data redundancy, making it preferable for smaller datasets. Conversely, online data augmentation, or augmentation on the fly, generates augmented images during model training. This method increases data diversity and saves disk space but can extend computation times and complicate training, making it more suitable for larger datasets. Online augmentation, which involves transforming mini-batches of images provided to the Convolutional Neural Network (CNN).

Generative Adversarial Networks (GANs) can be a powerful tool for data augmentation in Machine Learning tasks, particularly in situations with limited labeled data [53]. Two neural networks; a generator, and a discriminator, that are trained in competition with one another make up the generative model, GAN. The discriminator's goal is to distinguish between real and generated data, while the generator's objective is to create data that closely resembles the training data. Through this adversarial training process, the generator enhances its capability to generate realistic data.

IV. RESULT ANALYSIS

ALL-IDB, C-NMC 2019, and the ASH Image Bank. These datasets provide critical resources but also present significant limitations that challenge their application in deep learning. For instance, The ALL-IDB dataset [11] is specifically designed for Acute Lymphoblastic Leukemia (ALL) research and contains annotated microscopic blood cell images. It is widely used for segmentation and classification tasks due to its targeted focus and expert annotations. However, the dataset's scope is narrow, as it only includes ALL cases and healthy samples, omitting other leukemia types such as Acute Myeloid Leukemia (AML). Its small sample size further limits its utility for deep learning, requiring large datasets to train complex architectures effectively. Moreover, the dataset lacks variability in image acquisition conditions, such as differences in staining, illumination, and noise, reducing its ability to generalize to real-world clinical settings. These shortcomings highlight the need for more comprehensive and diverse datasets representing a broader spectrum of leukemia cases and imaging conditions.

On the other hand, the ASH Image Bank [12] provides a publicly available repository of hematological images and is often used as a supplementary dataset in leukemia research. Its comprehensive nature, covering a wide array of hematological topics, makes it a valuable resource for training and validation. However, this dataset lacks the standardization required for specific tasks such as leukemia classification. Images vary in quality, resolution, and format, complicating their direct use in deep learning pipelines. Furthermore, inconsistent and incomplete annotations limit its utility for supervised learning, where precise labels are critical for model performance. These factors necessitate preprocessing steps to normalize the images and detailed re-annotation efforts to align the data with the requirements of deep learning techniques.

The C-NMC 2019 dataset [13] offers a larger collection of segmented microscopic images, comprising 15,135 images from 118 patients. It includes real-world noise, such as staining imperfections and illumination errors, making it more representative of clinical scenarios. Additionally, the dataset is annotated by expert oncologists, ensuring high accuracy in its ground truth labels. Despite these strengths, the C-NMC 2019 dataset faces significant challenges. The dataset exhibits class imbalance, primarily focusing on two categories: normal cells and leukemia blasts, leading to biases in model training. Furthermore, the morphological similarity between these two classes makes feature extraction and classification particularly difficult for deep learning models. Limited patient diversity, with images derived from a relatively small number of individuals, further restricts the generalization of models trained on this dataset. Addressing these issues requires datasets with more balanced class distributions and increased patient diversity to ensure broader applicability.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405020

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

Fig 3: Enhancing acute leukemia classification

These datasets collectively highlight the pressing need for more robust data resources for deep learning applications. The key requirements include larger sample sizes to train deep learning architectures effectively, greater diversity in patient demographics and disease subtypes to improve generalization, and detailed and consistent annotations. Additionally, incorporating real-world variability in imaging conditions, such as noise, lighting, and staining differences, can enhance the robustness of models in clinical settings. Advanced techniques like data augmentation can artificially expand datasets by introducing variations such as rotations, flips, and brightness adjustments, thereby mitigating the issue of limited data. Synthetic data generation using methods like Generative Adversarial Networks (GANs) can also create realistic images to address dataset scarcity. Furthermore, integrating multiple datasets through cross-dataset training can enhance model performance by exposing it to diverse data sources. Researchers can significantly improve models' accuracy, robustness, and applicability in leukemia diagnosis by addressing the underlined limitation in combination with advanced deep learning techniques. The most suitable or effective dataset for leukemia classification using deep learning may vary depending on data quality, diversity, and relevance to clinical scenarios. However, the ALL-IDB dataset is often cited in the literature due to its specific focus on leukemia and its established use in research studies.

V. CONCLUSIONS

This paper provides insights from a systematic mapping study (SMS) and a systematic literature review (SLR) focused on deep learning techniques for leukemia diagnosis. By analyzing thirty studies, the review highlights a variety of methodologies that demonstrate promising levels of accuracy and sensitivity in detecting leukemia from blood smear images. The findings emphasize the importance of ongoing research to overcome current challenges and enhance the development of deep learning models that are accurate, reliable, and interpretable. Such advancements are essential for improving patient care and supporting informed decision-making in clinical settings. Integrating large language models (LLMs) into healthcare systems holds great potential to further transform diagnostics and treatment. LLMs' ability to adapt to evolving data and contexts can drive advancements in disease prediction, personalized care, and operational efficiency. However, addressing challenges like dataset limitations, ethical concerns, and practical implementation hurdles is critical to successfully adopting these technologies. With continued innovation and interdisciplinary collaboration, AI and machine learning can significantly reshape healthcare delivery, leading to improved outcomes and more equitable access to care.

REFERENCES

[1]. Huang, J.; Chan, S.C.; Ngai, C.H.; Lok, V.; Zhang, L.; Lucero-Prisno, D.E., III; Xu, W.; Zheng, Z.J.; Elcarte, E.; Withers, M.; et al. Disease burden, risk factors, and trends of leukaemia: A global analysis. Front. Oncol. 2022, 12, 904292. [Google Scholar] [CrossRef] [PubMed]

[2]. MSD Manual. Acute Lymphoblastic Leukemia (ALL). 2023. Available online: https://www.msdmanuals.com/professional/hematology-and-oncology/leukemias/acute-lymphoblastic-leukemia-all?ruleredirectid=745 (accessed on 8 November 2024).

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

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

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

[3]. Medicine, Y. Diagnosing Leukemia. Available online: https://www.yalemedicine.org/conditions/leukemia-diagnosis (accessed on 10 November 2024).
[4]. Cruz-Benito, J. Systematic Literature Review & Mapping. 2016. Available online: https://zenodo.org/records/165773 (accessed on 10 November 2024).
[5]. Theodore Armand, T.P.; Nfor, K.A.; Kim, J.I.; Kim, H.C. Applications of Artificial Intelligence, Machine Learning, and Deep Learning in Nutrition: A Systematic Review. Nutrients 2024, 16, 1073. [Google Scholar] [CrossRef] [PubMed]
[6]. Hosseini, S.H.; Monsefi, R.; Shadroo, S. Deep learning applications for lung cancer diagnosis: A systematic review. Multimed. Tools Appl. 2024, 83, 14305–14335. [Google Scholar] [CrossRef]
[7]. Ahmed, N.; Yigit, A.; Isik, Z.; Alpkocak, A. Identification of Leukemia Subtypes from Microscopic Images Using Convolutional Neural Network. Diagnostics 2019, 9, 104. [Google Scholar] [CrossRef]
[8]. Labati, R.D.; Piuri, V.; Scotti, F. All-IDB: The acute lymphoblastic leukemia image database for image processing. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2045–2048. [Google Scholar] [CrossRef]
[9]. The American Society of Hematology. Available online: http://www.hematology.org (accessed on 15 October 2024).
[10]. Agrawal, R.; Satapathy, S.; Bagla, G.; Rajakumar, K. Detection of White Blood Cell Cancer using Image Processing. In Proceedings of the 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 30–31 March 2019; pp. 1–6. [Google Scholar] [CrossRef]
[11]. Gupta, A.; Gupta, R. ALL Challenge Dataset of ISBI 2019 [Data Set]. The Cancer Imaging Archive. 2019. Available online: https://www.cancerimagingarchive.net/collection/c-nmc-2019/ (accessed on 10 November 2024).
[12]. Kassani, S.H.; Kassani, P.H.; Wesolowski, M.J.; Schneider, K.A.; Deters, R. A Hybrid Deep Learning Architecture for Leukemic B-lymphoblast Classification. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 16–18 October 2019; pp. 271–276. [Google Scholar] [CrossRef]
[13]. Loey, M.; Naman, M.; Zayed, H. Deep Transfer Learning in Diagnosing Leukemia in Blood Cells. Computers 2020, 9, 29. [Google Scholar] [CrossRef]
[14]. Blood Cell Images. Available online: www.kaggle.com/paultimothymooney/blood-cells (accessed on 19 October 2024).
[15]. Mathur, P.; Piplani, M.; Sawhney, R.; Jindal, A.; Shah, R.R. Mixup Multi-Attention Multi-Tasking Model for Early-Stage Leukemia Identification. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1045–1049. [Google Scholar] [CrossRef]

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