CLASSIFICATION OF ACUTE LYMPHOBLASTIC LEUKEMIA CELLS USING ARTIFICIAL INTELLIGENCE

Ayşe Berika VAROL MALKOÇOĞLU[1], İsmail İŞERİ[2*]

[1] Department of Computer Engineering, Istanbul University-Cerrahpaşa, Istanbul, Turkiye
[2] Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Turkiye

Keywords	Abstract
Artificial Intelligence, Machine Learning, Cancer Detection.	Due to the morphological similarity between immature lymphoblasts (cancerous cells) to lymphocytes (non-cancerous cells), detecting Acute Lymphoblastic Leukemia poses a significant challenge for pathologists. These cells, which exhibit a similar pattern, can lead to various errors during the diagnosis of the disease. In this study, the cancerous and non-cancerous cells were classified using 3 different artificial intelligence approaches. In the first approach, the classification process was carried out by training Convolutional Neural Networks in 4 different architectures. In the second approach, a hybrid approach was proposed by combining the convolution layer of the CNN model as the feature extractor with the Support Vector Machine, Naive Bayes and Random Forest algorithms as the classifier. The classification processes were carried out by training the proposed second approach. In the third approach, the classification process was performed using transfer learning process and ResNet50 and VGG16 networks. In all experiments, the effects of hyper-parameter and dataset changes on model performance were also examined. The results obtained by these three approaches were compared using the Accuracy, Precision, Recall, F-score, and AUC performance measures. It was determined that the most successful results were obtained with the 1st approach using the Dataset3.

AKUT LENFOBLASTİK LÖSEMİ HÜCRELERİNİN YAPAY ZEKA İLE SINIFLANDIRILMASI

Anahtar Kelimeler	Öz
Yapay Zeka, Makine Öğrenimi, Kanser Tespiti.	Olgunlaşmamış lenfoblastların (kanserli hücreler) lenfositlere (kanserli olmayan hücreler) morfolojik benzerliği Akut Lenfoblastik Lösemi kanserinin tespitinde patologlar için zorlu bir problemdir. Benzer bir desene sahip olan bu hücreler, hastalığın teşhisi sırasında çeşitli hatalara neden olabilmektedir. Bu sebepten çalışma kapsamında kanserli ve kanserli olmayan hücreler 3 farklı yapay zeka yaklaşımı kullanılarak tespit edilmiştir. İlk yaklaşımda Evrişimsel Sinir Ağları 4 farklı mimaride eğitilerek sınıflandırma işlemi gerçekleştirilmiştir. İkinci yaklaşımda, özellik çıkarıcı olarak CNN modelinin evrişim katmanı, sınıflandırıcı olarak ise Destek Vektör Makinesi, Naive Bayes ve Rastgele Orman algoritmaları birleştirilerek hibrit bir yaklaşım sunulmuştur. Önerilen ikinci yaklaşım eğitilerek sınıflandırma işlemleri gerçekleştirilmiştir. Üçüncü yaklaşımda ise transfer öğrenme süreci ile ResNet50 ve VGG16 ağları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Tüm deneylerde modellerdeki hiperparametre ve veri setleri değişikliklerinin performans üzerindeki etkileri incelenmiştir. Bu üç yaklaşımla elde edilen sonuçlar Doğruluk, Kesinlik, Geri Çağırma, F-skor ve AUC performans ölçüleri kullanılarak karşılaştırılmış ve en başarılı sonuçların Dataset3 kullanılarak Yaklaşım1 ile elde edildiği tespit edilmiştir.

* İlgili yazar / Corresponding author: ismail.iseri@omu.edu.tr, +90-362-312-1919

CLASSIFICATION OF ACUTE LYMPHOBLASTIC LEUKEMIA CELLS USING ARTIFICIAL INTELLIGENCE

Ayşe Berika VAROL MALKOÇOĞLU[1], İsmail İŞERİ[2†]
Department of Computer Engineering, Istanbul University-Cerrahpasa, Istanbul, Turkiye
Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Turkiye

---

Highlights

- Detection of cancer cells,
- Hybrid model proposal and comparison with other approaches
- Effects of different hyper-parameters on the models

Purpose and Scope

The purpose of the paper is to analyze the impact of hyper-parameters in different approaches on the result and to propose a helpful system for pathologists. The scope is the detection of ALL cancer with pathology images.

Design/methodology/approach

In the study, similar studies were analyzed through a literature review. A hybrid model is proposed by combining the convolution layers of CNN architecture with classical machine learning algorithms. Then, the performances of CNN architecture, hybrid model and transfer learning approaches are tested and compared for different hyper-parameters.

Findings

The performance of 3 approaches with different hyper-parameters on low quality data is analyzed. Accordingly, the proposed hybrid model (Approach 2) was found to perform better than the others. When the effect of different size datasets on the models is analyzed, it is seen that the dataset with synthetic and original data produces better results than the others. It is also observed that our model performs well even on low quality datasets.

Originality

The classification of cancerous cells with the proposed hybrid model (Approach 2) has not been previously observed in the literature. The performance of this approach is analyzed and compared with others, and it is shown that it can produce remarkable results for classification processes in healthcare and similar fields. Therefore, it is thought that the proposed model can be used as an auxiliary system for pathologists.

---

1. Introduction

Acute Lymphoblastic Leukemia (ALL) is a blood cancer where the uncontrolled growth of immature lymphoblast cells eventually leads to the accumulation of deadly numbers of leukemia cells. This type of cancer is usually seen in children. As shown in Figure 1, according to the research conducted by World Health Organization (WHO) in 2022, leukemia was detected in 31.9% of children with cancer aged 0-14 years (Global Cancer Observation, 2024; Uzunhan and Karakaş, 2012). This disease generally seen in children between the ages of 2 and 5, occurs as a result of overproduction of the white blood cells (lymphocytes) in the form of lymphoblast in the bone marrow, and can lead to rapid death if it is not detected and not treated (Pediatric Treatment Editorial Board, 2021; Inaba et al., 2013). Therefore, the detection of ALL is crucial. During the detection of the disease, various tests such as physical examination, blood count, chest X-Ray, chemical blood study, and bone marrow biopsy are performed. Apart from these tests, morphological examination method is also used. However, during detection carried out with this method, some difficulties are encountered. This is due to the fact that immature blood cells (lymphoblasts) and normal blood cells (lymphocytes) are morphologically similar.

To fulfill this challenging task, pathologists try to diagnose the diseased area using traditional methods. Samples taken from the disease site are examined under the microscope by pathologists to detect the molecular basis of the disease, and then the most appropriate diagnosis is determined and treatment begins (Madabhushi, 2009). However, undesired human-related errors may occur during the diagnosis because the procedures are performed by humans (Sipes & Li, 2018). In recent years, several digital systems have been developed to prevent such possible errors and help pathologists. These developments have gained momentum with the digitization of data. The data previously diagnosed by pathologists were digitized and the concept of "digital pathology" emerged. Thus, with the development of digital pathology, the number of artificial intelligence-based studies has increased

---

† Corresponding author: ismail.iseri@omu.edu.tr, +90-362-312-1919

in recent years. To achieve better results than existing conventional disease diagnosis methods, digital pathology images are provided as input to various artificial intelligence models. These models are trained and then digital pathology images are classified and analyzed using Machine Learning (ML) algorithms (Madabhushi, 2009; Heffner *et al.,* 2021). Therefore, with the artificial intelligence models developed, it is aimed to analyze the data obtained from patients more objectively and quickly.
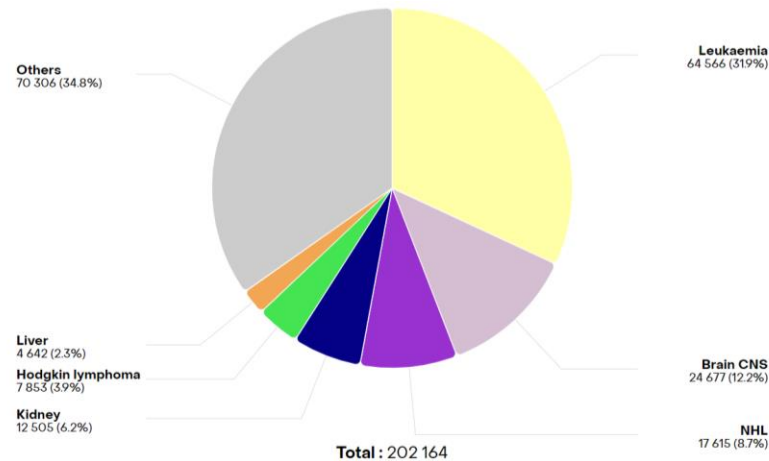

Figure 1. Cancer statistics of children between the ages of 0-14 in the World shared by WHO in 2022

In a study conducted by Selçuk and Özen (2015), 130 cancerous and 130 non-cancerous ALL cells were used. The researchers used image processing techniques on these data and classified the cells as cancerous or non-cancerous only with the Support Vector Machine (SVM) algorithm. In some other studies (Sipes & Li, 2018; Kumar *et al.,* 2018; Bhuiyan *et al.,* 2019), performance comparisons were carried out using multiple algorithms on ALL data. The common feature of the datasets they used was that the data were pre-classified as cancerous and non-cancerous. However, image scales and dataset size were different in each of these three studies. In the study conducted by Sipes and Li (2018), ALL-IDB and Cella Vision datasets were combined to classify all cancer types, and images were classified in Red Green Blue (RGB) color space with K-Nearest Neighbor (K-NN), Neural Network (NN), and Convolutional Neural Networks (CNN) algorithms. By separating the dataset, they used as 60% education, 20% validation, and 20% test, they assessed their performances. They achieved the best result by the CNN model with a 92% (± 3) accuracy rate. Similarly, K-NN and Naive Bayes (NB) algorithms were applied in the study carried out by Kumar *et al.* (2018), but their images were used in gray level color space. In this way, they observed that K-NN and NB produced 92.8% and 79% accuracy values, respectively. Bhuiyan *et al.* (2019) used microscopic images of resized 108 white blood cells and evaluated the classification performance by using Random Forest (RF), SVM, Logistics Regression (LR), and Decision Tree (DT) machine learning algorithms. Results showed that with an accuracy rate of 99.05%, the SVM algorithm was more successful than other algorithms. In another study using the ALL dataset (Duggal *et al.,* 2017), to obtain a large dataset, the image data was subjected to 180 degrees random and vertical rotation during the training phase and performance evaluations were carried out using AlexNet and T-CNN algorithms. However, the change of the hyper-parameters was not observed. In addition, this study also compared the performances of AlexNet and T-CNN algorithms by adding the SD layer, which analyzed the pixel density developed in the study, to these algorithms. As a result of this process, it was revealed that with 95.5% accuracy and 95.4% F-scoring rate, the best values were achieved with T-CNN-SD.

In the study of Zhao *et al.* (2017), where different pathology data were used, scholars utilized CNN and RF algorithms to detect 5 types of white blood cells. Jiang *et al.* (2018), examined 3 types of lymphoma and classified them with the SVM algorithm. In that study, the average accuracy performance value was calculated as 97.96%. In another study (Banik *et al.,* 2019), on the other hand, as in the study of Zhao *et al.* (2017), different blood cells were classified separately from each other. However, in this study (Banik *et al.,* 2019), scholars used data in RGB color space and benefited from the CNN algorithm. They similarly balanced the different unbalanced dataset by using a random data return method to eliminate unstable data, and they determined that the classification accuracy was over 90%. Perlberg & Kramer (2019) and Oliveira & Dantas (2021) used the same dataset as in this study. Perlberg & Kramer (2019) used ResNeXt to differentiate ALL-type cancerous cells from non-cancerous cells and used the Squeeze-and-Excitation module, which improves the channel dependencies of CNN models with almost no computation. This module was used by Hu *et al.* (2018) in the ImageNet competition, helping to improve the 2016 best result by approximately 25%. They performed an effective classification using the ResNeXt convolutional neural network with the Squeeze-and-Excitation module, and thus obtained an F1 score of 88.91%. Oliveira and Dantas (2021) classified cancerous and non-cancerous cells using the VGG16/VGG19 and Xception

model. By augmentation the data and running it on VGG16/VGG19 and Xception models, they obtained an F1-score value of 92.60% with the parameters they determined.

As observed in the literature, several machine learning studies have been examined for ALL cancer detection in biopsy images. In this study, different artificial intelligent approaches were examined. In the first approach, the performance of CNN models with different number of layers were analyzed. In the second approach, well-known machine learning methods (SVM, NB, RF) were used as classifiers instead of NN, and the convolutional feature extraction from the first approach was used as a feature extractor. In the third approach, the transfer learning-based study was carried out using the VGG16 and ResNet50 models. Briefly, the classification of the disease was carried out using digital pathology data and by artificial intelligence methods. Additionally, performances of the approaches under different hyper-parameters were examined and evaluated using experimental methods.

2. Materials and Methods

2.1. Description of The Problem

Leukemia occurs as a result of the decay of white blood cells in the maturation process. Figure2 shows the functioning of normal white blood production in a body. Lymphoid stem cells produced from stem cells form lymphoblasts. With the maturation of these lymphoblasts, white blood cells, which are called B lymphocytes and T lymphocytes, are formed. However, during ALL disease, due to an unknown reason in the blood formation process, the lymphoblast stage hesitates and shows an uncontrolled increase here. Lymphoblasts, which accumulate in the blood, tissues, and bone marrow, inhibit the production of normal bone marrow cells. This inhibition leads to a deficiency of red blood cells, thrombocytes, and normal white blood cells, ultimately causing cancer (Uzunhan & Karakaş, 2012). Although this situation is detected by pathologists, the obtained results vary according to the skill of the pathologist and the quality of the blood sample. Therefore, it is aimed to design assistive systems for pathologists by detecting ALL cancer with artificial intelligence approaches. To detect ALL cancer using artificial intelligence applications, it is necessary to differentiate between lymphocytes and lymphoblasts. However, as shown in Figure 2, it is known that it is difficult to distinguish between lymphoblasts and lymphocytes due to the morphological similarities of the cells.
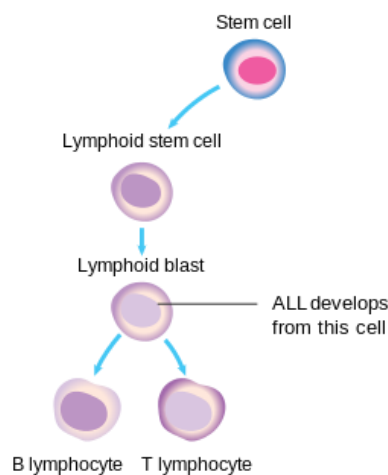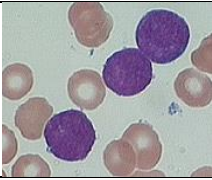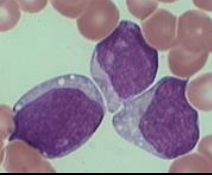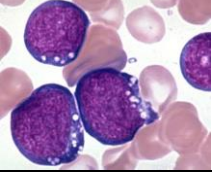

Figure 2. Formation diagram of white blood cells (Anonymous, 2023)

For this reason, it is essential to identify the types of cancerous cells. According to the FAB (French - American - British) classification technique, which is frequently used in the morphological differentiation of ALL, there are 3 types of disease forms consisting of L1, L2, and L3 whose explanations are given in Table 1. In this study, the classification was made using the data of L1 and L2 forms, which are the childhood types of ALL disease.

As it seems that, it is quite difficult to separate lymphoblasts in L1 form from lymphocytes. While lymphocytes have formally smooth edges and a single nucleus, lymphoblasts have similar sizes, small diameter homogeneous structure, few nucleoli, and narrow cytoplasm structure. Distinguishing between these two cell types by eye and detecting them objectively is a very challenging process. In this study, we aim to separate and successfully classify lymphocytes from immature lymphoblasts.

Table 1. ALL disease types

Disease Type	L1
The diameters of cells are small, they are homogeneous, have narrow cytoplasm, and nucleole is too small to be seen. It is the type of ALL seen in children. It accounts for 85% of all ALL patients.	

Disease Type	L2
The diameters of cells are different, they are heterogeneous, and have prominent cytoplasm and nucleolus. It is the type of ALL seen in adults. It constitutes 14% of all ALL patients.	

Disease Type	L3
The diameters of cells are large, they are homogeneous, similar to Burkitt lymphoma and are caused by B lymphocytes. It is also known as the "starry sky" because of the points on it, and it accounts for 1% of ALL patients.	

2.2. Dataset Preparation

An open-source dataset of ALL cancer type, prepared by the Cancer Imaging Archive was used in this study (Duggal *et al.*, 2017; Clark *et al.,* 2013; Duggal *et. al.,* 2016a; Duggal *et. al.,* 2016b; Karlik & Olgac, 2011; Mourya *et al.*, 2019). The selected dataset contains the pathology images of the most common cancer type diagnosed in childhood. This dataset consists of 10.661 labeled images at a resolution of 450x450 pixels in RGB format. Among these 3389 instances (Figure 3a) were labeled as non-cancerous and 7272 instances (Figure 3b) were labeled as cancerous cells.

In the machine learning phase, 3 different datasets containing samples with equal number of cancerous and non-cancerous labels were prepared by using the original data. In the first of these datasets, there were a total of 2532 samples (1266 cancerous and 1266 non-cancerous). In the second dataset there were a total of 5646 samples (2823 cancerous and 2823 non-cancerous). In the third dataset, there were a total of 12.140 samples (6070 cancerous and 6070 non-cancerous). While creating the third dataset, the data augmentation process described in section 2.3 was used, and this dataset contained both real and synthetic data. In addition to the different dataset sizes, the dimensions of the images used were also resized. The first set of images was reduced to 48x48 (Figure 3c and Figure 3d) and the second set of images was reduced to 32x32 (Figure 3e and Figure 3f). For the model, the obtained datasets were separated randomly as 70% training, 15% validation, and 15% test. In this way, the performance of the change in image sizes and different dataset sizes was also observed.

2.3. Data Augmentation

In the literature, various data augmentation processes are applied to increase the generalization ability of the models especially in deep learning and some machine learning studies. For this reason, considering that it would have a positive effect in our study, the data in the dataset was augmented by image processing methods and 12.140 sets were created. This augmented dataset contains both synthetic and real data and includes 6070 cancerous and 6070 non-cancerous samples.

During the data augmentation process, the total dataset was increased to 12.140 units by applying rotating, shearing, and polishing operations on the original images. New sample parts produced from the original image data because of the magnification process are shown in Figure 4. The parameters and values applied during the data augmentation process are as follow:
- rotation_range=30
- shear_range=0.2
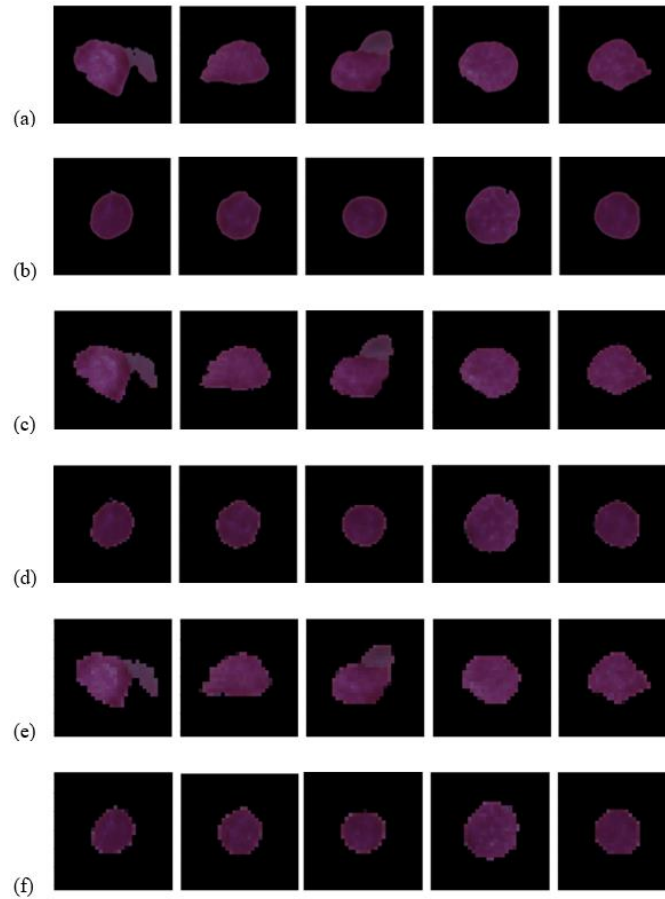- brightness_range= [0.1,1.0]

Figure 3. Cancerous (a) and non-cancerous (b) cells at 450x450 scale. Cancerous (c) and non-cancerous (d) cells at 32x32 scale. Cancerous (e) and non-cancerous (f) cells at 48x48 scale
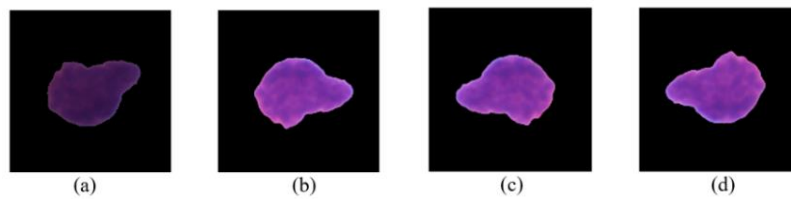
Figure 4. Increased view of cancerous cell Images of 450 x 450 scale. Original image (a), synthetic data (b), (c), (d)

2.4. Used Platforms and Libraries

In the experiments, a hardware with 12 GB RAM, 311 GB disk, and Tesla K 80 GPU features was used on the Google Colab platform. Applications developed using Python language were stored by Google Drive. Within the scope of this study, the following libraries were utilized:

- Keras and TensorFlow for deep learning,
- Scikit-learn for machine learning,
- Matplotlib for data visualization,
- NumPy for mathematical operations,
- OpenCV for image processing and image augmentation

2.5. Hyper-parameters

2.5.1. Activation Function

The main task of the activation function is to map a processed value to a specific range by limiting the output value of the neuron. If the learning of the network is carried out without the use of the activation function, it will have carried out limited learning (Karlik & Olgac, 2011). In this study, Rectified Linear Unit (ReLU) activation functions

493

were used in the outputs of the convolution layers and in the hidden layers of the fully connected layer. Sigmoid activation function was used in the output layer of the fully connected layer.

ReLU: It is an activation function that performs a thresholding operation for each input value (Eq. 1). It works faster than Sigmoid and Hyperbolic Tangent by reducing the cost of calculation. It is generally preferred in multilayer networks. However, when the majority of inputs consist of negative values, a lot of information is lost because their return will produce a value of zero, which prevents the network from learning (Si *et al.,* 2018; Strusa & Dolezel, 2019). This function is generally preferred in the convolution layer due to its low-cost load.

$$f(x) = \begin{cases} x & if & x \geq 0 \\ 0 & if & x < 0 \end{cases} \quad (1)$$

Sigmoid: It is an activation function (Eq. 2) that produces positive y output values between 0 and 1 values. It is commonly used in backpropagation algorithms (Sibi *et al.,* 2019). This function is preferred in the output layer and is particularly useful for binary classification tasks, where the result is either 0 or 1.

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

2.5.2.   Optimization Function

Optimization functions are used to find the most suitable solution for nonlinear problems. In this study, Adaptive Moment Estimation (ADAM), Stochastic Gradient Descent (SGD) and Root Mean Square Error Probability (RMSProp) were used. The features of these selected optimization functions are as follows:

ADAM: It is a stochastic-purpose optimization function that is a primary gradient of its functions and needs little memory. This function was created by combining the advantages of AdaGrad, which works well for sparse parameters, and RMSProp, which is effective for non-stationary objectives. The difference of it from the RMSProp function is that it directly estimates the parameter updates by calculating the first and second moment averages of the m_t in Eq. 3 and *i, v_t* gradients in Eq. 4. (Kingma & Ba, 2014). This dual moment estimation helps in achieving a more balanced and adaptive learning rate (Eq. 5.).

$$m_t = \frac{M_t}{1-\beta_1^t} \quad (3)$$

$$v_t = \frac{V_t}{1-\beta_2^t} \quad (4)$$

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{v_t}+\epsilon} m_t \quad (5)$$

SGD: It is an optimization method that minimizes the cost function during parameter updates by employing the gradient descent technique**.** Initially, randomly determined parameters are updated in the opposite direction of the gradient each time. This update occurs for each individual training sample, allowing SGD to potentially reach the local minimum faster and work more efficiently than other methods (Yazan & Talu, 2017). In Equation 6, x^i and represent the features and labels of the samples, respectively. The learning rate is denoted by *n* and *θ* refers to the parameter applied for the entire training set.

$$\theta = \theta - n\nabla_\theta J(\theta; x^i; y^i) \quad (6)$$

RMSProp: It   produces   parameter   updates   using   a   scaled-up   degree   of   momentum. It prevents the rapid decline of the learning rate in AdaGrad function, as with the AdaDelta function. Thanks to this function, by reducing oscillations, the algorithm is accelerated, and more accurate results are produced (Ruder, 2016). In this method, the learning rate is denoted by ***n***. The sum of the squares of the previous gradients, represented as ($E[g^2]_t$), is calculated to adaptively adjust the learning rate for each parameter.

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{E[g^2]_t+\epsilon}} g_t \quad (7)$$

2.6. Approaches

2.6.1. Approach1: Convolutional Neural Network

CNNs are basically a special artificial neural network model that consist of convolution, pooling, flattening and fully connected layers. In recent years, it has come to the forefront with its classification successes especially on large datasets (Si *et al.,* 2018). This study aims to enhance the classification performance of models by different hyper-parameters used in the layer numbers, filter sizes, and training of CNN models. To achieve this goal, 4 different CNN models were designed within the scope of this study. In the designed CNN models, ReLU activation function was used in the convolution layers and Sigmoid activation function was used in the output layers. The general structure of the designed CNN models was transferred to the convolutional layer as shown in Figure5 and the feature map was resized by sending it to the pooling layer. The resized feature map was converted to one-dimensional vector in the flattening layer and transferred to the fully connected layer for classification. There were 4 different CNN models created in this approach. These models consisted of 3, 6, 9 and 10 convolution layers and were called CNN-1, CNN-2, CNN-3, and CNN-4, respectively. Each CNN model used filters of different sizes and different step numbers. With these models, the effect of the increase in the number of layers on performance was measured.
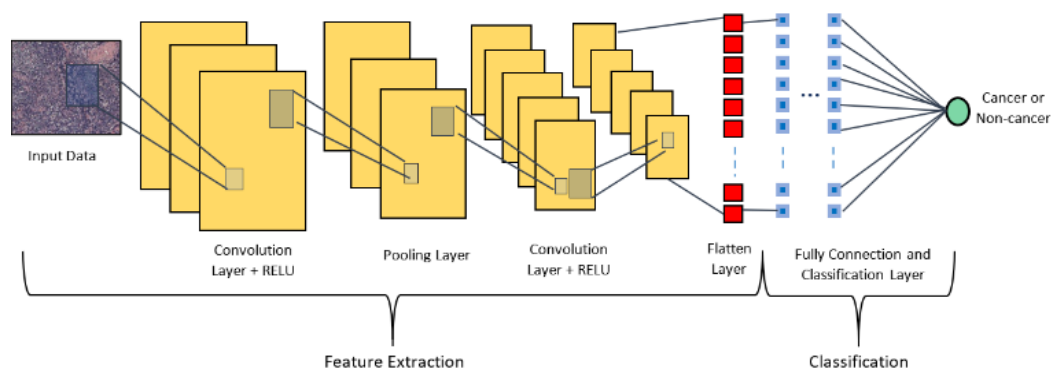

Figure 5. The structure of the convolutional neural network model

2.6.2. Approach2: Convolutional Feature Extraction and Machine Learning

Machine learning is an approach that can learn and predict using data and mathematical techniques. Feature extraction is made from data with the help of different mathematical techniques (Strusa and Dolezel, 2019). Classification, clustering or regression can be done by using these features obtained from data in various algorithms. In this approach, feature extraction was made using the convolution process and NB, RF and SVM classification algorithms were used instead of the fully connected layer. Feature vectors from the CNN-1, CNN-2, CNN-3, and CNN-4 models designed on Approach1 were used by NB, RF, and SVM classifiers for training and testing. The new feature extraction models were called Feature Extraction Model-1 (FEM-1), Feature Extraction Model-2 (FEM-2), Feature Extraction Model-3 (FEM-3), and Feature Extraction Model-4 (FEM-4), respectively. The vectors obtained from the created feature extraction models were given as inputs to NB (Sibi *et al.,* 2013), RF (Kingma & Ba, 2014) and SVM (Yazan & Talu, 2017) algorithms. With this approach, feature vectors were classified by other machine learning algorithms instead of NN algorithm. By comparing the performance results of these hybrid models, a comparison was made between the 4 models used for feature inference and the 3 classifiers. The basic structure of the established classification network is shown in Figure 6.
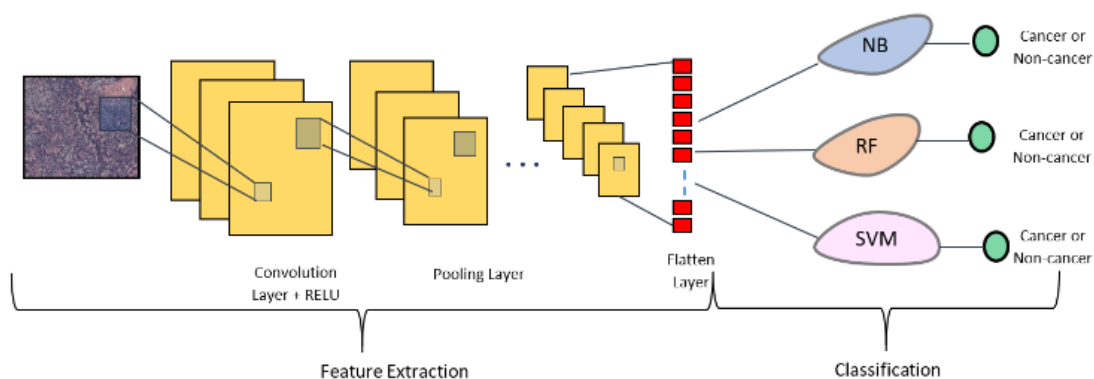

Figure 6. Classification of data extracted with CNN models with ML algorithms

2.6.3. Approach3: Transfer Learning

Transfer Learning (TL) is known as the ability of the system to transfer information learned from previous tasks to new tasks. It aims to reuse information from one or more sources for a new given task. Unlike traditional learning techniques, which start to learn things from scratch every time, it performs its learning by storing information (Ruder, 2016).

In this study, VGG16 and ResNet50 models, which are well-known classic deep learning models, were used for feature extraction through transfer learning. The VGG16 model is a model with 13 convolution layers, 5 maximum pooling layers and 3 full connection layers, resulting from improvements to the AlexNet model. It uses 11x11, 5x5, 3x3, and 1x1 filters in the convolution layers. The ResNet50 model, on the other hand, features a structure that feeds the values to the next layer. The first convolution layer consists of a 7x7 filter. After the 3x3 maximum pooling layer that follows, it has a structure known as learning now, consisting of 3 blocks and convolution layers with 1x1, 3x3, 1x1 filter, respectively. The convolution layers of the VGG16 and ResNet50 models with the specified structures were connected to the classification layer designed in the first approach. For the classification process, 2 fully connected layers and classifier layers with sigmoid function was used. This method used in the study is shown in Figure7.
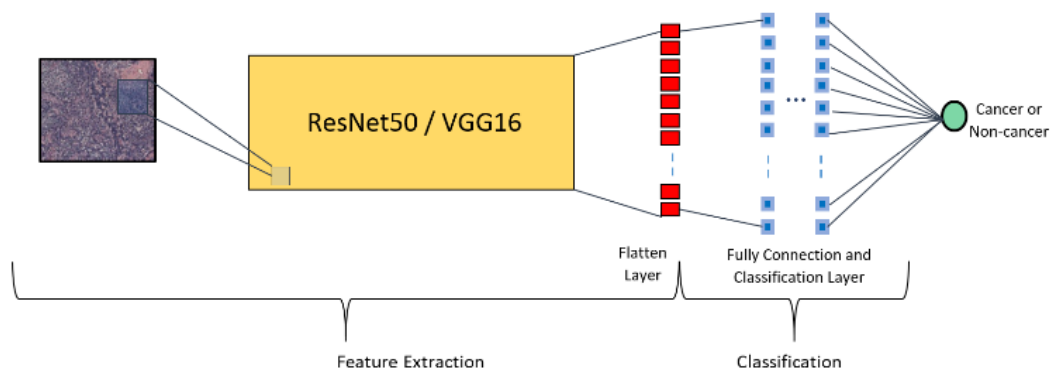

Figure 7. Combining the classification layer of CNN model created with the features obtained from VGG16 and ResNet50 models

3. Experimental Results

3 different approaches have been proposed to predict cancerous cells, and within each approach, models with different hyper-parameters have been developed. A series of extensive experiments were conducted to evaluate the performance of these models. This section is organized as follows:

- Description of our implementation details in all experiments conducted in this section.
- Each of the developed approaches was trained and tested with different hyperparameters.
- Performance analysis and comparison of the models included in the proposed approaches.
- Identifying the model that produces the best result.
- Performance analysis of the proposed model against recent studies done.

In this study, 3 different datasets containing labeled pathology images were used. Each dataset was randomly split, 70% for training, 15% for validation, and 15% for testing. The models trained sequentially with these datasets were compared using 2 different image scales in 32x32 and 48x48 sizes, 3 different optimization functions including Adam, SGD and RMSProp, and 3 different epochs as 25, 50, 100 values. In addition, the use of different classification algorithms within the approaches increased, the diversity in the experiments. In this way, when the determined hyper-parameters were applied to all the designed models, 324 different training-test processes were carried out. The results of the experiments carried out with 3 approaches and various hyper-parameters designed within the scope of the study are presented in the tables. Tables contain only the results of test data. These test results were evaluated using Accuracy, Precision, Recall, and F-score metrics. As seen in Table 2, training and testing processes were carried out with different hipper parameters using Dataset1. In experiments performed with this dataset, it was generally seen that more successful results were produced with the SGD function.

Table 2. Performance metrics of Approach1 and Approach2 on Dataset 1 containing 2352 data

Models	Hyper-parameters	Approach1		Approach2	
		NN (CNN)	RF	NB	SVM
FEM 1	Opt. Func.	SGD	SGD	Adam	SGD
	Image size	32x32	32x32	32x32	32x32
	Epoch	50	50	100	100
	Accuracy	0.817	0.690	0.790	0.815
	Precision	0.840	0.74	0.850	0.830
	Recall	0.820	0.690	0.790	0.820
	F-Score	0.810	0.670	0.780	0.810
FEM 2	Opt. Func.	SGD	SGD	SGD	SGD
	Image size	32x32	48X48	48X48	48X48
	Epoch	100	50	25	50
	Accuracy	0.818	0.656	0.741	0.803
	Precision	0.840	0.800	0.830	0.810
	Recall	0.820	0.660	0.740	0.800
	F-Score	0.820	0.610	0.720	0.800
FEM 3	Opt. Func.	SGD	Adam	SGD	SGD
	Image size	32x32	32x32	32x32	32x32
	Epoch	100	25	100	100
	Accuracy	0.812	0.630	0.781	0.821↗
	Precision	0.840	0.73	0.850	0.850
	Recall	0.810	0.63	0.780	0.820
	F-Score	0.810	0.59	0.770	0.820
FEM 4	Opt. Func.	RMSProp	Adam	RMSProp	RMSProp
	Image size	48x48	32x32	48x48	48x48
	Epoch	25	100	25	25
	Accuracy	0.798	0.683	0.744	0.789
	Precision	0.860	0.780	0.830	0.850
	Recall	0.800	0.690	0.740	0.790
	F-Score	0.790	0.610	0.730	0.780

The symbol (↗) in the tables represents the highest accuracy value in each approach.

Training and testing processes were carried out with different hipperparameters using Dataset2, as shown in Table 3 and Dataset3, as shown in Table 4. Experiments conducted with both datasets generally produce more successful results with the Adam and RMSProp optimization function. The best results produced by the experiments carried out with the third approach are shown in Table 5. Models developed with this approach were trained and tested with all datasets, similar to the other approaches.

Table 3. Performance metrics of Approach1 and Approach2 on Dataset 2 containing 5646 data

Models	Hyper-parameters	Approach1		Approach2	
		NN (CNN)	RF	NB	SVM
FEM 1	Opt. Func.	RMSProp	RMSProp	Adam	Adam
	Image size	48x48	48x48	32x32	48x48
	Epoch	50	50	100	25
	Accuracy	0.827	0.690	0.820	0.838
	Precision	0.86	0.700	0.860	0.870
	Recall	0.83	0.650	0.820	0.840
	F-Score	0.82	0.670	0.820	0.830
FEM 2	Opt. Func.	Adam	Adam	Adam	SGD
	Image size	32x32	48X48	32x32	48X48
	Epoch	25	50	50	25
	Accuracy	0.810	0.626	0.741	0.794
	Precision	0.750	0.680	0.660	0.850
	Recall	0.940	0.630	0.740	0.790
	F-Score	0.830	0.600	0.790	0.790
FEM 3	Opt. Func.	Adam	SGD	Adam	SGD
	Image size	32x32	32x32	32x32	32x32
	Epoch	25	100	25	50
	Accuracy	0.780	0.652	0.730	0.810
	Precision	0.850	0.700	0.790	0.860
	Recall	0.780	0.650	0.730	0.810
	F-Score	0.770	0.630	0.710	0.800
FEM 4	Opt. Func.	RMSProp	Adam	RMSProp	RMSProp
	Image size	32x32	48x48	32x32	32x32
	Epoch	25	25	25	25
	Accuracy	0.843	0.780	0.828	0.849↗
	Precision	0.870	0.810	0.870	0.860
	Recall	0.840	0.780	0.830	0.850
	F-Score	0.840	0.780	0.820	0.850

Table 4. Performance metrics of Approach1 and Approach2 on Dataset 3 containing 12140 data

Models	Hyper-parameters	Approach1		Approach2	
		NN (CNN)	RF	NB	SVM
FEM 1	Opt. Func.	SGD	Adam	Adam	Adam
	Image size	48x48	32x32	32x32	32x32
	Epoch	50	50	25	25
	Accuracy	0.842	0.717	0.840	0.846
	Precision	0.870	0.720	0.870	0.870
	Recall	0.840	0.720	0.840	0.850
	F-Score	0.840	0.720	0.840	0.840
FEM 2	Opt. Func.	RMSProp	Adam	Adam	Adam
	Image size	48x48	32x32	32x32	32x32
	Epoch	25	25	25	25
	Accuracy	0.841	0.702	0.839	0.806
	Precision	0.860	0.640	0.770	0.850
	Recall	0.840	0.620	0.970	0.810
	F-Score	0.840	0.760	0.860	0.800
FEM 3	Opt. Func.	Adam	Adam	Adam	Adam
	Image size	32x32	32x32	32x32	32x32
	Epoch	50	50	50	50
	Accuracy	0.870↗	0.747	0.851	0.859
	Precision	0.870	0.750	0.850	0.860
	Recall	0.870	0.750	0.850	0.860
	F-Score	0.870	0.750	0.850	0.860
FEM 4	Opt. Func.	Adam	RMSProp	RMSProp	RMSProp
	Image size	32x32	32x32	32x32	32x32
	Epoch	50	25	50	25
	Accuracy	0.854	0.834	0.829	0.845
	Precision	0.850	0.850	0.850	0.850
	Recall	0.850	0.830	0.830	0.800
	F-Score	0.850	0.830	0.830	0.790

Table 5. Performance metrics of Approach3 on all Dataset

Models	Hyper-parameters	Approach3		
		2352	**5646**	**12140**
ResNET50	Opt. Func.	Adam	Adam	SGD
	Image size	48x48	32x32	32x32
	Epoch	100	25	100
	Accuracy	0.786	0.794	0.819↗
	Precision	0.830	0.830	0.860
	Recall	0.790	0.790	0.820
	F-Score	0.77	0.790	0.810
VGG16	Opt. Func.	RMSProp	RMSProp	SGD
	Image size	48x48	48x48	32x32
	Epoch	100	100	25
	Accuracy	0.815↗	0.812↗	0.791
	Precision	0.840	0.830	0.840
	Recall	0.820	0.810	0.790
	F-Score	0.810	0.810	0.780

When the experimental results obtained are examined:

- The experimental results obtained in the third approach were lower compared to those in the first and second approaches. In contrast, there were common inferences.
- Considering all approaches, the best result was obtained with the dataset containing 12.140 pieces.
- In addition, it was seen that the decrease in image sizes did not adversely affect the test results. This situation, cost and time savings were ensured.
- Clear information could not be obtained with the optimization function and the number of epochs. However, when looking at the overall success, it can be stated that the Adam and RMSProp functions are more in the tables, so that good results can be produced with these optimization functions.

The best results produced by these 3 approaches were also evaluated in terms of other calculated metrics. Accordingly, it was determined that the Precision metric was around 85% for the best results in all 3 approaches. This value shows that models successfully perform accurate positive classifications. In other words, it demonstrates how reliable the data classified as cancerous cells is. When the Recall metric was examined, it was observed that the result produced in the first table was lower than the others. Although these results, which should not be underestimated, they indicate how many of the positive classes the model is able to capture accurately. While this situation is similar to the Precision metric, the Recall metric does not calculate whether negative samples are classified as positive incorrectly. When ensuring that an image is classified as positive, the results obtained from the Precision metric must be considered. The harmonic mean of these two metrics is obtained with the F-score metric. For the best Precision and Recall, the F-score metric is expected to converge to 1. It is seen that the values obtained in this study are very close to 1. The cost and running time of the models are as important as their performance. When looking at the vector lengths, which are one of the most important factors affecting these values, it is seen that the vector lengths produced by 4 different CNN / FEM (vector lengths are equal because of the same feature extraction) models vary by the image scale. This affects the different vector lengths in the convolution and image scale layers, as shown in Table 6. The vector length decreases as the image scale decreases and the number of layers increases. The reduction in vector length reduces model running time and saves time.

Table 6. Vector length in flattening layer of models in different convolution layers

Models	Conv. Layers	Image Scale	Vector lengths
1	3	32x32	1x50176
		48x48	1x123904
2	6	32x32	1x73728
		48x48	1x204800
3	9	32x32	1x1024
		48x48	1x4096
4	10	32x32	1x64
		48x48	1x576

Finally, these models were also evaluated in terms of AUC metrics. The ROC curve and AUC value of each approach that produces the best values for the 3 datasets are shown in Figure 8. Although the produced values were close to each other, the test results performed with the Dataset3, which contained synthetic and original data, were more successful. In general, it was observed that data augmentation in deep learning models positively affected the results of the model. As a result of the experiments, it was observed that the test performed with Dataset3 and Approach1 produced the most successful results. It was seen that with this model, cancerous cells were detected better than others. It was also understood through the result produced by the AUC metric that cancerous and non-cancerous cells were successfully separated from each other.

In the study, three different approaches were trained and tested with various hyper-parameters. The best result was obtained with the Approach1 trained with Dataset3. It was observed that the 32x32 scaled data used in this experiment and the Adam optimization function positively affected the results. Looking at the approaches in general, it was observed that data augmentation positively affected the performance and Approach1 performed better than other approaches. In addition, it was observed that the increase in the convolution layer in Approach1 decreased the vector size and thus saved the training time.
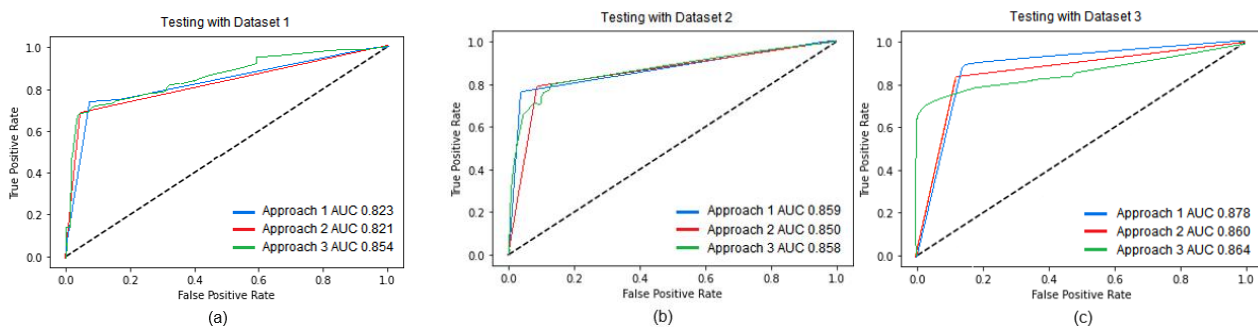
Figure 8. ROC curve and AUC value all approach. (a) for Dataset 1. (b) for Dataset 2. (c) for Dataset3

The results obtained from this study were compared with similar studies in the literature (Table 7). The dataset used in the study was previously developed by Duggal *et al.* (2017) using 2 different CNN models and they found that the stain layer they developed was close to the results obtained in this study. Oliveira and Dantas (2021) successfully predicted cancerous cells by improving the VGG16/VGG19 model with the same data set. However, in both studies, the quality of the images is quite high, which is far from reality. In our study, however, the image quality was deliberately reduced and the training-test success of a data in this situation was observed in different models.

Similarly, Perlberg and Kramer (2019) aimed to predict ALL cancer types with different datasets, and for this, they performed an effective classification using the Squeeze-and-Excitation module and the ResNeXt convolutional neural network. Although this module was not used in our study, an F1-score of 87.0% was obtained. This result is very close to the work of Perlberg and Kramer, although no extra optimization was used. When looking at other studies focusing on the same disease but using different datasets, it was seen that in Selçuk and Özen (2015), an accuracy rate of 93.91% was reached as a result of the training and testing process done with the SVM algorithm. Sipes *et al.* (2018) showed that by comparing the performance of different algorithms, as in this study, the most successful result was obtained with the CNN model. In Kumar *et al.* (2018), the performance of K-NN and NB models were compared and the K-NN algorithm was found to be more successful. In Bhuiyan *et al.* (2019), RF, SVM, LR, and DT models were compared and it was found that SVM algorithm was more successful. Additionlly, in another study by Banik *et al,* (2019), it was shown that augmented data yielded successful results, as in this study.

Table 7. Classification of pathology data to the leukemia cancer

Studies	Work	ML Model	Success Rate *(Accuracy)*
Selcuk & Ozen	Classification of ALL cells	SVM	%93,91
Sipes & Li	Classification of ALL cancer and comparison of performance	KNN NN CNN	%81(±5) %85(±3) %92(±3)
Kumar *et al.*	Categorization of ALL data extracted by image segmentation with K-NN and NB	KNN NB	%92,8 %78-79
Bhuiyan *et al.*	Classification performance comparison of ALL blood cells with RF, SVM, LR and DT algorithms	RF SVM LR DT	%98 %99,05 %97,18 %98,18
Duggal *et al.*	Increasing the dataset of ALL cell and adding SD layer that determines the pixel density of AlexNet and T-CNN models and comparing the success rate	AlexNet T-CNN AlexNet+SD T-CNN+SD (SVD)	%87,9 %92,4 %88 %95,5
Zhao *et al.*	Automatic recognition system for automatic detection and classification of WBCs	CNN	-
Jiang *et al.*	Classification of CLL, FL and MCL lymph types	SVM	%97,96
Banik *et al.*	Classification of 4 different white blood cells with 8-layer CNN model	CNN	Neutrophil 90,55 Eosinophil 96,12 Monocyte 95,33 Lymphocyte 99,8
Perlberg & Kramer	ResNeXt to differentiate ALL-type cancerous cells from non-cancerous cells and used the Squeeze-and-Excitation modüle with same dataset	ResNeXt	-
Oliveira & Dantas	Classified cancerous and non-cancerous cells using the VGG16/VGG19 model with same dataset	VGG16 VGG19	%92,48 %91,59
Varol Malkocoglu & İseri	Classification and performance comparison of ALL cells using 3 approaches and different parameters	Approach1 (CNN-3)	Accuracy %87,0 and AUC: %87,8 (with Dataset3)

The contributions of this study to the literature are as follows:

- The performance of models with different hyper-parameters on low-quality data was analyzed.
- The classification of cancerous cells by the proposed Approach2 has not been previously observed in the literature. The performance of this approach was examined and compared with others, and it was observed that it could produce remarkable results for classification processes in health and similar fields.
- The effect of data sets of different sizes on the models was examined. Accordingly, the data set with synthetic and original data produced more successful results than the others.
- It has been observed that our model performs well even on low-quality datasets. In low-quality datasets, the image scale is small, thus reducing the vector length. The reduction in vector length reduces model run time and saves time.
- Comparing with other studies were made, similarities differences are discussed.

4. Conclusion

In this study, applied research was conducted on the classification of digital pathology images related to ALL disease with deep learning, machine learning, and transfer learning approaches. Experimental studies were carried out on the Google Colab platform using a digital pathology images that contain patient information. These studies were performed on a dataset consisting of 10,661 previously labeled ALL images. After making the selected dataset more balanced compared to the lesser class (3389 non-cancerous), a new dataset containing 12.140 data was created by applying rotation to the images. In addition to this dataset, two more clusters consisting of 2352 and 5646 data were created with only the original data. 70% of each dataset was devoted to training, 15% to validation, and 15% to testing. On the created datasets, two different image scales with 32x32 and 48x48 dimensions were resized. Batch size value was determined as 32 and iteration values were determined as 25, 50, and 100. In addition, 4 different models were obtained by gradually increasing the number of layers. These models had 3, 6, 9, and 10 convolution layers, respectively. The ReLU activation function was used in hidden and fully connected layers, while the Sigmoid activation function was used in the output layer. The two fully connected layers used in the model consisted of 256 and 128 neurons, respectively.

When the results obtained with these developed approaches were examined, it was determined that using the dataset consisting of both synthetic and real data affected the success of the model positively. It was observed that the increase in convolution layers used for feature extraction and the Adam or RMSProp optimization functions used during the training of models positively affected the success of the model. The results produced by both the first approach and the second approach were quite close to each other and significant. However, when looking at the third approach in which the transfer learning method was used, it was found that the performance values were lower than the other two approaches.

When we look at similar studies in the literature, it is seen that different models and datasets are used. Although the models in this study were trained with low resolution pathology images, the model results were found to perform as well as the existing studies. This shows that similar performance can be achieved by designing more compact models. In this way, successful systems can be designed to save costs and profit from speed.

Consequently, when the performance values of the artificial intelligence approaches are taken into consideration, it can be said that, in terms of digital pathology, artificial intelligence-based methods can be used at the pathology departments of hospitals especially in the diagnosis phase. The use of a system to assist pathologists during cancer diagnosis is crucial in terms of time, diagnostic accuracy and workload. With this model, human error can be reduced and time can be saved, reducing the workload of pathologists. Additionally, this study measured the effect of changes in hyperparameters on the performance of the model. It is believed that the results of this evaluation will guide researchers in their hyperparameter choices. In the future, by using the holistic form of pathology data, regions of interest (RoI) will be detected and classified with our best approach.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Anonymous. (2023, Nov. 10). Diagram showing the cell that ALL starts [Online]. Available:https://en.wikipedia.org/wiki/Acute_lymphoblastic_leukemia

Banik, P. P., Saha, R. and Kim, K. D., 2019. Fused Convolutional Neural Network for White Blood Cell Image Classification, In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 238-240 , Okinawa, Japan.

Bhuiyan, M. N. Q., Rahut, S. K., Tanvir, R. A. and Ripon, S., 2019. Automatic Acute Lymphoblastic Leukemia Detection and Comparative Analysis from Images In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 1144-1149, Paris, France, 2019.

Clark, K. , Vendt, B. , Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. and Prior, F., 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging, vol. 26, no. 6, pp. 1045-1057.

Duggal, R., Gupta, A. & Gupta, R., 2016b. Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks, CME Series on Hemato - Oncopathology, All India Institute of Medical Sciences (AIIMS), New Delhi, India.

Duggal, R., Gupta, A., Gupta, R. and Mallick, P., 2017. SD-layer: stain deconvolutional layer for CNNs in medical microscopic imaging, In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 435-443, QC, Canada.

Duggal, R., Gupta, A., Gupta, R., Wadhwa, M. & Ahuja, C., 2016a. Overlapping cell nuclei segmentation in microscopic images using deep belief networks, In Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, Guwahati, India.

Global Cancer Observation. (2024, Jan. 29). International Agency for Research on Cancer [Online]. Available:https://gco.iarc.fr/today/en/dataviz/pie?mode=cancer&group_populations=1&age_end=2

Heffner, S., Colgan, O. and Doolan, C. Digital Patology, [Online]. Available: https://www.leicabiosystems.com/pathologyleaders/digital-pathology/, Accessed on: April 5, 2021

Hu, J., Shen, L. & Sun, G., 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141.

Inaba, H., Greaves M. & Mullighan, C. G., 2013. Acute lymphoblastic leukaemia. The Lancet, vol. 381, no. 9881, pp. 1943-1955.

Jiang, H., Li, Z., Li, S. and Zhou, F., 2018. An Effective Multi-Classification Method for NHL Pathological Images, In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 763-768, Miyazaki, Japan.

Karlik, B., and Olgac, A. V., 2011. Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems, 1(4), 111-122.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kumar, S., Mishra, S. & Asthana, P., 2018. Advances in Computer and Computational Sciences (second edition), Springer, Singapore, pp. 655-670.

Madabhushi, A. (2009) Digital pathology image analysis: opportunities and challenges. Imaging in medicine, vol. 1, no.1, pp. 7-10.

Mourya S., Kant S., Kumar P., Gupta A., Gupta R. (2023, May. 5). ALL Challenge dataset of ISBI 2019 [Data set]. The Cancer Imaging Archive. [Online]. Available: https://www.cancerimagingarchive.net/collection/c-nmc-2019/

Oliveira, J. E. M. & Dantas, D. O., 2021. Classification of Normal versus Leukemic Cells with Data Augmentation and Convolutional Neural Networks. In VISIGRAPP (4: VISAPP), pp. 685-692.

PDQ® Pediatric Treatment Editorial Board PDQ Childhood Acute Lymphoblastic Leukemia Treatment. Bethesda, MD: National Cancer Institute, [Online]. Available: https://www.cancer.gov/types/leukemia/patient/child-all-treatment-pdq. , Accessed on: April. 10, 2021

Prellberg, J. & Kramer, O., 2019. Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging, pp. 53-61, Springer, Singapore.

Selçuk, O. and Özen, F., 2015. Acute lymphoblastic leukemia diagnosis using image processing techniques. In 2015 23rd Signal Processing and Communications Applications Conference (SIU), pp. 803-806, Malatya, Turkey.

Si, J., Harris, S. L. & Yfantis, E., 2018. A Dynamic ReLU on Neural Network. In 2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS), pp. 1-6, IEEE.

Sibi, P., Jones, S. A. and Siddarth, P., 2013. Analysis of different activation functions using back propagation neural networks. Journal of Theoretical and Applied Information Technology, 47(3), 1264-1268.

Sipes, R. and Li, D., 2018. Using Convolutional Neural Networks for Automated Fine-Grained Image Classification of Acute Lymphoblastic Leukemia. In 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA), pp. 157-161, Hong Kong, China.

Sipes, R. and Li, D., 2018. Using Convolutional Neural Networks for Automated Fine-Grained Image Classification of Acute Lymphoblastic Leukemia, In 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA), pp. 157-161, Hong Kong, China.

Stursa D. and Dolezel, P., 2019. Comparison of ReLU and linear saturated activation functions in neural network for universal approximation. In 2019 22nd International Conference on Process Control (PC19), pp. 146-151, IEEE.

Uzunhan, T. A. and Karakaş, Z., 2012. Çocukluk Çağı Akut Lenfoblastik Lösemisi. Çocuk Dergisi, vol. 12, no. 1, pp. 6-15.

Yazan, E. and Talu, M. F., 2017. Comparison of the stochastic gradient descent-based optimization techniques. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). 1-5. IEEE.

Yöntem, A. & Bayram, İ., 2018. Çocukluk Çağında Akut Lenfoblastik Lösemi. Arşiv Kaynak Tarama Dergisi, vol. 27, no. 4, pp. 483-499.

Zhao, J., Zhang, M., Zhou, Z., Chu, J. and Cao, F. (2017) Automatic Detection and Classification of Leukocytes Using Convolutional Neural Networks, Medical & Biological Engineering & Computing, vol. 55, no. 8, pp. 1287-1301