Computer Systems Science & Engineering
DOI:10.32604/csse.2022.019333
Article

Tech Science Press

Classification and Diagnosis of Lymphoma's Histopathological Images Using Transfer Learning

Schahrazad Soltane*, Sameer Alsharif and Salwa M.Serag Eldin

College of Computers and Information Technology, Computer Engineering Department, Taif University, Taif, Kingdom of Saudi Arabia
*Corresponding Author: Schahrazad Soltane. Email: sch.soltane@gmail.com

Received: 10 April 2021; Accepted: 14 May 2021

Abstract: Current cancer diagnosis procedure requires expert knowledge and is time-consuming, which raises the need to build an accurate diagnosis support system for lymphoma identification and classification. Many studies have shown promising results using Machine Learning and, recently, Deep Learning to detect malignancy in cancer cells. However, the diversity and complexity of the morphological structure of lymphoma make it a challenging classification problem. In literature, many attempts were made to classify up to four simple types of lymphoma. This paper presents an approach using a reliable model capable of diagnosing seven different categories of rare and aggressive lymphoma. These Lymphoma types are Classical Hodgkin Lymphoma, Nodular Lymphoma Predominant, Burkitt Lymphoma, Follicular Lymphoma, Mantle Lymphoma, Large B-Cell Lymphoma, and T-Cell Lymphoma. Our proposed approach uses Residual Neural Networks, ResNet50, with a Transfer Learning for lymphoma's detection and classification. The model used results are validated according to the performance evaluation metrics: Accuracy, precision, recall, F-score, and kappa score for the seven multi-classes. Our algorithms are tested, and the results are validated on 323 images of 224 × 224 pixels resolution. The results are promising and show that our used model can classify and predict the correct lymphoma subtype with an accuracy of 91.6%.

Keywords: Classification; confusion matrices; deep learning; k-fold cross validation; lymphoma diagnosis; residual neural network; transfer learning

1 Introduction

Lymphoma is one of the mortal cancerous diseases that is formed by abnormal cell mutation in the immune system. Many different histological subtypes exist, whose diagnosis is typically based on sampling (biopsy) [1]. According to the American Cancer Society [2], the expected new lymphoma cases will top 85,720 cases in 2021, while the death rate will jump to 21,000 cases in both sexes. Compared to other types of cancer, lymphoma records a moderate rate of incidence and death. However, those numbers are still worrying as the diversity and development of lymphoma types surge in the shadow of diagnosis difficulties. According to Leukemia & Lymphoma Society [3], more than 60 subtypes of lymphoma are classified as Hodgkin and non-Hodgkin lymphoma based on their morphological and biological structures. Those types are originating from T-cells or B-cells forming T or B lymphocytes, respectively, where the latter is responsible for 85% of total cases. Histologically, this breakdown, along with various features of each subtype, requires an expert specialist to be determined. However, similar patterns and some morphological structures are difficult to be distinguished by the naked eye [4]. Therefore, Machine Learning (ML) could be exploited to increase diagnosis precision based on extracting precise features from digitized images.

The digitization of histological images is prepared from stained microscopic tissue slices in different steps. First, microtomical techniques are used to slice tissue blocks into different thickness tissue slides for the purposes of analysis and classification [5]. Next, samples are commonly, stained by Hematoxylin and Eosin (H&E), which both stain different parts of the cell with a different color for computation and visualization. Finally, a microscopic embedded camera is used for digitizing histological slides. Indeed, the revolution of modern digital computers opens the door for the rapid development of digital image processing. Therefore, different computer algorithms are used to extract valuable information from such images. Moreover, the digital image can be segmented to indicate interesting regions such as tumors or cancerous cell formation. Advanced Machine Learning algorithms like Support Vector Machine (SVM), Naïve Bayes, or Deep Learning algorithms are used in digital images and computer vision applications to automatically detect and classify significant patterns based on extracting discriminant features.

Feature extraction is a technique applied to the digital image to find the most relevant information that can be the training set for pattern recognition and classification. The digital image contains scattered and redundant data challenging to recognize. The data is reduced to the most discriminant feature vector. Indeed, there are many methodologies that can be used to extract such features, and mostly statistical tools such as mean, standard deviation, and random numbers used to find the most likelihood gathering of data that can be considered as valuable features. In biomedical digital images like digitized lymphoma slides, each type can be recognized morphologically by patterns that could be taken as features for recognition and classification purposes.

For instance, Follicular lymphoma (FL) and Mantle cell lymphoma (MCL) are both non-Hodgkin subtypes that share some standard biological features such as cellular cleaved nuclei and a tiny amount of cytoplasm. However, FL is characterized by centroblast, which is enlarged B-cells rapidly spread in the germinal centers, which is the absent feature in MCL [6]. Hence, centroblast is a discriminant morphological feature that can be observed when the tissues get stained by H&E. However, these lymphomas are rare and are not fully characterized; MCL is characterized by an aggressive clinical evolution with few survivors in a long time. The diagnosis of a large B-cell lymphoma LBCL and classic Hodgkin lymphoma CHL is often straightforward. However, in some cases, these simple diagnoses can be quite complex. The diagnostic difficulty may be due to the evaluation of their morphologic and immunophenotypic features. The more complex cases of follicular lymphoma versus LBCL, or regrouping more than one lymphoma type such as peripheral T-cell lymphomas PTCL with large B-cell proliferations or PTCL with Hodgkin/Reed-Sternberg-like cells. More efforts from an expert are needed to help the detection of such diseases.

The main contribution of this paper is to classify seven subtypes of lymphoma, Non-Hodgkin (NHK) and Hodgkin (HK) Lymphoma, which are HK-Classical Hodgkin Lymphoma, HK-NLP Nodular Predominant, NHK-BcellBurkitt, NHK-BcellFollecular, NHK-BcellMantel, NHK-LargeBcell, and NHK-Tcell. These subtypes are aggressive, rare, silent-growing, and morphologically similar lymphomas [7], which add complexity factors to the classification process. Therefore, we used the transfer learning technique with RestNet50 as a pre-trained model, and the results show that this approach was able to classify up to 91.6% of testing images.

The rest of this paper is organized as follows: Related work and recent techniques applied in the domain are discussed in Section 2. The proposed methodology and material setup are given in Section 3. Section 4 presents results and discussion. Finally, the paper and most relevant results are concluded in Section 5.

2 Related Work

Lymphomas, the most common type of cancers, have more than 60 subtypes. For decades, cancer diagnosis procedure has required an expert person to identify any abnormalities in color, texture, or morphology. The human mistake of identifying cancer cells is highly probable, and the process of studying each sample is time-consuming. Therefore, the need to build an accurate and automated diagnosis support system for lymphoma identification is essential [8]. The revolutionary of digital computers opens the door for research works to rapidly develop complex computational algorithms such as a Deep Learning (DL) using the Convolutional Neural Networks (CNN) to help to solve different biomedical problems.

The CNN shows remarkable progress in image classification tasks and diverts the compass of current image classification research toward such competitive methodology [9–11]. A review study that focuses on applying CNNs to image classification tasks is presented in Rizwan et al. [12]. The study covers their development from their predecessors to recent works in deep learning techniques by reviewing the contributions and challenges of over 300 publications. A study about recent advances in image processing techniques in classifying NHL was presented in Azevedo Tosta et al. [13]. Authors gave the most used segmentation techniques such as thresholding, region-based methods, and K-means clustering algorithm. Besides, it includes reviewing validation techniques and potential directions of research in the segmentation of this neoplasia. Another survey on researches published between 2010 and 2017 on the detection of NHK-L types is reviewed and compared in Battula et al. [14]. Many image-based algorithms have been discussed to find any abnormal changes in the images, such as the change in color, shape, size, texture, etc. Color deconvolution is the main technique used in preprocessing, and different classifiers are used in predicting the correct type of lymphoma. Nascimento et al. [15] presented computer-aided biopsy analysis support system of NHK-L classification. The approach is used for the classification of information extracted from morphological and non-morphological features of nuclei of lymphoma images. It achieved higher AUC and ACC values for all cases, the obtained rates were between 95% and 100%.

A multiple instance learning setting was exploited in Lippi et al. [16] where support vector machines and random forests are used as classifiers at both single VOIs (instances) and patients (bags) levels. Results were presented on two datasets comprising patients that suffer from four different types of malignant lymphomas, namely diffuse large B cell lymphoma, follicular lymphoma, Hodgkin’s lymphoma, and mantle cell lymphoma. The study indicated that texture analysis features extracted from positron emission tomography, combined with multiple-instance machine learning algorithms, can be discriminating for different malignant lymphomas subtypes. They achieve 97% for recall and 94% for the precision of analyzing 60 patients. Orlov et al. [17] classify three subclasses of lymphoma: SLL, Follicular lymphoma, and Mantle lymphoma. They use spectral analysis with a weighted neighbor distance algorithm. The model obtains high accuracy of 99%, but only 30 lymphoma images were used. Recent studies using Deep Learning have shown promising results. Brancati et al. [18] used ResNet27 model in the classification of three NH lymphoma subtypes. The study was done by adopting a residual convolutional 27 layers neural network. Their performances have been evaluated on the public datasets of digital histological images and have been compared with those obtained by using different deep neural networks (UNet and ResNet). The experimental results show an improvement of 5.06% based on the F-score for the detection task and an improvement of 1.09% inaccuracy measure for the classification task. A fully automated system based on Deep NN is presented in Tambe et al. [19]; such method predicts the class label of 75 images in the given test dataset. El Hachi et al. [20] used Deep Learning with a convolutional neural network (CNN) algorithm to build a lymphoma diagnostic model for three categories of diseases, diffuse large B-cell lymphoma, Burkitt lymphoma, and small lymphocytic lymphoma and Benin case. The test results showed excellent diagnostic accuracy at 95% using 240 images for the test set.

Other studies concentrated on predicting certain types of lymphoma subtypes such FL as it grows silently and is usually diagnosed in its later stages. Somaratne et al. [21] proposed a deep learning model that uses transfer learning with fine-tuning to improve the identification of FL on images from new sites that are different from those used during training. The proposed approach improves the prediction accuracy from 12% to 52% compared to the initial prediction of the model for the same testing images.

With respect to the previous research works, we used transfer learning with RestNet50 pretrained model to classify seven lymphoma subtypes for diagnosis purposes with obtained accuracy of 91.6%. Types were chosen according to their silent growing and difficulty of their diagnoses. Under some circumstances, detection and classification of B-cell or CHL lymphomas might be quite complex due to uncertainty in the evaluation of morphological and immune-phenotypical features along with biological continuum. Moreover, some lymphomas such as Composite Follicular and Mantle Cell are rare and not fully characterized, more efforts are needed to help to detect such diseases and save a life for many people [22].

3 Materials and Method

3.1 Lymphoma Image Dataset (L.I.D)

In our proposed diagnosis system shown in Fig. 7, we built Lymphoma Image Dataset (L.I.D) with RGB images of 224 × 224 pixels size to match the ResNet50 image size. Our Lymphoma Image Dataset is composed of 323 color images. This dataset is divided into seven different categories of lymphoma characterized by color and textured features. Two classes, Hodgkin Lymphoma (HK-L) and Non-Hodgkin Lymphoma (NHK-L).

The Hodgkin Lymphoma class has two subtypes; (1) Classical Hodgkin Lymphoma, where most cells are B lymphocyte and so-called Reed-Sternberg (RS) cells, the Nodular Lymphoma Predominant (HK-NLP), and (2) Non-Hodgkin Lymphoma classified as B or T cell Lymphoma. Five subtypes of them are selected, NHK-BcellBurkitt, NHK-BcellFollicular, NHK-BcellMantel, NHK-LargeBCell, and NHK-TCellLymph. Fig. 1 shows some images from each class. As shown in Fig. 2, the L.I.D has the same number of images for each class to build our classifier efficiency. Some images in our dataset are extracted from Refs. [23,24].

Our approach uses Residual Neural Network architecture (ResNet) with the Transfer Learning to classify complex textural images of lymphoma cancer. The implemented system trains five different models as lymphoma detectors on L.I.D dataset and selects the optimal model that shows the best classification results. This model is validated according to the performance metrics of confusion matrices. Finally, through the third stage of our work, network performance was evaluated against the used data set.

3.2 Training and Testing Protocols

The evaluation of the proposed ResNet model is done by using K-fold cross-validation method to assure that any biases in the dataset are detected [25]. This method helps to obtain a reliable estimation of the model's generalization error and gives predictions of how the model performs on unseen data. At the first step, data is randomly divided into training and testing sets. The data is split into K equal subsets. For each iteration, a different data are reserved for testing, and all the others are used for training the new classifier. A single subset is retrained as the validation data for testing the model. The process was repeated K times (iterations) and generated a model for each iteration. Cross-validation is used to pick the model that gives the best accuracy on various data partitions. Generally, averaged accuracies overall testing sets were taken for selection. Fig. 3 summarizes our proposed protocol of *K*-fold Cross-Validation following the 5F-CV Algorithm.

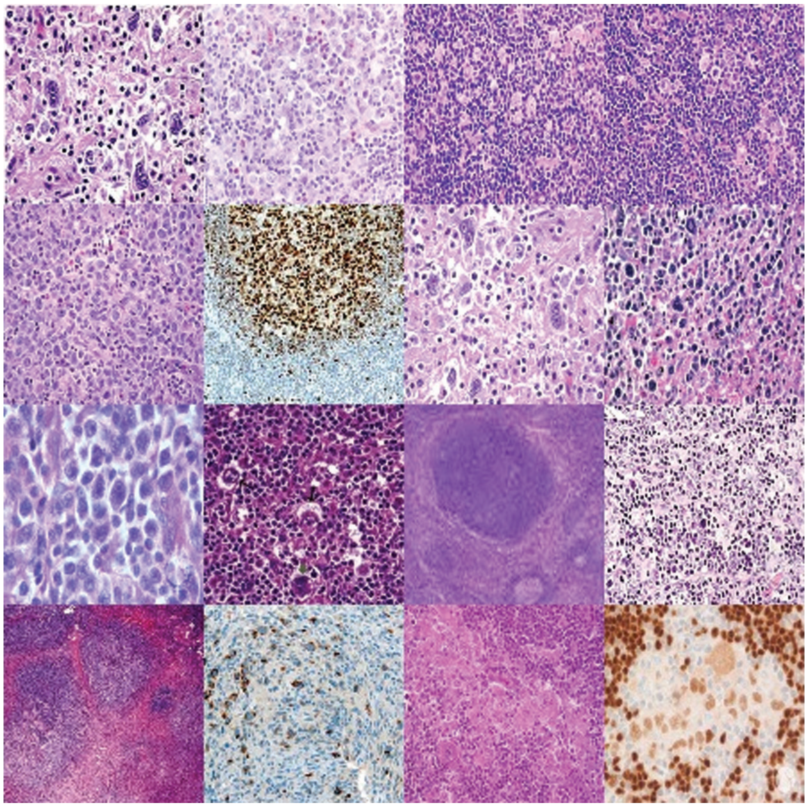


Figure 1: Representation of 16 RGB images 224 × 224 pixels size from Lymphoma Image Dataset (L.I.D)

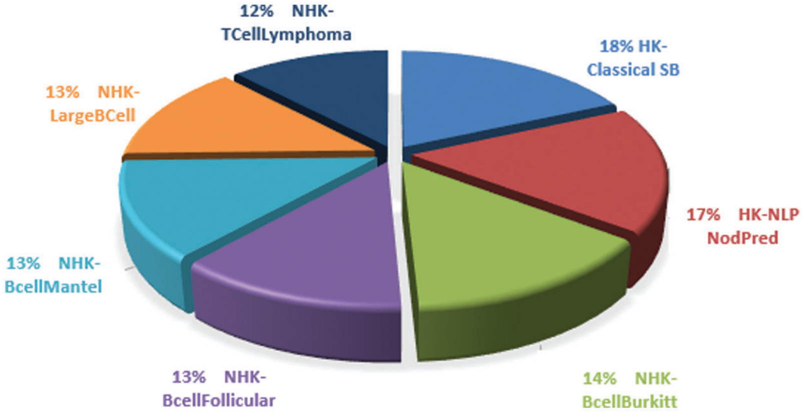
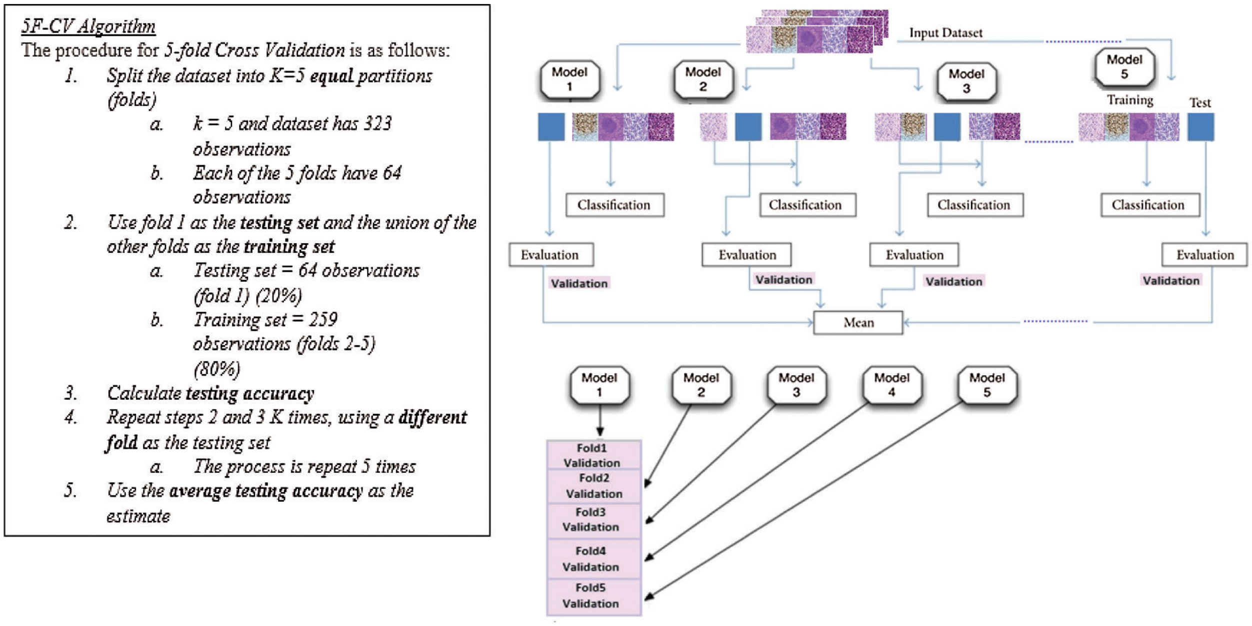


Figure 2: Distribution of labeled images: 323 total images divided by 7, about 50 images by class

The *K* value should be chosen carefully for the data sample. A poorly chosen value of *K* may result in a misrepresentative idea of the model skill. As *K* gets larger, the difference in size between the training set and the resampling subsets gets smaller. Typically, one performs *K*-fold C.V using $K = 5$ or $K = 10$. These values have empirically shown to yield test error rate estimates that suffer neither from excessively high bias nor very high variance. We tested both values; the best results were obtained with $K = 5$. This choice is due to the reduced number of data used.



5F-CV Algorithm

The procedure for *5-fold Cross Validation* is as follows:

1. *Split the dataset into K=5* ***equal*** *partitions (folds)*
    a. *k = 5 and dataset has 323 observations*
    b. *Each of the 5 folds have 64 observations*
2. *Use fold 1 as the* ***testing*** *set and the union of the other folds as the* ***training*** *set*
    a. *Testing set = 64 observations (fold 1) (20%)*
    b. *Training set = 259 observations (folds 2-5) (80%)*
3. *Calculate* ***testing accuracy***
4. *Repeat steps 2 and 3 K times, using a* ***different*** ***fold*** *as the testing set*
    a. *The process is repeat 5 times*
5. *Use the* ***average*** *testing* ***accuracy*** *as the estimate*

Figure 3: 5 Folds cross-validation for training and testing protocols

In the second step, our method used a transfer learning technique for feature extraction and classification. Residual Neural network architecture was adopted through our work using a ResNet50 deeper network with 50 layers. This network gave a significant performance in various medical imaging applications [26,27]. Its architecture achieved an optimal tradeoff between a speed and efficiency.

3.3 Residual Neural Network (ResNet)

The deep Residual Learning framework for image classification task supports several architectural configurations [28]. It allows achieving a suitable ratio between speed and quality of the model. This allows one to use deeper ResNet, which are faster to train and computationally less expensive than conventional CNNs and other architectures [29]. ResNet network converges faster compared to different architectures. However, the problem arises as the network depth increases resulting in fast decaying of accuracy. He et al. [30] solve this problem using the Deep Residual Learning framework with identity shortcuts. This identity shortcut (x) can be directly used when the input and output are of the same dimensions as represented in Eq. (1) and shown in Fig. 4.

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{1}$$

Designing the networks of Fig. 5 as follow:

1. *Use 1*1 or 3*3 filters mostly,* Tab. 1.
2. *Downsampling with CNN layers with stride 2, 50 layers for ResNet50,*
3. *Global average pooling layer and a 1000-way fully connected layer with Softmax in the end.*

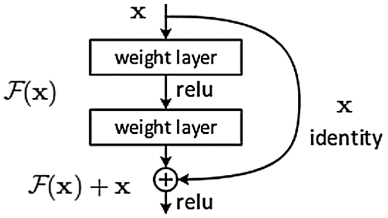

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 4: Residual block

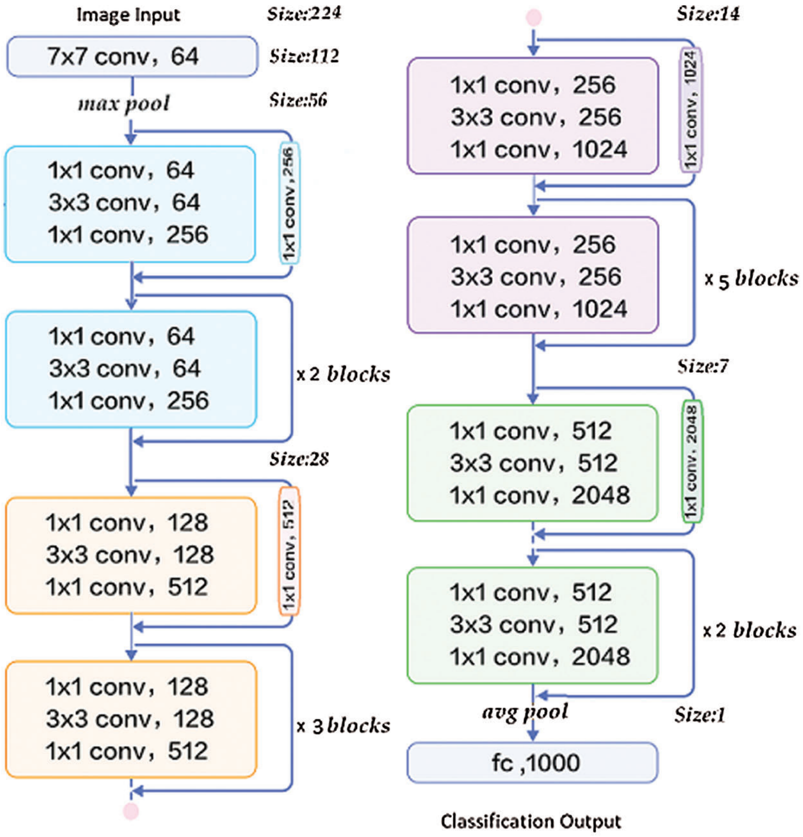


Figure 5: ResNet50 block diagram

Table 1: The standard confusion matrix (2 × 2)

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

True positives (TP) and true negatives (TN) are the correct predictions, while false negatives (FN) and false positives (FP) are the incorrect predictions (misclassified cases).

ResNet uses four modules made up of residual blocks. Each module uses blocks with the same number of output channels. The additional 1 × 1 convolutional layer is to change the stride or number of channels in Fig. 6. The network used a 7 × 7 convolutional layer with 64 output channels followed by 3 × 3 maximum pooling. A batch normalization layer is added after each convolutional layer. The ResNet50 network is designed as follows: The first module has 3 residual blocks, 4, 6, and 3 for the 2nd, 3rd, and 4th module, respectively. Each first residual block has 1 × 1 and 3 × 3 convolutional layers with the same number of output channels, and the number of output transformed in the last block, adding 1 × 1 convolution for the first module block as a shortcut. The resolution decreases while the channel number increases until the global average pooling aggregates all features. The global average-pooling layer extracted 2048 features, followed by the fully connected layer output and using the Softmax regression classifier as the final step of the process, as illustrated in Fig. 5.

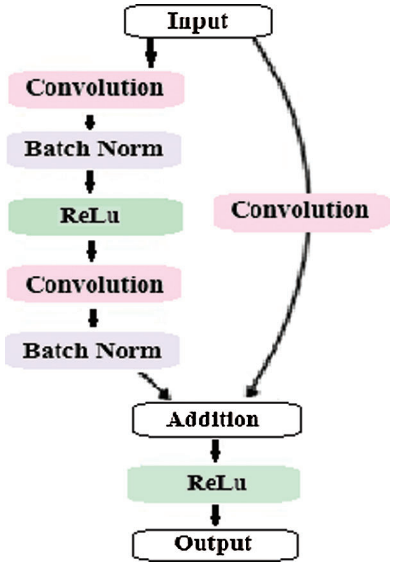
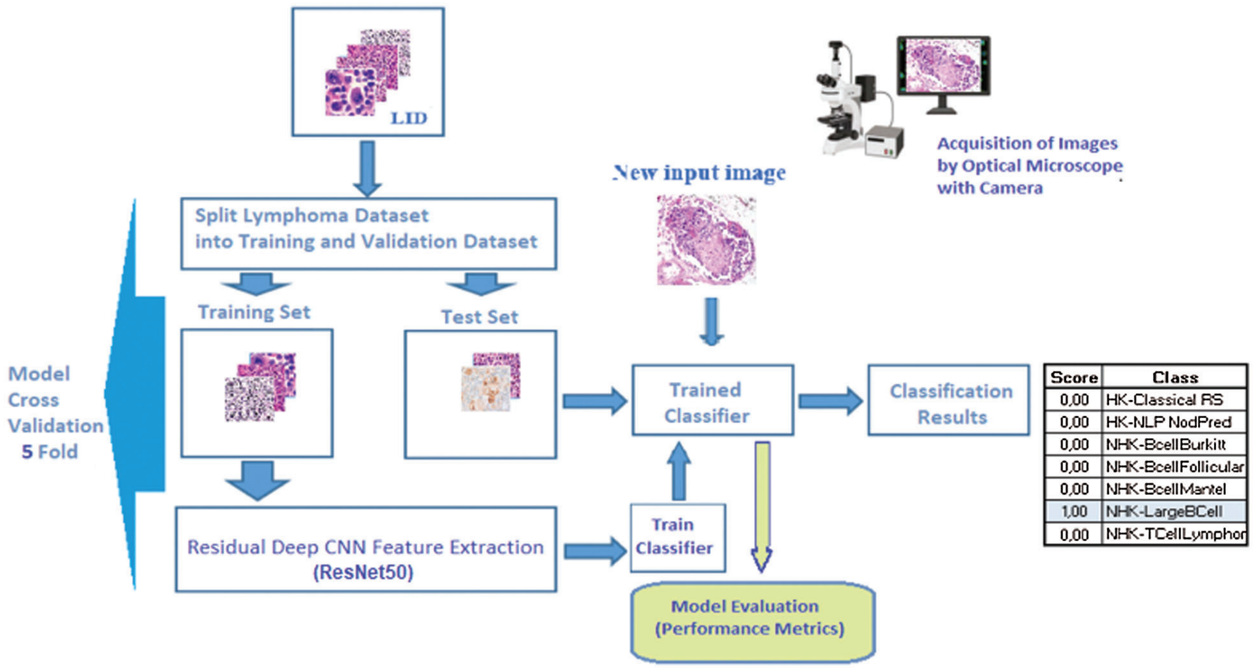


Figure 6: ResNet block with 1 × 1 convolution



Score	Class
0,00	HK-Classical RS
0,00	HK-NLP NodPred
0,00	NHK-BcellBurkitt
0,00	NHK-BcellFollicular
0,00	NHK-BcellMantel
1,00	NHK-LargeBCell
0,00	NHK-TCellLymphor

Figure 7: Global block diagram of our approach

3.4 Evaluation Metrics

The effectiveness of the proposed method was evaluated and validated using the confusion matrices and its performance metrics, as shown in Fig. 8. The confusion matrices tell how the currently selected classifier performs in each class and identifies the poorly performed areas. The classification model forecasts the type of each data instance, attributing to each observation its predicted label (positive or negative): thus, at the end of the classification, every observation corresponds in one of the following four cases [31]. A confusion matrix of size N × N associated with the classifier shows the predicted and actual classification. Tab. 1 shows a confusion matrix for binary classification N = 2. In our case, we use multi-classes (7 classes) where N = 7.

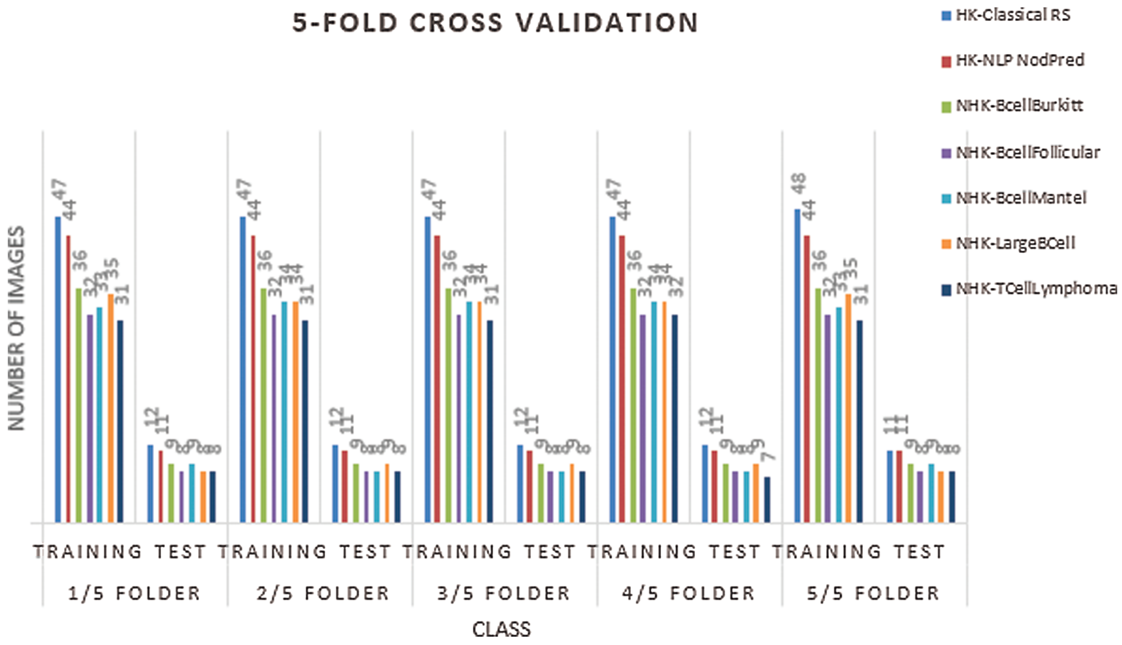


Figure 8: L.I.D Organization and categories distribution of training and test sets for each model

The trained network was tested on the test dataset and assessed quantitatively through four accuracy parameters extracted from the confusion matrix: Recall (True Positive/Sensitivity Rate), Precision (Positive Predictive Value), F-score (F1), and Kappa score, calculated in Tab. 2. Fig. 7 summarizes the global schema of our method.

4 Results and Discussion

All our experiments were developed in Matlab R2020a and trained using Intel Core i5-7200U CPU. In this section, we explore the results obtained using the pre-trained network, ResNet50, as a classifier to categorize an image into seven lymphoma diseases: Classical Hodgkin Lymphoma, Nodular Lymphoma Predominant, Burkitt Lymphoma, Follicular Lymphoma, Mantle Lymphoma, LargeBCell Lymphoma, and TCell Lymphoma.

To evaluate the model performance across the random dataset and assuring that any biases in the dataset will be detected, our approach uses a 5-fold cross-validation method (Fig. 7). It helps us to obtain reliable estimates of the model's generalization error and gives how well it performs on unseen data. The complete dataset is divided into five subsets. It is trained and tested five times. In each iteration, one subset was used to test the model, and the other four subsets were used for training the classifier. We assessed the average performance of the model by calculating the estimated error for each subset. 80% of the L.I.D are used for the training set, and 20% are used for the testing set, as shown in Fig. 8. The training and validation accuracy at the end of 40 epochs is given in Fig. 9. It illustrates the curves of model 3 (iteration3) for (1) training and (2) testing sets.

Table 2: Accuracy Metric Parameters [32–34]

Recall/Sensitivity Rate	$\frac{T_p}{T_p+F_N} \times 100\%$
Precision/Positive Prediction value	$\frac{T_p}{T_p+F_P} \times 100\%$
F-score	$2 \times \frac{Recall \times Precision}{Recall+Precision}$
Kappa-score	$\frac{Agree-Chance\ AG}{1-Chance\ AG}$ where : • Agree is: $\frac{\sum T_p}{\sum T_p+\sum F_N}$ • Chance AG is product of probability: $\left[\left(\frac{T_p+F_N}{\sum F_p+\sum F_N}\right) \times \left(\frac{T_p+F_P}{\sum F_p+\sum F_N}\right)\right]$ Total observations = $\sum F_p + \sum F_N$

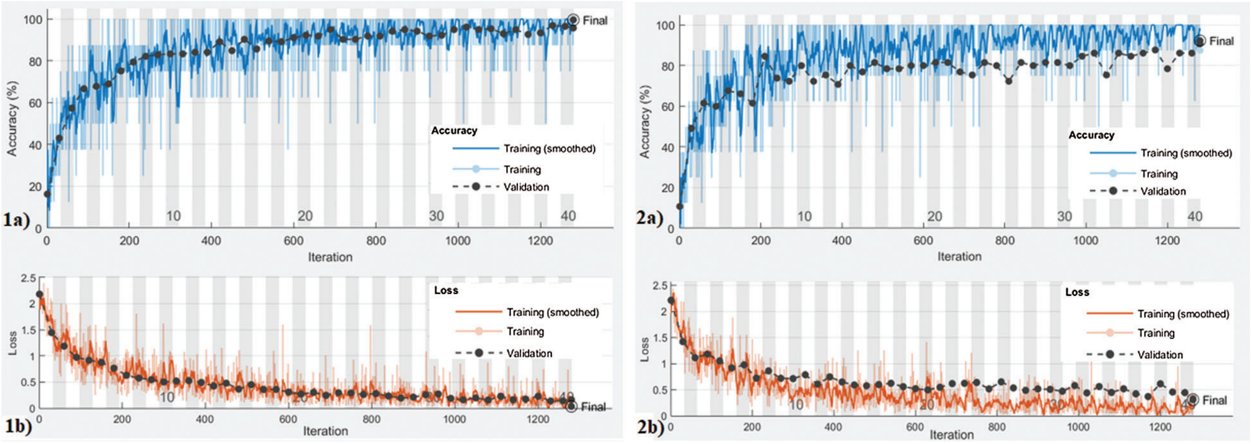


Figure 9: Training and Validation progress using ResNet50: 1a) Training accuracy curves, 2a) Validation accuracy curves with 40 epochs: case Model 3 (Fold1). 1b) Training loss of Model 3 (fold1) 176 mn time training, 2b) Validation loss

The classification results of the training and test sets for the 5 models are summarized in Tab. 3. We can see that, for each subset, the training classification accuracy was almost 99%. The results show that the network performs very well. The training accuracy is high for every model of ResNet50, with an averaging accuracy of 0.9923. It means the network is thriving in classifying 99.23% of the Images from all the training sets. For the test set, the minimum accuracy was noticed at iteration 4 (Fold4), with 89.36%. The maximum error (misclassifications) observed is 10.6%. The average classification accuracies were 91.61% without over-fitting the data. Our approach has been able to achieve the highest validation accuracy of about 92%. The key benefits of k-fold CV are the evaluation of the classification performance and elimination of over-fitting issues that occurred in ResNet models.

Table 3: Classification accuracy for the 5 Fold CV on the training and testing L.I.D dataset

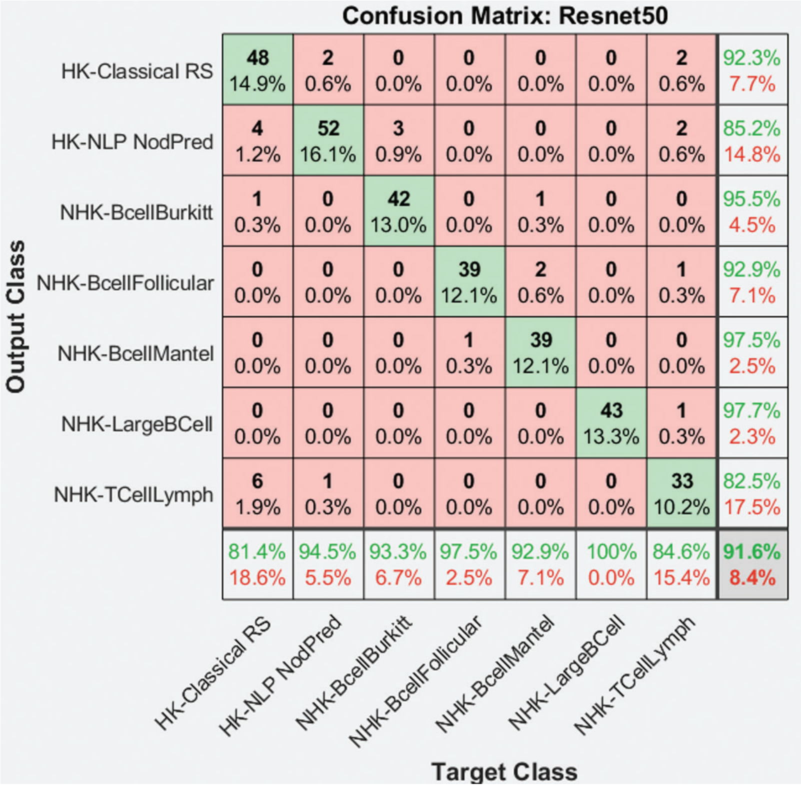
	Training	Validation	
	Tacc (%)	Vacc (%)	Error (%)
K = 1	99,61	92,31	7,69
K = 2	98,84	90,77	9,23
K = 3	99,22	93,79	6,21
K = 4	99,23	89,36	10,64
K = 5	99,23	91,82	8,18
Avg acc	99,23	91,61	8,39

Once we have trained our model, we want to see other metrics before concluding the usability used. Classification accuracy is typically not enough evidence to decide where your designed classifier is good enough to make robust predictions. The problem with accuracy is that it cannot discriminate among diverse kinds of misclassifications. Additionally, we computed other performance metrics such as Macro and Weighted of precision and recall, F-score (F1), and Kappa-score using the confusion matrix [34,35]. These terms are generally defined for multi-classification problems where the outcome is either positive or negative. As we have seven classes and dealing with the multi-class problem, we computed the four metrics while calculating TN, TP, FP, and FN of each class separately.

To create the confusion matrix, we have made the predictions over the test set. The actual class is taken directly from the original test dataset, whereas predicted class values are obtained from the classifier on the test dataset. The raw represents the true class labels, and the column is the predicted class. Along the first diagonal represent the correct classification, whereas all off-diagonal show misclassification (FP and FN). Fig. 10 illustrates the ResNet50 confusion matrix for the seven categories of lymphoma classes. It can be seen that the model has predicted 2 samples of class HK-NLP, 2 samples of NHK-TCell as wrongly predicted HK-Classical class, and 0 samples for other classes. Their classification was correct.

We have computed and summarize the results of TP, TN, FP, and FN of each of the seven lymphoma subtypes in Tab. 4. We can show that the model perfect for classifying NHK-LargeBCell with TP = 43, FP = 0, and FN = 1. The samples for this class are 43. This means no misclassification for this disease. We observe that the misclassifications are high for HK-Classical (TP = 48, FP = 11, and FN = 4). However, it gives better classification results for all other classes. A total of 27 misclassifications for 323 samples were noticed for the model. The rest of the 296 observations were correctly classified, with high accuracy of 91.6%.

The results obtained to compare the performance of the classifier shows that the approach performs exceptionally well. We observe that all precision values are superior to 0.93 (93%), with the highest score for NHK-BcellFollicular and NHK-LargeBcell, which have respectively, a precision of 0.98 (98%) and 1 (100%). The low results are observed for HK-Classical RS with 81.4% and NHK-TCell class with 85%. Combining the per-class precision into a single number, the average precision for multi-classes (seven classes) is called Macro-Precision. Macro-Precision is 92% for the multi-classes. It means that low than 8% of predictive data are false positive.


Confusion Matrix: Resnet50
Output Class
Target Class

	HK-Classical RS	HK-NLP NodPred	NHK-BcellBurkitt	NHK-BcellFollicular	NHK-BcellMantel	NHK-LargeBCell	NHK-TCellLymph	
HK-Classical RS	48 14.9%	2 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.6%	92.3% 7.7%
HK-NLP NodPred	4 1.2%	52 16.1%	3 0.9%	0 0.0%	0 0.0%	0 0.0%	2 0.6%	85.2% 14.8%
NHK-BcellBurkitt	1 0.3%	0 0.0%	42 13.0%	0 0.0%	1 0.3%	0 0.0%	0 0.0%	95.5% 4.5%
NHK-BcellFollicular	0 0.0%	0 0.0%	0 0.0%	39 12.1%	2 0.6%	0 0.0%	1 0.3%	92.9% 7.1%
NHK-BcellMantel	0 0.0%	0 0.0%	0 0.0%	1 0.3%	39 12.1%	0 0.0%	0 0.0%	97.5% 2.5%
NHK-LargeBCell	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	43 13.3%	1 0.3%	97.7% 2.3%
NHK-TCellLymph	6 1.9%	1 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	33 10.2%	82.5% 17.5%
	81.4% 18.6%	94.5% 5.5%	93.3% 6.7%	97.5% 2.5%	92.9% 7.1%	100% 0.0%	84.6% 15.4%	91.6% 8.4%

Figure 10: Confusion matrix results using ResNet50 for the 7 multi classes. HK-NLP has 2 observations wrongly predicted as class HK-Classical. Validation Accuracy = 296/323 = 0.916, Error = 0.084

Table 4: Performance metrics for predictive model evaluation

	TP	TN	FP	FN	Precision	Recall	F-score	Kappa
HK-Classical RS	48	260	11	4	0,814	0,923	0,865	0,914
HK-NLP NodPred	52	259	3	9	0,945	0,852	0,897	0,914
NHK-Bcellgurkitt	42	276	3	2	0,933	0,955	0,944	0,915
NHK-BcellFollicular	39	280	1	3	0,975	0,929	0,951	0,915
NHK-BcellMantel	39	280	3	1	0,929	0,975	0,951	0,915
NHK-Largel3Cell	43	279	0	1	1,050	0,977	0,989	0,915
NHK-TCeIlLymph	33	277	6	7	0,846	0,825	0,835	0,915
		Macro Mean			0,920	0,919	0,919	0,915
		Weighted mean			0.916	0,918	0,916	

We have computed our method using just four lymphomas (NHK-LargeBcell, NHK-BcellFollicular, NHK-Mantel, and HK-NLP) to compare our approach to Brancati et al. [18]. Our positive predictive value (precision) is highest with 97.73% against 94% for the method.

We also observe that the significant results using the model are promising and give high recall values of 0.955, 0.975, and 0.977 for NHK-BcellBurkitt, NHK-BcellMantel, and NHK-LargeBcell, respectively. The Macro-Recall of the 7 multi-classes is about 91.9%. It means that 8% of predictive data are false negatives.

As shown in Fig. 11, the comparison of the two performance metrics (Precision and Recall), the precision and recall can not give the correct performance of the model. We observe that the precision rate for HK-NPL is high 94.55% when the recall is low 85.25%. The same results for the NHK-Follicular class. Low recall and high precision show that we miss many positive examples (high FN), but those we predict as positive are indeed positive (low FP). Adding to the previous parameters, the harmonic mean of the precision and recall, we calculated, F-score (F1) to measure a test accuracy and reach the best value. The model achieves a Macro-F1 of 91.9% for the 7 multi-classes.

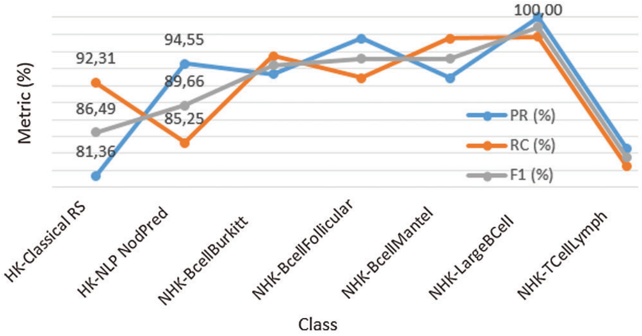


Figure 11: Performance metrics comparison

When averaging the macro-F1, we gave equal weights to each class. The Weighted-F1 weight the score of each class by the number of samples (images) from each class. The results are equally convincing. The model gives a Weighted-F1 of 91.6% for the 7 multi-classes. The problem is that parameter F1 ignores the count of True Negatives.

In contrast, the Kappa score measures, for the multi-classes, the degree of agreement between the true values and the predicted values by the classifier. Sun [34] commented that the Kappa score is high only if the classifier is doing well on both the negative and the positive elements. The values obtained to confirm the previous excellent results, as shown in Tab. 4, the average Kappa score for the seven multi-classes, is 91.5%. Since 0.915 means perfect agreement.

We have test 10 new lymphoma images with our trained classifier. Fig. 12 illustrates the 9 images of the results. Tab. 5 shows the score of each new test image (1 means 100%). Nine images from 10 were successfully classified.

We believe that the results obtained are very relevant. For most of the images, the system predicted almost 100% of the disease class. Two cases require the opinion of an expert. Image 6 that we believe to group two lymphoma types T-cell lymphomas with large B-cell proliferation and whose system has predicted both classes well NHK-Tcell and NHK-LargeBcell with scores of 40% and 44%, respectively. This result is very interesting. The score is shared almost equally. It is due to the presence of two lymphoma in the same image. However, a patient can develop two forms of lymphoma since the disease can mutate.

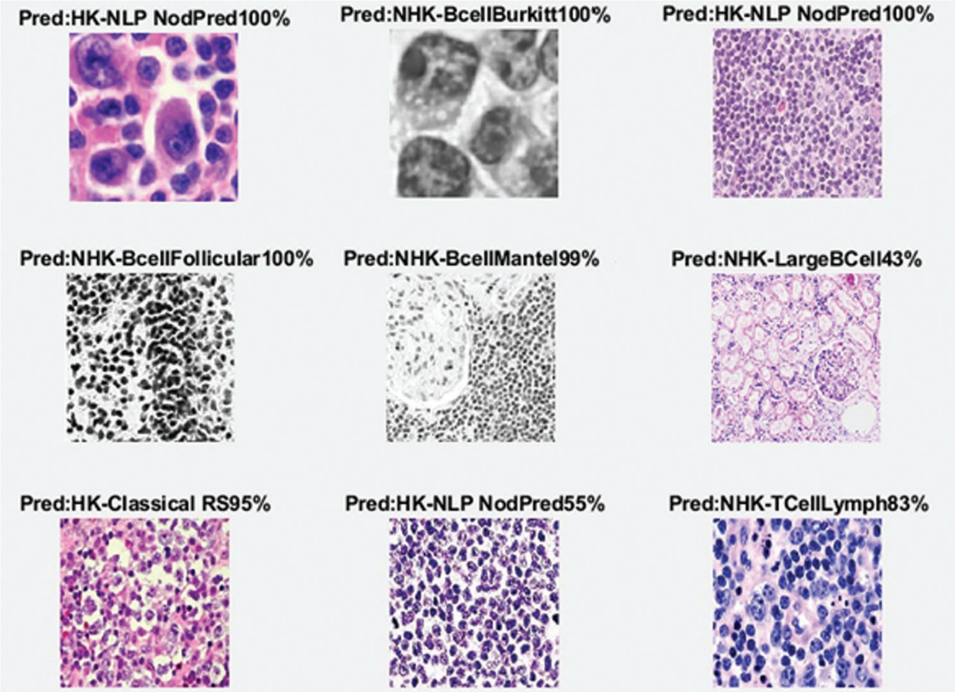


Figure 12: A new set of Lymphoma test images: good classification of 8 images, Error detection (img8) HK-NLP score = 55%

Table 5: Validation score of new dataset

	Score								
Lymphoma Class	Img1	img2	img3	img4	img5	img6	img7	img8	img9
HK-Classical RS	0,00	0,00	0,00	0,00	0,01	0,16	0,95	0,00	0,03
HK-NLP NodPred	**1,00**	0,00	**1,00**	0,00	0,00	0,00	0,01	**0,55**	0,12
NHK-BcellBurkitt	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00
NHK-BcellFollicula	0,00	0,00	0,00	**1,00**	0,00	0,00	0,02	0,00	0,00
NHK-BcellMantel	0,00	0,00	0,00	0,00	**0,99**	0,00	0,00	0,00	0,00
NHK-LargeBCe11	0,00	0,00	0,00	0,00	0,00	**0,44**	0,02	0,00	0,03
NHK-TCellLymph	0,00	0,00	0,00	0,00	0,00	**0,40**	0,00	**0,45**	**0,83**

5 Conclusion

We have proposed an approach to assist pathologists in classifying lymphoma disease using Transfer Learning. The approach classifies seven different lymphomas types that rare, aggressive, and difficult to be diagnosed. The Deeper Residual Neural Network, ResNet50, is performing and showing its capability of classifying the seven types of lymphomas with accuracy that can top to 91.6%. The disadvantage is in the training time, where for one fold and 323 images, our system process in 176 mn can be solved using parallel processing. Our experimental results suggest that the proposed approach can classify the correct lymphoma disease accurately. However, to improve accuracy in our future works, we plan to increase the number of images. We also intend to analyze the image locally, combining several diseases in the same image when the patient develops different lymphoma types.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] U.S. National Cancer Institute, [Online]. Available: http://www.cancernet.gov/cancertopics/.

[2] R. L. Siegel, K. D. Miller and A. Jemal, “Cancer statistics, 2020,” *CA A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[3] The Leukemia & Lymphoma Society (LLS), [Online]. Available: https://www.lls.org/lymphoma/hodgkin-lymphoma/diagnosis/.

[4] O. Sertel, G. Lozanski, A. Shanaah and M. Gurcan, “Computer aided detection of centroblasts for follicular lymphoma grading using adaptive likehood based cell segmentation,” *IEEE Transactions on Biomedical Engineering*, vol. 157, pp. 2613–2619, 2010.

[5] M. N. Gurcan, L. E. Boucheron, A. Can and A. Madabhushi, “Histopathological image analysis,” *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[6] P. Dieusaert, “Guide pratique des Analyses Medicales,” in *MALOINE*, 6eme . edition, France, pp. 462–571, 2015.

[7] G. P. Canellos, T. A. Lister and B. Young, “The Lymphomas, ” in *Sanders*, 2nd. edition, London, pp. 321–465, 2006.

[8] G. Bueno, M. M. Fernandez-Carrobles, O. Deniz and M. Garcıa-Rojo, “New trends of emerging technologies in digital pathology,” *Pathobiology*, vol. 83, no. 2–3, pp. 61–69, 2016.

[9] Q. Song, L. Zhao, X. Luo and X. Dou, “Using deep learning for classification of lung nodules on computed tomography images,” *Journal of Healthcare Engineering, Hindawi*, vol. 2017, pp. 7 pages, 2017.

[10] T. Purwaningsih, T. Nurhikmat and P. B. Utami, “Image classification of golek puppet images using convolutional neural networks algorithm,” *International Journal of Advances in Soft Computing & Its Applications*, vol. 11, no. 1, pp. 34–45, 2019.

[11] M. G. Ribeiro, L. A. Neves, G. F. Roberto, A. A. Tosta, S. A. Martins *et al.,* “Analysis of the influence of color normalization in the classification of non-hodgkin lymphoma images,” in *Proc. of the 31st Conf. on Graphics, Patterns and Images (SIBGRAPI)*, Paraná, Brazil, vol. 1, pp. 369–376, 2018.

[12] I. Rizwan, I. Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” *Journal Informatics in Medicine Unlocked*, vol. 18, pp. 1–12, 2020.

[13] T. A. Azevedo Tosta, L. A. Neves and M. Z. do Nascimento , “Segmentation methods of H&E-stained histological images of lymphoma,” *Review Informatics in Medicine Unlocked*, vol. 9, pp. 35–43, 2017.

[14] P. Battula and S. Sharma, “Automatic classification of non-hodgkin’s lymphoma using histological images: Recent advances and directions,” in *Proc. of the IEEE Int. Conf. on Advances in Computing, Communication Control and Networking*, Greater Noida, India, pp. 634–639, 2018.

[15] M. Z. do Nascimento , A. S. Martins, T. A. Azevedo Tosta and L. A. Neves, “Lymphoma images analysis using morphological and non-morphological descriptors for classification,” *Computer Methods and Programs in Biomedicine*, vol. 163, pp. 65–77, 2018.

[16] M. Lippi, S. Gianotti, A. Fama, M. Casali, E. Barbolini *et al.,* “Texture analysis and multiple-instance learning for the classification of malignant lymphomas,” *Computer Methods and Programs in Biomedicine*, vol. 185, pp. 1–8, 2020.

[17] N. V. Orlov, W. W. Chen, D. M. Eckley, T. J. Macura, L. Shamir *et al.,* “Automatic classification of lymphoma images with transform based global features, ” in *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society.* vol. 14, pp. 1003–1013, 2010.

[18] N. Brancati, G. De Pietro, M. Frucci and D. Riccio, “A deep learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images,” *IEEE Access*, vol. 7, pp. 44709–44720, 2019.

[19] R. Tambe, S. Mahajan, U. Shah, M. Agrawal and B. Garware, "Towards designing an automated classification of lymphoma subtypes using deep neural networks," in *Proc. of the ACM Joint Int. Conf. on Data Science and Management of Data*, India, pp. 143–149, 2019.

[20] H. El Hachi, T. Belousova, L. Chen, A. Wahed, I. Wang *et al.,* "Automated diagnosis of lymphoma with digital pathology images using deep learning," *Annals of Clinical and Laboratory Science*, vol. 49, pp. 153–160, 2019.

[21] U. V. Somaratne, K. W. Wong, J. Parry, F. Sohel, X. Wang *et al.,* "Improving follicular lymphoma identification using the class of interest for transfer learning," in *Proc. of the Digital Image Computing: Techniques and Applications (DICTA)*, Hyatt Regency Perth, Australia, pp. 1–7, 2019.

[22] K. Wah Wen, B. Fakhri, J. Menke, R. Ruiz-Cordero, R. M. Gill *et al.,* "Complexities in the diagnosis of large B-cell lymphomas, classic hodgkin lymphomas and overlapping peripheral T-cell lymphomas simplified: An evidence-based guide," *Annals of Diagnostic Pathology, Elsevier*, vol. 46, pp. 1–10, 2020.

[23] D. Ramnani, WebPathology. [Online]. Available: https://www.webpathology.com/.

[24] American Society of Hematology, [Online]. Available: http://imagebank.hematology.org/.

[25] T. T. Wong and N. Y. Yang, "Dependency analysis of accuracy estimates in k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 11, pp. 2417–2427, 2017.

[26] B. N. Narayanan, V. S. P. Davuluru and R. C. Hardie, "Two-stage deep learning architecture for pneumonia detection and its diagnosis in chest radiographs," in *Proc. of the SPIE Medical Imaging Conference*, Houston, Texas, 11318, pp. 10 pages, 2020.

[27] A. Mahmood, M. Bennamoun, S. An and F. Sohel, "Residual network based features for image classification," in *Proc. of the Int. Conf. on Image Processing (ICIP)*, Beijing, China, pp. 1597–1601, 2017.

[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.

[29] D. Theckedath1 and R. R. Sedamkar, Detecting affect states using vgg16, resnet50 and se-resnet50 networks. In: *SN Computer Science* . Springer, pp. 1–79, 2020.

[30] K. He, X. Zhang, S. Ren and J. Sun, "Identity mappings in deep residual networks," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 630–645, 2016.

[31] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[32] M. Sokolova, N. Japkowicz and S. Szpakowicz, "Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation, " in *Advances in Artificial Intelligence Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 1015–1021, 2006.

[33] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, pp. 6–21, 2020.

[34] S. Sun, "Meta-analysis of Cohen's kappa," *Health Services and Outcomes Research Methodology*, vol. 11, pp. 145–163, 2011.

[35] D. Hand and P. Christen, "A note on using the f-measure for evaluating record linkage algorithms," *Journal of Statistics and Computing*, vol. 28, no. 3, pp. 539–547, 2018.