INTERNATIONAL JOURNAL ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 20th Oct 2016.

Link : https://ijiemr.org/downloads/Volume-05/Issue-11

Title : **IMPLEMENTING OF HUMAN BLOOD CELL CLASSIFICATION AND LEUKEMIA PREDICTION USING CONVOLUTIONAL NEURAL NETWORKS**

volume 05, Issue 11, Pages: 50-64

Paper Authors: [1]**Mr.M.Prabhakar Rao,**[2] **Ms.K.Radhika,**

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IMPLEMENTING OF HUMAN BLOOD CELL CLASSIFICATION AND LEUKEMIA PREDICTION USING CONVOLUTIONAL NEURAL NETWORKS

[1]Mr.M.Prabhakar Rao,[2] Ms.K.Radhika,
[1,2] Assistant Professor,Dept. of CSE,
Malla Reddy Engineering College (Autonomous), Secunderabad, Telangana State

ABSTRACT: Recognition of diseases is still the biggest challenge in hematological sectors; therefore more accurate algorithms are required to substantiate the pathological laboratory technicians. In a manual or more of a traditional method in the detection of Leukemia, doctors check the microscopic images of the blood samples. This is tedious and time-consuming approach and the accuracy of the detection is mostly depending on the experience and skills of the individual, which is not reliable all the time. The alternative solution is an automated process or system which can analyze the same blood sample images. It takes in the required portions of the blood sample image and does some filtering process to the images. Kmean clustering followed by SVM classification is one such approach used for finding the affected cells. The main drawback is that it just classifies the Lymphocytes and Myelocytes white blood cells for leukemia detection. To overcome these drawbacks, this research work specifically focusing on the classification of various types of Blood cells and extraction of accurate blood cells, ranking of the blood cells to cause effective prediction of leukemia blood caner through the blood cell classification respectively. The proposed Convolutional neural network (ConvNET) classification results the higher performance compared to the state of art approaches.

Keywords: Human blood cells, hematology, deep learning, CNNs.

1. INTRODUCTION

The disorder leukemia is a major type of cancer caused by the abnormality in producing WBC [1]. As per the researches on cancer, it is estimated that around 132,574 cases will be associated with the lymphoid system alone in the year of 2020. It was 104,239 and 117,649 in the years 2010 and 2015 respectively [2]. In the hematopoietic system, the bone marrow is producing the blood. The most specific blood components are plasma, sugar, white blood cells and red blood cells. The WBC are categorized into three types: Lymphocytes, monocytes, and granulocytes. Amongst them, the granulocytes are differentiated into three types as basophil, neutrophil, and eosinophil. The increase in the count of WBC causes diseases, tumors, and cancers like leukemia. Existence in the blood of some of these cells is expressed in terms of accurate count rather in percentage. Recognition and classification of the types of WBC help in discriminating the diverse types of leukemia. The most general types[3] are Acute Lymphatic Leukemia (ALL), Acute Myelogenic Leukemia (AML), Chronic Lymphatic Leukemia (CLL), and Chronic Myelogenic (CML) [4][5]. Acute leukemia grows faster than chronic types. For this purpose, it is essential to characterize the WBC type. The progression in the population of a specific kind of WBC in peripheral blood and change in their texture, shape, and color define what sort of leukemia it is [6].

In [7] authors used the standard algorithm ECOC with OVA and OVO. ECOCECS does the multi-class classification by finding the distance between centroid of classes. The centroids are detected by the typical formula (25) of finding average. The existed ECOCAMD [8] finds the accurate mean. Same way, the ECOCECS algorithm finds the distance using the default Euclidean metric. Instead, the ECOCAMD uses a novel distance metric (23 and 27) on both data-mismatch and Distk,i equations. This in turn ensures the accurate distance between neighboring classes and enables to pick most suitable classes to a partition. Generally, the discrimination of blood cells is done with the microbiological examinations by technicians at hematological laboratories. This work is an attempt to lend a hand to the hematopathologists for classifying WBC in ascertaining the leukemia and stepping forward the medication accordingly. Lots of algorithms are being presented with this intention [9]. Moreover, this proposal accomplishes enhanced performance by means of simplicity and accuracy nearer to the manual microscopic test method. Since healthcare sectors believe on accurate predictions, this work bestows its motive.

The major contributions of the paper as follows:

- Preprocessing of the test and training images has been performed, so noises form the dataset will be removed.
- Effective method of feature extraction, feature classification and feature selection operations has been performed.

- The ConvNET method was implemented classification on public available dataset for both blood cell classification and leukemia prediction, the results shows that the proposed SVM classification gives the better performance compared to other approaches.

Rest of the paper is organized as follows; section 2 deals with the various literatures with their drawbacks respectively. Section 3 deals with the detailed analysis of the proposed method with its operation. Section 4 deals with the analysis of the results with the comparison analysis. Section 5 concludes the paper with possible future enhancements.

2. LITERATURE SURVEY

The processes of segmenting a required component of any blood smear image sample and making multi-class decisions based on the features extracted from them are done in diverse methods. Some of the methods are analyzed as a starting point of this proposal: In [10] authors classified the classes of WBC with 91% of accuracy using Artificial Neural Networks (ANN) classifier. Their proposal shows that doing proper segmentation and giving appropriate features the accuracy of classification can be increased. In [11] authors achieved 88.33% of accuracy in identifying blood cells of varying types of leukemia using bi-modal thresholding and Morphological Image Processing. Bi-modal thresholding helps in segmenting the nucleus as well as cytoplasm regions based on the difference in their intensity levels.

In [12] authors classified the types of WBC using HSV color model and geometric features of the blood cells. Their work proves that the appropriate selection of features can increase the classification accuracy of up to 99%. In [13] authors used L*a*b* color space images and fuzzy-c means (FCM) algorithm for segmenting the WBC. They extracted a set of color, texture and geometrical features and this proposal achieves 97.73% of the correct rate in classifying the types of WBC with SVM classifier. In [14] authors have come up with a system for categorizing the types of leukemia based on WBC segmentation and classification. Segmentation is done with FCM clustering and the L1, L2, and L3 types of ALL are classified with SVM classifier. Their system is 77.52% accurate in classification. In [15] authors used neural networks based classifier to classify the types of WBC in the accuracy rate of 76%.

The classification is done by considering the shape, intensity and textural features of segmented WBC.

In [16] authors experimented three different segmentation methods specifically k-means clustering, HSV color space based segmentation and markercontrolled watershed algorithm for segmenting WBC in blood smear images. Then the leukemia types are classified into AML, ALL, CML, and CLL by using SVM classifier. The accuracy rates of the classifier with these different algorithms are not shown.

In [17] authors used Gram-Schmidt orthogonalization and snake algorithm for segmenting the nucleus region of WBC. Leukemia and its subtypes are classified with ANN and SVM classifiers. According to their research SVM classifier achieves 76% of classification accuracy with GLCM features, which is far better than the ANN classifier.

In [18] authors worked on blood smear images converted into HSV color pattern and segmented the images with the k-means clustering algorithm. The segmented WBC are classified into ALL affected or healthy cells by applying binary Support Vector Machine (SVM) classifier with the accuracy rate of 77%. Then WBC is further classified into ALL subtypes using Multi-SVM classifier with 75.33% of accuracy. This proposal has experimented with only 21 blood smear images.

In [19] authors have proposed a new image processing based framework which incorporates two algorithms. One is Haar wavelet for image transformation and the other one is K- Nearest algorithm for image classification, the main objective in the proposed system is to develop a malaria parasite detection system in which pathology admin will transfer the patient's scanned RGB report. To build up an expert system for patients after uploading image transformation, feature extraction and image classification. The feature extractions are done by uploading image and scale those images onto 256*256 pixels and transform the original image using Haar wavelet algorithm. It is used to compress the images and store those pictures for further classification. Image classification is done by using KNN algorithm by calculating Euclidean distance with the help of extracted features. In Euclidean distance systems will form clusters of multiple stages among these clusters suitable cluster will be considered as a final malarial stage. Then the K-Nearest Neighbour algorithm which is a method that does not use the estimation of parameters is used. Input consists of K-closest nearest sample in the feature. So, this system is interactive, hence is faster and more accurate than manual process. This system will help limit the human mistake while recognizing the presence of malaria parasites in the blood sample by using Image Processing and limit human blunder by automation.

In [20] authors performed Image segmentation and feature extraction using minimum distance classifier was used to identify the parasites in the blood sample. Feature extraction uses two phases in architectural model: Training phase and Recognition phase which helps to recognize the Malaria parasite. In this work, they focus on automated detection and quantification of malaria detection, the strategy to determine infected images using machine learning to improve the predictive value for detection of infected cells. The image is acquired that may contain impurities and noise. It is converted to gray scale and the zones are segmented by recognizing the similar properties. The image is threshold by creating binary images for grey-level ones by converting all pixels below some threshold to zero and all pixels above to one. Then, the image was enhanced to make it more suitable for further processing based on intensity property. Erosion and Dilation are applied to remove a considerable amount of noise. After this, the images are segmented using watershed segmentation. It tends to separate touching objects so that overlapped RBCs will be separated and will be helpful for counting the RBCs. Using the CIE system they have specified any color in terms of its coordinates and have measured the sensitivities of three broad band's by suiting spectral colors to certain mixtures of three colored lights. After the segmentation the mean perimeter of the RBCs are found with the help of Matlab function region props helps to quantify properties of image regions. Then the parasite compares whether it is greater than the mean value of RBC cells. A circle is plotted around the infected RBC and they are calculated.

3. PROPOSED METHOD

The research proposes a smart method to facilitate the detection and classification of blood cancer. By doing so physicians and oncologists can provide proper timely treatment to the patients. This can actually increase the survival rates of patients. Fig 1 depicts the proposed architecture. The proposed method includes initial data acquisition process followed by data pre-processing step. From the pre-processed data, a data set is created for training aspect. Subsequently, certain features from the data set are then ranked using machine learning algorithms. After ranking, features are selected and then it is applied to a feature classifying algorithm. Finally, a prediction model is created based on a classification algorithm.

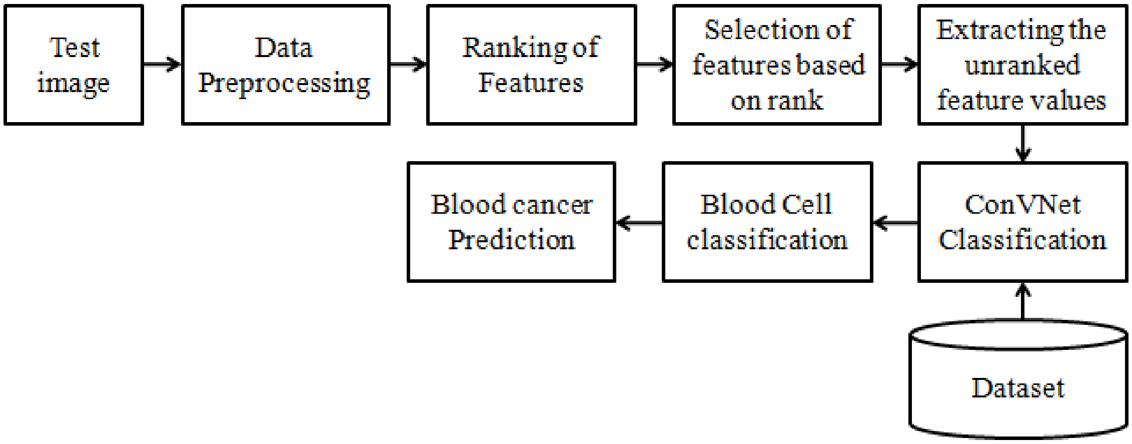

Fig 1: Architecture of the Prediction Model using Feature Ranking and ConVNet Algorithm

3.1 Data Pre-processing:

The datasets created will also have information regarding the various stages of blood cancer. The information is regarding the cell details. This cell details are very much helpful in classification of stages in blood cancer.

3.2 Ranking of features and selection of features:

Following the data collection and pre-processing, the data have to be now ranked and then selected. The ranking is based on certain features of the data sets. Ranking algorithms are used for this step. An amalgamation of several search techniques are utilized for feature selection and the result is a feature subset. Additionally a scorecard is maintained to keep track of evaluation process. Machine learning literature proposes a variety of feature ranking and selection methods. During this stage, irrelevant features are discarded. In the architecture, we propose two steps for the feature ranking and selection. First step is the subset generation followed by a subset evaluation stage as depicted in Fig 2 [20]. Subset evaluation stage involves a filter method. The Table 1 below gives certain performance domains based on which the evaluation and ranking of feature subsets are done [20].

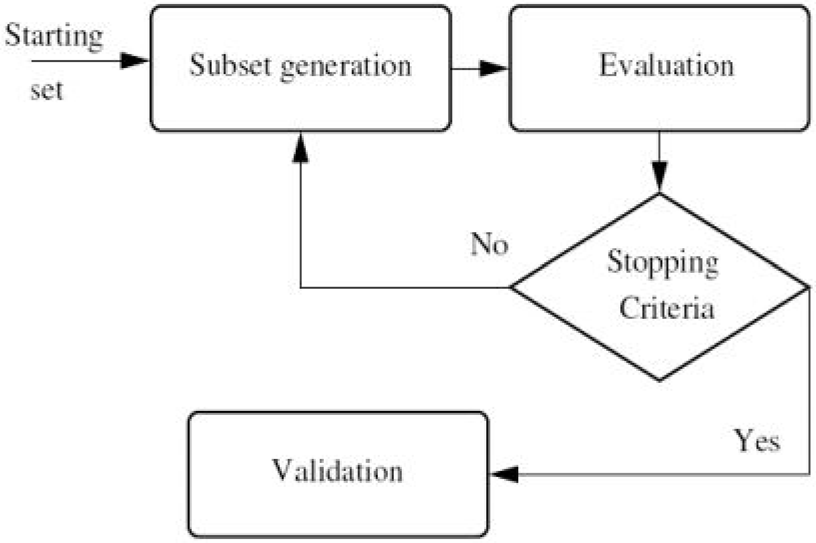

Figure 2: Feature Selection-Subset creation and Subset evaluation

3.3 Classification:

The figure 3 show the workflow that shows the entire process carried out for the classification of leukemia from the blood sample images. The process starts with loading the dataset from the file to the program. The separating the images for testing and training process as shown in figure 3, then these loaded images are once verified before passing those images into the modified ConVNet.

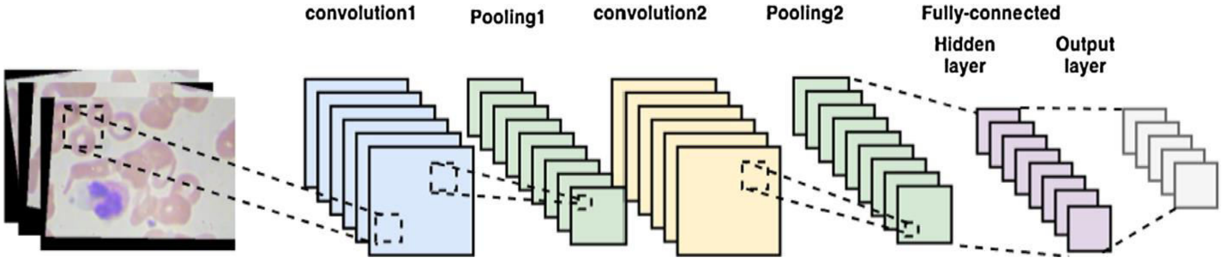

Figure 3: Architecture of ConVNet

The model is then defined with the various CNN layers for the classification of leukemia. After the creation of the model, the fitting of the model takes place. Fitting is the process where the parameters of the model is declared. [11] The parameters like the loss function used in the network and the metrics that are to be calculated are given. The training is then started with the provided parameters and the metrics to be calculated. Initially, the network is trained for about

100 epochs and then the accuracy, the loss is obtained from training the model [14-15]. These values are then plotted in the form of graph to better visualize the result. This process is again repeated for another 100 epochs and the graph is plotted. After the training is done to a sufficient accuracy the model is then tested with the testing set for the evaluation. Then the model is made to predict the unseen image. This is the entire process from the start to the end.

The model is built by arranging various layers of CNN in order, such that the model exhibits high performance. The input image size is 200x200 with the RGB color format. The input layer is then passed on to a convolution layer having filters that are specified in our function declaration and then attaches the Batch Normalization layer, after thatan activation function ReLU is used, and then another convolution layer is stacked in. Now, this combination is repeated and for the output layer the input is flattened. [21] This step is necessarily to be done to match the layer sizes of both the layers that are going to be converged. As an add on max-pooling layer is added and also dropout layer is used to avoid any over fitting that could occur in the network. There is a total of six layers used in the design of this complex model. (figure 2) Each of those layers contains several layers such as a Convolution layer, Activation layer, dropout layer and then the max-pooling layer. Finally, these layers are converged to the fully connected dense layer which is made to process the images to their respective labels. This classification involves a total of four classes. [22]

For cancer imaging, convolutional neural networks identify spatial connection between pixels. This process is done in a hierarchical manner. According to Vogado et al, medical imaging of the human body is done by convolving medical images. This process of convolution is done through scholarly channels which in turn create a chain of mapping highlights [18]. The process of convolution s carried out in numerous layers and this determines the capacity of convolution. The highlight achieved through the process are the interpretations that could be the high level of exactness. The CNN has several layers which are discussed as follows:

Input Image Format Layer: A variety of pixels values determine the image input. These pixels depend on the size of the picture according to Lopez-Rincon et al [9]. For instance a 3 × m × n cluster of numbers depicts a shaded information picture. Here ‘3’ signifies the values blue, green, and red with the values of the pixels for each shading running from 0–255; also, m and n are the elements of the picture[9]. On account of a grayscale picture, the picture size is characterized by 2D exhibit (m × n), where the force of the pixels likewise extends from 0–255 [9].

Convolution Layer: Convolutional Layer is the most prominent and important layer of CNN. This layer using the convolutional channels captures highlights from the provided input image. These channels are square exhibits with numbers which characterise loads and parameters. The top left corner of the image is actually the initial position of the channel in the convolutional process. In this process, the duplication of picture pixel lattice progressively. In addition, the channel grid is also progressed. According to Mocan et al, adding the pixel lattice and channel grid rehashes the sliding channels to one [11].The strides indicate the amount of cell movements to one side during each and every progression [11]

Pooling Layer: Besides convolution layers, CNNs fairly often use supposed pooling layers. They are used primarily to scale back the dimensions of the tensor and speed up calculations. These layers are unit straightforward - we want to divide our image into totally different regions, and then perform some operation for each of these elements. As an example, for the scoop Pool Layer, we tend to choose the most worth from every region and place it within the corresponding place within the output. As within the case of the convolution layer, we've got2hyper parameters accessible — filter size and stride. However, if you 'reper forming pooling for a multi-channel image, the pooling for every channel has to be done individually.

Fully Connected Layers: Located at the end of the neural network, these layers are connected with all the activation location in all the layers preceding it. These layers compile the information extracted from previous layers to generate the final output. Fully connected layers classify the information into various classes after feature extraction. After Convolution layer it is the most time-consuming layer.

Softmax output Layer: The softmax function is commonly used in the final layer of CNNs. The status of a particular class data is shown and a value is generated about which class is closer to it. The probability value of each class is extracted by performing the calculation of probability in network. For these procedures cross entropy is used. Finally, it will summarizes the information and compares with respect to the database and gives the accurate classification of blood cells and then also predicts the potential Leukemia blood cancer.

4. EXPERIMENTAL RESULTS

4.1 Datasets:

A large data set is required for making a clear and accurate prediction. The National Center for Biotechnology Gene Expression Omnibus database was referred. The database included molecular cell details for peripheral blood monocular class and bone marrow. The datasets were divided into three groups' namely-HG-U133A microarray (dataset1), the HGU133 2.0 microarray (dataset2), and Illumina RNA-seq (dataset3). The samples included in the dataset were mainly Acute Lymphocytic Leukemia(ALL), Acute Myeloid leukemia (AML),Chronic Lymphocytic leukemia(CLL), Chronic Myeloid leukemia(CML), Myelodysplastic syndrome(MDS) and other non- leukemia diseases. Duplicate values of the data sets were excluded and pre-filtered.

4.2 Performance evaluation:

The random samples are taken for training and testing the classifier. Out of 242 images 142 images are used for training purpose and 100 images are used for testing the performance. Amongst these 100 images 22 are lymphocytes, 21 are monocytes, 19 are eosinophils, 18 are neutrophils, and 20 are basophils. In the proposed system, ConvNET technique uses Gaussian kernel function along with binary learner classifiers for making prediction.

Table 1: The features extracted from the sample output images.

Features	Basophil	Eosinophil	Lymphocyte	Monocyte	Neutrophil
Contrast	0.002183809	0.000984252	0.001107283	0.002337598	0.001599409
Homogeneity	0.998908	0.999508	0.999446	0.998831	0.9992
Correlation	0.926737	0.857913	0.871784	0.896941	0.86688
Energy	0.968013	0.99209	0.990258	0.974986	0.986388
Centroid	37.7185	132.7655	126.3688	154.6381	165.5903
Area	984	226	282	746	393
Perimeter	119.927	56.577	57.66	151.585	91.204
MinorAxisLength	32.05057	14.05057	17.33324	25.49919	18.047
MajorAxisLength	40.37866	21.57597	20.91896	46.49351	31.06539

The features which are considered as most appropriate are gathered from the segmented WBC nucleus region of blood smear images. This appropriateness reduces errors in classification. Table 1 summarizes the features extracted from a sample image. The pulled out features are represented as a feature matrix and fed into the ConvNET classifier algorithm.

In Table 2, TP represents the True Positives and E represents Error predictions. From this confusion matrix the performance measures of a classifier such as an accuracy, error-rate, sensitivity, specificity and, F-measure are calculated using the values of True Positives (TP- The total number of TP's of a particular class is the sum of right predictions in the corresponding row or column), False Negatives (FN- The total number of FN's of a particular class is the amount of error predictions in the corresponding row excluding the TP), True Negatives (TN- The sum of all TN's of a particular class is the sum of all rows and columns excluding that class's row and column) and, False Positives (FP- The total number of FP's of a particular class is the amount of error predictions in the corresponding column excluding the TP).

Table 2: Performance of the ECOCAMD classifier on predicting each type of WBC

	Lymphocyte	Monocyte	Eosinophil	Neutrophil	Basophil
Accuracy	98	100	99	98	99
Sensitivity	95.45	100	94.73	100	95
Specificity	98.75	100	100	97.56	100
F-Measure	95.45	100	94.74	94.74	97.44
Error Rate	2	0	1	2	1

The above results are compared with the manual predictions made by a human expert. The performance measures calculated through the confusion matrix are 98.81% ((100+397)/(100+397+3+3)) of overall accuracy, 97.09% (100/(100+3)) of sensitivity, 99.25% (397/(397+3)) of specificity and, 97.09 % ((2×100)/((2×100)+3+3))of F-measure.

Table 3: Scrutiny of Performance of Various Classifiers

Technique	**Accuracy**	**Specificity**	**Sensitivity**
ConvNET	98.46%	97.33%	95.59%
SVM[15]	94.30%	81.81%	100%

ANN [17]	94.04%	95.92%	97.60%

Performance of the proposed ConvNET system is compared with number of existing systems, which is shown in Table 3. Very less research works are there in classifying the WBC along with Leukemia blood cancer and this proposal presents much improved accuracy than all of them.

5. CONCLUSION

A predominant facet of distinguishing the category of leukemia is recognizing the sort of WBC got damaged and causes to Leukemia blood cancer. Focusing on WBC of a peripheral blood smear, the proposed system segments the cells using data preprocessing based thresholding and mathematical morphology methods. Then the essential features are extracted and passed to the proposed multi-class feature extraction, feature selection and feature ranking respectively. The proposed model is to identify the types of leukemia using ConvNET. The obtained results show the effectiveness of ConvNET for identifying the types of leukemia. The accuracy is greater than 98% for a dataset with a small number of classes. In the future works, the ConvNET can be designed in such a way to effectively classify and predict the different forms of blood sample images. In the future, the segmentation and classification accuracy of the system proposed can be improved with customized classification algorithms, provided that the manual process of analyzing the blood smears is made as a digitized online task. Changes can be made in the classifier by exploring on the direction of making modifications in the distance since it is also playing a crucial role in classification.

REFERENCES

[1] Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. Science 349, 255-260, doi:10.1126/science.aaa8415 (2015).

[2] van Ginneken, B. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. Radiological Physics and Technology 10, 23-32, doi:10.1007/s12194-017-0394-5 (2017).

[3] de Bruijne, M. in Med Image Anal Vol. 33 94-97 (Elsevier, 2016).

[4] Kerr, W. T., Lau, E. P., Owens, G. E. &Trefler, A. The future of medical diagnostics: large digitized databases. Yale J Biol Med 85, 363-377 (2012).

[5] Kukar, M., Kononenko, I. &Grošelj, C. Modern parameterization and explanation techniques in diagnostic

decision support system: A case study in diagnostics of coronary artery disease. Artificial intelligence in medicine 52, 77-90 (2011).

[6] Šajn, L. &Kukar, M. Image processing and machine learning for fully automated probabilistic evaluation of medical images. Computer methods and programs in biomedicine 104, e75--e86 (2011).

[7] Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115-118, doi:10.1038/nature21056 (2017).

[8] Yamamoto, Y. et al. Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach. Scientific Reports, 46732, doi:10.1038/srep46732 (2017).

[9] Badrick, T. Evidence-based laboratory medicine. The Clinical Biochemist Reviews 34, 43 (2013).

[10] Luo, Y., Szolovits, P., Dighe, A. S. & Baron, J. M. Using Machine Learning to Predict Laboratory Test Results. American journal of clinical pathology 145, 778-788, doi:10.1093/ajcp/aqw064 (2016).

[11] Gehlot, Shiv, Anubha Gupta, and Ritu Gupta. "SDCT-AuxNetθ: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis." *Medical image analysis* 61 (2020): 101661.

[12] Vijayakumar, T. "Neural network analysis for tumor investigation and cancer prediction." *Journal of Electronics* 1.02 (2019): 89-98.

[13] Thanh, T. T. P., et al. "Leukemia blood cell image classification using convolutional neural network." *International Journal of Computer Theory and Engineering* 10.2 (2018): 54-58.

[14] Rajpurohit, Subhash, et al. "Identification of acute lymphoblastic leukemia in microscopic blood image using image processing and machine learning algorithms." 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.

[15] He, Jian, et al. "Deep convolutional neural networks for predicting leukemia-related transcription factor binding sites from DNA sequence data." *Chemometrics and Intelligent*

Laboratory Systems 199 (2020): 103976.

[16] Naz, Isra, et al. "Robust discrimination of leukocytes protuberant types for early diagnosis of leukemia." *Journal of Mechanics in Medicine and Biology* 19.06 (2019): 1950055.

[17] Naz, Isra, et al. "Robust discrimination of leukocytes protuberant types for early diagnosis of leukemia." *Journal of Mechanics in Medicine and Biology* 19.06 (2019): 1950055.

[18] Mostavi, Milad, et al. "Convolutional neural network models for cancer type prediction based on gene expression." *BMC medical genomics* 13 (2020): 1-13.

[19] Kilicarslan, Serhat, Kemal Adem, and Mete Celik. "Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network." *Medical hypotheses* 137 (2020): 109577.

[20] Ahmed, Nizar, et al. "Identification of leukemia subtypes from microscopic images using convolutional neural network." *Diagnostics* 9.3 (2019): 104.