Convolutional Neural Network-Based Deep Learning Model Evaluation of Imbalanced Leukocyte Classification Datasets

Article *in* Jisuan Lixue Xuebao/Chinese Journal of Computational Mechanics · October 2023

CITATIONS
0

READS
124

4 authors, including:

Hayder Al-Ghanimi
Hilla University College
18 PUBLICATIONS **1** CITATION
SEE PROFILE

Ahmed Al-Ghanimi
University of Babylon
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Chinese Journal Of Computational Mechanics

Convolutional Neural Network-Based Deep Learning Model Evaluation of Imbalanced Leukocyte Classification Datasets

Jameela Ali Alkrimi[1], Hayder Al-Ghanimi[2], Raja Salih Mohammed[3], Ahmed Al-Ghanimi[4]

[1]College of Dentistry, University of Babylon, Babylon, Iraq
[2]Department of Medical Instrumentation Techniques Engineering, Hilla University College, Babylon, Iraq
[3]Al Mansur Institute of Medical Technology. Middle Technical University Baghdad, Iraq
[4]Department of Computer Science, Faculty of Pharmacy, Babylon University, Babylon, Iraq

ABSTRACT
Leukocyte classification has a significant diagnostic role and provides helpful facts to pathologists for the diagnosis and treatment of many diseases based on the type of cells. As its types have varying ratios in blood causing an imbalanced data set. This is currently a key difficulty in machine learning algorithms, particularly in medical data analysis. This article evaluates the performance classification model using deep learning of convolutional neural network (CNN) algorithms. It isolation, detects and classifies leukocytes in a blood smear microscopic picture. It is applied to identify seven types of leukocytes based on the number of nuclei. The numbers of nuclei were a discriminative feature for both classification and detection. To extract leukocytes and segment the nucleus from the cytoplasm, several image preparation approaches and algorithms are used. After that, a classification model based on CNN was used to classify types of leukocytes. The model experimental achieved 97.86% accuracy for binary classification, and 96.4% for multi-classification. Three metrics have been applied to evaluate the classification model. The threshold metrics, Ranking Metrics and Probability Metrics. The average scorecard criterion of the model is satisfied with an F- score of 0.973, precision of 0.971, and recall of 0.972. While, the process of ranking a list of items based on their relevance in a classification task gives are: Kappa St. 0.964, ROC 0.964, and AUC 0.999. Finally, the measure of the uncertainty of the classification model (MSE) gives 0.002,0.004 for binary and multi-classification.

Keywords: Evaluation Metrics, Convolutional Neural Networks, Deep learning, Imbalance data, Leukocytes, Pre-processing image.

1. INTRODUCTION

Convolutional neural networks (CNNs) has attained a phenomenal performance for several computer vision tasks including object detection, image classification, and semantic segmentation (Zhang, 2022). As it necessitates a thorough evaluation of all the regions across the features, CNNs' significant accuracy improvement comes at the expense of enormous computational complexity (Alkhelaiwi, 2021). There are numerous types of (ML) algorithms that require high-quality datasets for training and testing. It includes correctly labeling images, balanced dataset and sufficient data quantity (Esteva, 2019). Building normal and abnormal leukocyte dataset images requires subject matter experts to classify the images because of the importance and sensitivity in medical diagnosis (Kouzehkanan, 2022). Additionally, the collected medical images are extremely imbalanced because of the difficulty in obtaining them in case studies, which caused an imbalanced dataset (Pereira, 2020). Imbalanced classifications deem a challenge for predictive modeling in machine learning algorithms. In machine learning (ML), unbalanced classes occur often in medical field. Medical datasets usually have a high dimensional variable and an imbalanced class distribution (Khaldy, 2018). There is a discrepancy between the class sizes in this scenario. Intelligent medical diagnosis faces the challenge of classifying datasets that are unbalanced due to the high dimensionality of medical data and the small sample sizes (Zhang, 2022). As a result, the classification performance suffers, leading to flawed clinical recommendations. As a result of the highly skewed distribution of classes resulting in imbalanced dataset classification is a difficult predictive modelling task. Traditional ML mode is evaluation metrics which assumed a balanced distribution of classes perform poorly under these circumstances (Kulkarni, 2022). As most ML algorithms are used for classification assume an equal examples number for each class, predictive modeling of such classificatory is difficult (Allugunti, 2022).

Imbalance of classes happens when a class is highly shown in the dataset as compared to others. Most machine learning algorithms are sensitive

with imbalance data set. Imbalance data set include minority and majority class. The majority classes is a large proportion of the data set, where minority classes is a smaller proportion (Ding, 2022). The classifier has a tendency to be more biased towards the majority class, misclassifying the minority class. The models' prediction performance was poor. The minority class is more important than the majority class since the classification errors are more sensitive for the minority class [9]. Lack of data causes difficulty for the algorithms to identify any evident differences between them [10].

Evaluation of imbalanced classification is very sensitive in machine learning, where choosing the wrong metric to evaluate models is misled by the expected performance of classification model. Therefore, choosing an suitable metric is a common challenge in applied machine learning, particularly in difficult imbalanced classification problems [11]. Due to most of the standard metrics being widely used, it is assumed a balanced class distribution, and typically not all prediction errors, are equal to the imbalanced classification [12].

Leukocytes play a crucial role in the immune system, as it acts as a defending mechanism. Complete blood count uses leukocyte count to reveal the presence of any sly infections hiding inside the body and serve to notify the hematologist [13]. The number leukocyte and the calculation of different leukocyte genre play an important role in clinical tests and diagnoses: they are perceptive of the hematologists and reflect the hidden infection within the body [14]. Diseases including leukemia, immunological disorders, and proven forms of cancer are often diagnosed based leukocyte count [15]. They are two major subsets of leukocytes: Granulocytes and Agranulocytes. Granulocytes known as polymorpho nuclear leukocytes. It is mostly found in the blood. It makes up around 65% of the total WBCs in the blood. It consists of three types Eosinophils, Neutrophils, and Monocytes [16] [17]. Each of them has a different contribution to the immunity system. It is present Innate immunity. It helps to produce antibodies, recognize antigens, and kill the cells that may cause harm to the body [18]. Whereas, Agranulocytes make up 35% of the total WBCs in the blood. It consists Lymphocytes, Monocytes and Macrophages [19]. It is considered an Adaptive immunity system; it has ability prevent disease in the future. Specialized immune cells and antibodies that target and eliminate foreign invaders are involved. It is capable of mounting a new immune response and remembering what those substances look like [20]. Each of them have different features, the geometric and color texture is the main features in distinguish between them [21]. In chronic myeloid leukemia, bone marrow transition is dependent on the donors' granulocytes count, therefore leukocytes has a therapeutic and diagnostic [22] concluded neutrophil count has no significant impact on diagnosis.

The purpose of this research to evaluate CNN classifier algorithm based on deep learning with imbalanced data set for leukocytes datasets using three main metrics which as threshold metrics, Ranking Metrics and Ranking Metrics.

The paper is organized as follows, methodology is the next section, which presents some theoretical background about the concepts used in the study. Followed by a discussion on related studies and how they were used in our synchronization. Subsequently. Then section bellow, describes the obtained results. Later, presents a discussion on the obtained results. Finally, concludes the current study and describes some possibilities for future works.

2. THEORETICAL METHODOLOGY

The purpose of this study is to evaluate classification of CNN algorithms for imbalanced leukocytes data set. The system includes two main parts as show in Figure (1).

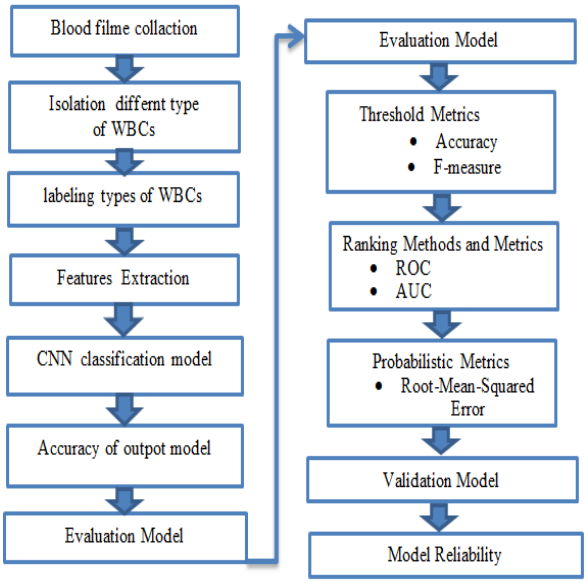


Figure 1. evaluation classification CNN algorithms system

The first part start with capturing blood smear images. These images where collacted from the Department of Laboratories at Marjan Teaching Hospital under the consent of the consultant pathologist Dr. A.Z.Naji. Several image processing algorithms and techniques apply in order to isolate the individual leukocytes from plasma, as shown in Figure 2. After, leukocytes images were labelled into two main groups, the first includes leukocytes with one nuclei named agranuloucytes composing 64 -75% of leukocytes. Whereas in the second group leukocytes containing more than one nuclei

named granulocytes, it is 25-34% of leucocytes [23,24]. Basing on this,extraction of the strong featuers that feed the CNN algorithms in both phase traning and testing in classification task. After that, the output of model optened. The most output of CNN model is accuracy. Due to the accuracy's flaws, which include bias toward data from the dominant class and a lack of discrimination or less informativeness [25]. And in order to make sure that that CNN model is accurately trained and that it outputs the right data and the CNN model's classification is accurate. The classification model must be evaluated.

The secound parts of classification model is evaluation,evaluation caauracy. The importance of model assessment measures is crucial for obtaining the optimal classifier during the training process. Three evaluation Metrics are applining, each Metrics has specific purpose. Where the first used for binary classification, the secound used to prototype selection classifiers during the learning process and used to create an optimized learning model. Finally, to measure the uncertainty of the classifications. In order to insure the CNN model was use the right data, and that the data is accurate then validation steps was appled.

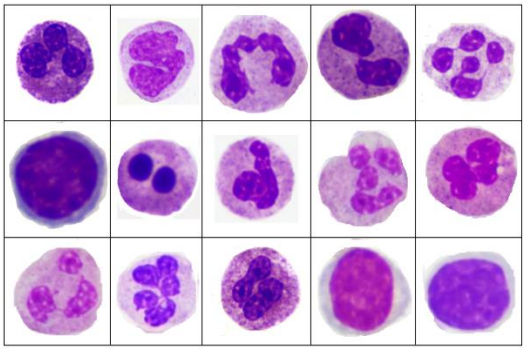


Figure 2. Types of WBCs

2.1 Dataset description

The dataset used in this paper was incloued 1000 WBCs images. It is results of segmantation 180 original blood film images, as described in [25] This dataset is basically divided into two parts. The first part is identified as the training part and the other is identified as the validation part. The training part and the validation part are divided in the ratio 80 : 20. Table 1, show the dataset categories discription, and the images of the dataset samples are shown in Table 2. These data will feed to CNN model.

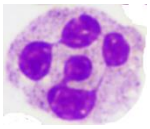
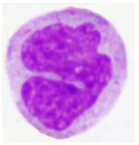
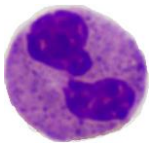
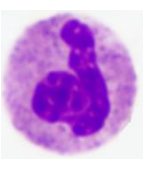
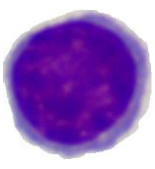
2.2 Featuers Extraction

The main featuers based in this study is number of nuclei in leukocytes,as there are several image preprocessing techniqe used in order to isilation the nuclei. The threshold tachnige as well as different image preprocessing methods such as, binarization, boundary tracing, resizing image and edge smoothing, edge tracing, applying seed filling. Each method has its own purpose in order to obtain the best isolation. Then the numerical feature are obtained. These featuers present in Exel file befor to preper CSV file. The normalization preprocessing dataset do to keep its numerical stability to classification models.

Table 1. The WBCs dataset for CNN classification model

WBCs types	No.of training images	No. of validating images	Total
Agranuloucytes	213	53	266
Granulocytes	587	147	734
Total	800	200	1000

Table 2. The WBCs dataset

WBCs Type	**Granulocytes**			**Agranuloucytes**	
WBCs Name	**Netrophil**	**Eosinophil**	**Basophil**	**Monocyte**	**Lymphocyte**
WBCs Image					
WBCs number	**556**	**153**	**25**	**38**	**228**

2.3 Convolutional Neural Network (CNN)

The literature has a wide range of CNN architectures. However, their fundamental components are very similar. A type of feed-forward neural network named a convolutional neural network often focuses on image processing data [26].The targets of convolutional layer to learn feature representations of the inputs. The CNN structure's architecture successfully preserves the original data's structure and produces a layered representation[27]. Layers of multilevel processing are systematic from left to right in a usual CNN structure [28]. Convolutional, fully connected, pooling and classifying layers are the four main types of layers used in CNN. The basic layers of the architecture are pooling layers and convolutional layers, and they are typically used in the first stages.

3. CLASSIFICATION ACCURACY

Accuracy is the proportion of correctly classified examples over the whole set of examples, So that the result of the classification was obtained through the application of the Weka platform, It is a set of machine learning techniques for data mining problems. The classification don in two phase. Firstly, binary classification of WBCs data into Granulocytes and Agranuloucyte based on the WBCs contains one or more than one nucleus. The secound phase, is multiple classification. The Granulocytes can classify into three types where Agranuloucyte classifyed into two types. The eccuracy for the first phase as show in table 3. Where the results of secound phase present in Table 4.

Table 3. The binary classification WBCs results

WBCs type	CNNs accuracy
Granulocytes	96.4%
Agranuloucyte	98.1%

Table 4. The multiple WBCs classification accuracy

WBCs type	Name of WBCs	Accuracy
Agranuloucyte	Monocyte	97%
	Lymphocyte	95.3%
Granulocytes	Netrophil	95.3%
	Eosinophil	97.1%
	Basophil	96.7%

4. EVALUATION CNN CLASSIFICATION MODEL

This study, several evaluation metrics are applying in order to evaluate the performance classification model and understand whether the model is working well with WBCs data. Becouse of the accuracy is not enough to evaluate a model and to avoid biased. These are the threshold metrics it include F-measure and accuracy, the ranking metrics and methods it include AUC and receiver operating characteristics (ROC) analysis, and the probabilistic metrics it include root-mean-squared error. Each them has a different purpose. These are Threshol metrics to check whether the model satisfies a predefined threshold, Ranking metrics to compare model instances or complete models against each other and Probability metrics to measure the uncertainty of the classifications model.

4.1 Threshold metrics

Threshold metrics are applicable to principally all classifier algorithms out there, To select the best threshold, calculate Precision + Recall and select the threshold of this spot. It is equal to selecting the model with the highest F1 score. Where F1 explene how many instances it classifies correctly. The reselt of evaluation classification modle as show in table 5.

Table 5. The evaluation threshold metrics

Classification type	F-score	Precision	Recall
Bainry classification	0.987	0.977	0.978
Multy classification	0.96	0.966	0.967

4.2 Ranking metrics

Ranking metrics compare model instances or complete models against each other. In this paper use ROC, AUC, confusion matrix and Kappa. ROC it measure of efficient the classification model and ability to separate positive classes from negative classes ,where AUC it measures the quality of the ML classification model's predictions irrespective of what classification threshold is chosen and Kappa it a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy [29]. The Kappa statistics,ROC and AUC reseult show in table 6. give is 0.9786 in binary classification and 0.964 in multy classification, where confusion matrix show in table 7 and 8 bellow.

Table 6. Ranking metrics results

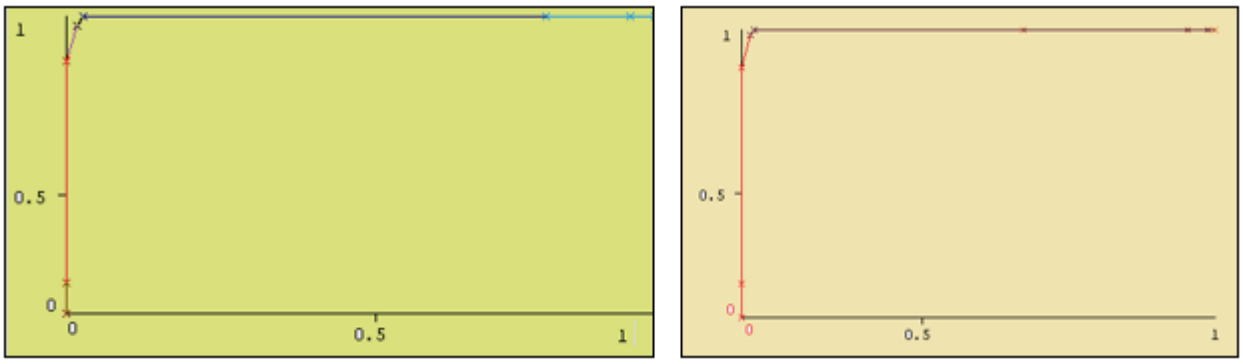
Classification type	Kappa.St	ROC	AUC
Bainry classification	0.963	0.961	0.9995
Multy classification	0.965	0.967	0.9996

Table 7. confusion matrix for binary classification

	Agranuloucytes	Granulocytes	Total
Agranuloucytes	210	56	266
Granulocytes	28	706	734
Total	238	762	1000

Table 8. confusion matrix for multy classification

	Monocyte	Lymphocyte	Netrophil	Eosinophil	Basophil	Total
Monocyte	35	0	0	1	2	38
Lymphocyte	3	220	0	0	5	228
Netrophil	1	0	546	6	3	556
Eosinophil	0	0	2	138	13	153
Basophil	1	0	0	3	21	25
Total	40	220	548	149	44	1000



a. AUC of multy classification b. AUC of binary classification

Figure 3. Area under the ROC Curve

4.3 Probability metrics

Probability metrics are a measure of the uncertainty of the classification model. The evaluate probability foucuse on probability-based ranking performance and probability estimation performance. The probability-based ranking is the Area Under ROC curve, where the Mean Square Error (MSE), is suited to evaluate the probability estimation accuracy performance. Table 9 show result of probability metrics for binary and multy classification.

Table 9. The results of Probability metrics for both classification types

Classification types	ROC	MSE
Binary	0.961	0.004
Multi-	0.966	0.002

5. CONCLUSION:

Nowadays, the main challenges in the medical domain classification is imbalanced data sets class. Because of the accuracy is not enough to evaluate a model. Specifically, in the applied deep learning algorithm results become biased towards the majority class. To alleviate this problem, three evaluation metrics applying in classification Leukocyte using convolutional neural network (CNN) classification algorithm. The classification task based into two steps, binary and multiclass classification. The average results show high accuracy, 97.2 and 96.28 respectively. There are threshold metrics, ranking methods and the probabilistic metrics. Each metric has a different purpose. The average scorecard criterion of the model is satisfied with more than 97%. While, the process of ranking a list of items based on their relevance in a classification task give is 0.9786 in binary classification and 0.964 in multy classification. Finally, the measure of the uncertainty of the classification model (MSE) gives 0.002,0.004 for binary and multy classification.

REFERENCES

1. Zhang, H. Z. “EPSANet: An efficient pyramid squeeze attention block on convolutional neural network”. In Proceedings of the Asian Conference on Computer Vision , pp. pp. 1161-1177.2022.
2. Alkhelaiwi, Munirah, et al. "An efficient approach based on privacy-preserving deep learning for satellite image classification." Remote Sensing 13.11 2221. 2021.
3. Esteva, A., Robicquet, A., Ramsundar, B.,

Kuleshov, V., DePristo, M., Chou, K., & Dean, J. “A guide to deep learning in healthcare”. Nature medicine, 25(1), 24-29. 2019.

4. Kouzehkanan, Zahra Mousavi, et al. "A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm." Scientific reports 12.1 (2022): 1123.
5. Pereira, Rodolfo M., et al. "COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios." Computer methods and programs in biomedicine 194 :105532. 2020.
6. Khaldy, M. A., and C. Kambhampati. "Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset." International Robotics & Automation Journal 4.1: 1-10. 2018.
7. Zhang, Hu, et al. "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network." Proceedings of the Asian Conference on Computer Vision. 2022.
8. Kulkarni, Atharva, et al. "Experimental evaluation of deep learning models for marathi text classification." Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021. Springer Singapore, 2022.
9. Allugunti, Viswanatha Reddy. "A machine learning model for skin disease classification using convolution neural network." International Journal of Computing, Programming and Database Management 3.1: 141-147. 2022.
10. Ding, Hongwei, and Xiaohui Cui. "A clustering and generative adversarial networks-based hybrid approach for imbalanced data classification." Journal of Ambient Intelligence and Humanized Computing : 1-16. 2023.
11. Stefanowski, L. What makes multi-class imbalanced problems difficult? An experimental study. Expert Systems with Applications, pp. 199, 116962. 2022.
12. Salehinejad, Hojjat, et al. "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.
13. Alkrimi, Jameela Ali, et al. "Classification of Imbalanced leukocytes Dataset using ANN-based Deep Learning." Journal of Physics: Conference Series. Vol. 1999. No. 1. IOP Publishing, 2021.
14. Sleeman IV, William C., and Bartosz Krawczyk. "Multi-class imbalanced big data classification on spark." Knowledge-Based Systems 212 (2021): 106598.
15. Stefanowski, Jerzy. "Classification of multi-class imbalanced data: Data difficulty factors and selected methods for improving classifiers." Rough Sets: International Joint Conference, IJCRS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings. Springer International Publishing, 2021.
16. Ding, Hongwei, and Xiaohui Cui. "A clustering and generative adversarial networks-based hybrid approach for imbalanced data classification." Journal of Ambient Intelligence and Humanized Computing 1-16. 2023.
17. Goossen, Linda H. "Pediatric and geriatric hematology and hemostasis." Rodak’s Hematology: Clinical Principles And Applications. 5th ed. Elsevier Inc, 829-46. 2016.
18. Ding, Hongwei, and Xiaohui Cui. "A clustering and generative adversarial networks-based hybrid approach for imbalanced data classification." Journal of Ambient Intelligence and Humanized Computing : 1-16. 2023.
19. Abbas, Abul, Andrew Lichtman, and Shiv Pillai. Cellular and molecular immunology E-book. Elsevier Health Sciences, 2014.
20. Borregaard, Niels, and Jack B. Cowland. "Granules of the human neutrophilic polymorphonuclear leukocyte." Blood 89.10: 3503-3521. 1997.
21. Mason, Hannah D., and Dorian B. McGavern. "How the immune system shapes neurodegenerative diseases." Trends in Neurosciences 2022.
22. Feng, Xiaokai, et al. "Correlation between white blood cell count at admission and mortality in COVID-19 patients: a retrospective study." 2020.
23. Santhosh Krishna, B. V., et al. "Detection of leukemia and its types using combination of support vector machine and K-nearest neighbors algorithm." Next Generation of Internet of Things: Proceedings of ICNGIoT 2021. Springer Singapore, 2021.
24. Baby, Diana, Sujitha Juliet Devaraj, and Jude Hemanth. "Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach." Turkish Journal of Electrical Engineering and Computer Sciences 29.8: 2742-2757. 2021
25. Othman, Amira, Meriem Sekheri, and János G. Filep. "Roles of neutrophil granule

proteins in orchestrating inflammation and immunity." The FEBS journal 289.14 : 3932-3953. 2022.

26. Hossin, Mohammad, and Md Nasir Sulaiman. "A review on evaluation metrics for data classification evaluations." International journal of data mining & knowledge management process 5.2: 1. 2015.
27. Akrimi, Jameela Ali, et al. "Classification red blood cells using support vector machine." Proceedings of the 6th international conference on information technology and multimedia. IEEE, 2014.
28. Gaonkar, Bilwaj, et al. "Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation." Medical imaging 2016: computer-aided diagnosis. Vol. 9785. SPIE, 2016.
29. Deng, Jiajun, et al. "Voxel r-cnn: Towards high performance voxel-based 3d object detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 2. 2021
30. Browne, Matthew, Saeed Shiry Ghidary, and Norbert Michael Mayer. "Convolutional neural networks for image processing with applications in mobile robotics." Speech, Audio, Image and Biomedical Signal Processing using Neural Networks : 327-349. 2008.
31. Kumar, Rajeev, and Abhaya Indrayan. "Receiver operating characteristic (ROC) curve for medical researchers." Indian pediatrics 48: 277-287. 2011.