AN EXPERIMENTAL ANALYSIS OF EVALUATION METRICS FOR MEDICAL IMAGE SEGMENTATION IN DEEP LEARNING APPROACHES

M.A. Sithi Banu[1], A. Dhavapandiammal[2], *P. Kalavathi[3]
Department of Computer Science and Applications,
The Gandhigram Rural Institute (Deemed to be University), Dindigul, Tamil Nadu, India

Abstract: *Deep learning (DL) models enabled fast growth in artificial intelligence research over the last decade, particularly in the medical industry. Several research studies have shown that these algorithms may generate accurate predictions and yield results comparable to medical experts. The DL models must be trained on substantial medical imaging data sets to accurately generate a prediction. Techniques in medical imaging like Ultrasound (US), X-ray, Computerized Tomography (CT), and Magnetic Resonance Imaging (MRI), seek to show the internal structures to detect and treat diseases. Medical Image Processing (MIP) tasks such as detection, segmentation, and classification are employed to analyze three-dimensional (3D) medical images. Segmentation is the most crucial task applicable to identifying abnormalities in medical images by extracting the region of interest. One of the essential steps when developing a successful DL model for segmentation involves evaluating its performance. The evaluation metrics provide insight into the model's performance and facilitate the comparison of various models or algorithms. An outline of several metrics is provided by this experimental analysis, including Accuracy, Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Receiver Operating Characteristic (ROC) curves, Hausdorff distance, Sensitivity, Specificity, etc., are utilized to assess the efficacy of the DL model in the medical image segmentation method.*

Keywords: Deep learning, Medical Image Processing, Evaluation metrics, Segmentation techniques.

1. INTRODUCTION

In the recent past, the widespread technology for processing digital images has greatly facilitated their use in production, research, and everyday life. Artificial intelligence is now playing a role in digital image processing, expanding into algorithmic research and standardization of related image processing technologies. It is extremely significant in many areas, especially in the medical industry [1]. Medical imaging is obtaining images of inside organs for therapeutic purposes, such as illness, detection, and research [2]. The main objective of Medical Image Processing (MIP) is to obtain pertinent data from images acquired from medical imaging devices. Many tasks, including feature extraction, image segmentation, image registration, classification, and visualization, are involved in MIP [3]. One of the best well-known techniques used in medical imaging analysis is deep learning. It has various network designs and is applied in numerous applications in healthcare, specifically with medical imaging. These approaches are continuously used in fields such as early identification of diseases, lowering healthcare personnel' density, reducing expert thoughts, and early treatment [4]. Deep Learning (DL) uses multiple-layered artificial neural networks to extract and analyze complex patterns from massive datasets [5]. Algorithms can be trained to identify and categorize abnormalities in a variety of medical imaging modalities, like MRIs, X-rays, CTs, and Ultrasounds. This process involves teaching algorithms to interpret enormous amounts of data [6]. After training, the system is capable of evaluating fresh clinical images and providing diagnostic data. According to studies, the application of algorithms in DL in MIP has produced promising outcomes, with high levels of accuracy being exhibited in identifying and diagnosing a variety of medical conditions [7]. The Deep Learning Approach (DLA) in MIP is a rapidly emerging area of research, that is extensively employed in medical imaging to determine the existence or non-existence of illness. It will enable the next generation of radiologists to make clinical decisions, reduce medical errors for clinicians, and increase efficiency when processing medical image analysis [8].

Various DLAs have been used in health care namely Boltzmann machines, Autoencoders, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and others [9-11]. CNN is a high-performance technique that excels at processing of image and computer vision (CV) tasks [12]. When DLA is used for medical images, CNNs are appropriate for classification, segmentation, object recognition, registration, and other tasks [13]. It has various types of models namely VGG19 [14], U-net [15], DenseNet [16], ResNet [17], Squeeze-Mnet [18], and AlexNet [19]. These models were utilized in several applications of medical imaging, like object detection, image segmentation, classification, and registration [20], [21]. In DL, algorithms significantly outperform traditional approaches. Evaluation of medical images using DL techniques is considerably superior to traditional methods [22]. Several evaluation measures are routinely used for analyzing the DL model's performance in medical diagnostics, including Accuracy, Dice, Jaccard, Specificity, Sensitivity, Receiver Operating Characteristic (ROC) curves, Confusion Matrices, and others [5], [23]. This work offers a summary of various performance metrics specifically for the segmentation of medical images based on DL approaches.

The structure of this paper is as follows: a study of various kinds of Performance measures for medical image processing in DL is discussed in Section 2, Section 3 describes the experimental analysis. Results and Discussion in Sections 4 and Conclusions and Future Initiatives are covered under Section 5 respectively.

2. Evaluation Metrics in Deep Learning for Medical Image Processing

Quantitative measurements called evaluation metrics are employed to assess the DL model's performance in medical imaging by contrasting anticipated and real results. Numerous metrics are essential for evaluating DL methods in medical imaging [23-25]. Commonly used metrics are displayed in Figure 1 and described below:

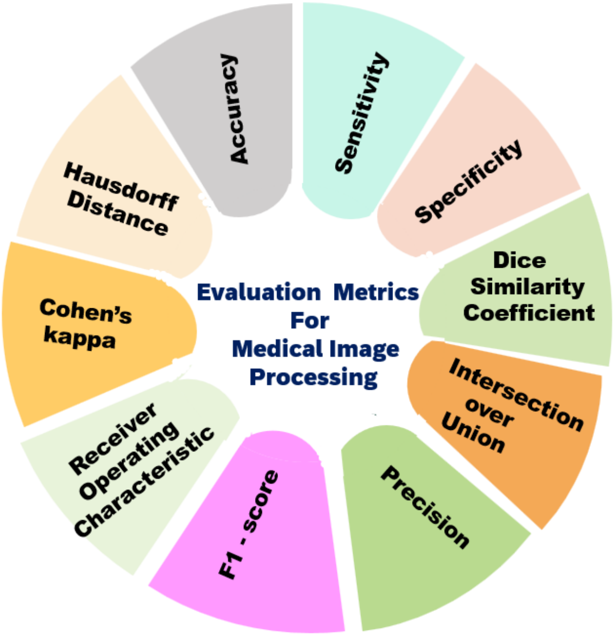
Accuracy
Sensitivity
Specificity
Dice Similarity Coefficient
Intersection over Union
Precision
F1 - score
Receiver Operating Characteristic
Cohen's kappa
Hausdorff Distance
Evaluation Metrics For Medical Image Processing

Figure 1. The Evaluation metrics for Medical Image Processing

- Score range: Evaluation metrics typically yield scores from 0 to 100% (or 0 to 1), where 0 indicates no performance and 1 or 100% signifies perfect performance.
- Common Metrics: Metrics are often defined using True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

- True Positive (TP): When a positive outcome is accurately predicted by the model, the actual result is also positive.
- True Negative (TN): The outcome was negative, as the model had accurately predicted.
- False Positive (FP): When a positive result is predicted by the model but the actual result is negative.
- False Negative (FN): The model predicted a negative result when the result was positive.

2.1 Accuracy

It is an essential component in figuring out how the model is assessed accurately. This measure is the most popular one for measuring the proportion of region of interest (ROI) in medical images. It is computed as the proportion of the total number of predictions the model made to the number of right predictions. Below is the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

2.2 Sensitivity

Another quantitative technique for evaluating a hyperparameter's relative importance to a model's accuracy is sensitivity analysis. It is also termed as recall; sensitivity measures the proportion of TP results among all actual positive cases. It indicates how well the algorithm identifies positive instances. It is known as the true positive rate.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2}$$

2.3 Specificity

This metric assesses the proportion of TN results among all actual negative cases. It displays how well the system can recognize negative instances. Specificity is a term used in medicine that describes the percentage of individuals who do not have the disease and were accurately predicted to not have it. A True Negative Rate (TNR) is another name for it.

$$\text{Specificity} = \frac{TP}{TP+FP} \tag{3}$$

2.4 Dice Similarity Coefficient (DSC)

The similarity between two sets of data is measured by this statistical method. It calculates the overlapping between a gold standard and predicted values pixel-by-pixel. Alternatively referred to as the "Sørensen–Dice coefficient".

$$\text{DSC} = \frac{2TP}{2TP+FP+FN} \tag{4}$$

2.5 Intersection over Union (IoU)

The IoU or Jaccard's Index, evaluates a segmentation model's ability to distinguish items in an image from their backgrounds. turning it into an essential parameter for evaluation. Similar to Dice but penalizes false positives more heavily.

$$\text{IoU} = \frac{TP}{TP+FP+FN} \tag{5}$$

2.6 Precision

The precision of a given algorithmic identification is the proportion of actual positive outcomes to all positive cases. It improves the measurement of the model's capacity to categorize positive samples and is essential to comprehend how accurate the model's favorable predictions are.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

2.7 F1-Score

A measure of statistics called the F1 score combines precision and recall to assess the efficacy of deep learning models. An alternate statistic for evaluating the DL model is the F1 score, which elaborates on a model's performance inside a class rather than evaluating it overall based on accuracy.

$$\text{F1} = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{7}$$

2.8 ROC Curve

A graph known as a Receiver Operating Characteristic (ROC) curve is a useful measure for comparing a deep learning model's performance. It plots the True Positive Rate (TPR) versus the False Positive Rate (FPR) at various classification thresholds in a graph that illustrates how successful the binary classification model performs.

$$\text{TPR} = Sensitivity$$
$$\text{FPR} = 1 - Specificity$$

2.9 Cohen's Kappa

A measurement that accounts for the potential of random agreement when quantifying the agreement between two raters or classifiers. Both multi-class and imbalanced class situations are extremely successfully handled by it.

$$\text{f}_{\text{c}} = \frac{(TN + FN)(TN + FP) + (FP + TP)(FN + TP)}{TP + TN + FN + FP}$$

$$\text{Kap} = \frac{(TP + TN) - \text{fc}}{(TP + TN + FN + FP) - \text{fc}} \tag{8}$$

2.10 Hausdorff Distance (AHD)

Measures the maximum distance between points in the predicted images and ground truth, useful for evaluating boundary accuracy. A similarity metric called Hausdorff distance (HD) can be used in deep learning to compare images and 3D volumes in medical images.

$$\text{AHD (A, B)} = max(d(A, B), d(B, A)) \tag{9}$$

The measurements that were previously mentioned are utilized for evaluating the most important task, segmentation in MIP. It is vital to realize that no one metric can adequately represent a DL method's performance. To accomplish a more comprehensive assessment, a variety of metrics should be used. A given application's needs and clinical relevance should guide the metrics selection process, making sure the assessment corresponds with real-world medical needs [23]. These evaluation techniques are crucial for determining how well DL methods improve medical imaging results.

3. Experimental Analysis

The efficiency of the DL models for MIP particularly for segmentation, is evaluated here. The workflow of this experimental analysis is shown in Figure 2.

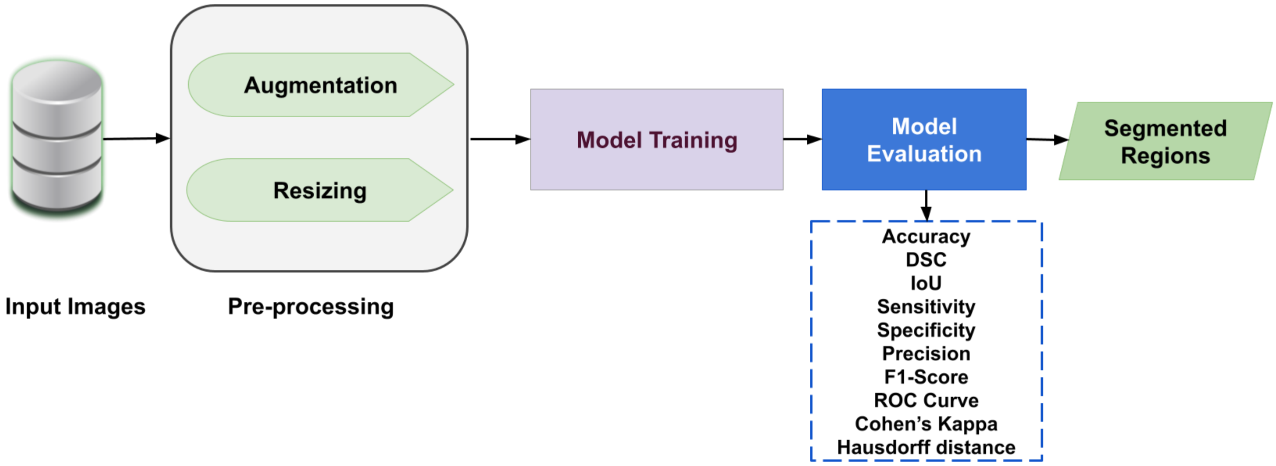

Figure 2. The Workflow Diagram

3.1 Datasets used

We combined datasets from various healthcare sectors that are publicly available to cover a wide range of medical imaging like Chest X-ray [26], CT [27], MRI [28], Ultrasound [29], and Dermatoscopy [30]**.**

3.2 Pre-processing

To train the DL models and medical images after undergoing some preprocessing. The required preprocessing steps for the input images used in the DL model training are augmentation and resizing. The primary limitation of DL models is the availability of huge datasets. The lack of appropriate medical imaging datasets caused many DL algorithms to produce inaccurate results [5] to address these situations, We have preprocessed the dataset by employing data augmentation, which has increased its amount. Resizing the images to 128 ×128 lowers the model's computational expense.

3.3 U-Net Architecture

U-Net is a CNN that is commonly used for image segmentation. It uses an "encoder-decoder" structure that was developed in 2015 for medical image segmentation. The encoder's path and the decoder's path are integrated. The encoding method uses convolutional layers that carry out 3x3 convolutional processes with the ReLU activation function to lessen the spatial resolution of the activation maps. These processes are followed by 2x2 max pooling processes that gradually reduce the spatial proportions of an input image to capture high-resolution, low-level features. The 2x2 transposed convolutions, also known as deconvolutions or upsampling layers, are used to upsample the encoding path feature maps. This increases the graphical resolution of the activation maps and allows the network to rebuild a segmentation map. The network combines feature maps derived from earlier layers with the required decoding path, conserving key spatial information while boosting segmentation precision using skip connections. U-Net concatenation encourages skip connections by concatenating map characteristics, high-level context, and low-level information and allowing for multi-scale segmentation. The last layer uses a 1x1 convolution to assign each feature vector to as many classes as required [26].

4. Results and Discussion

Several experiments are conducted to validate the principles of our evaluation guideline and to show metric behaviors across a range of modalities in medical imaging like Chest X-ray [26], CT [27], MRI [28], Ultrasound [29] and Dermatoscopy [30]. To segment medical images, the analysis made use of various medical imaging modalities. These images were given as a dataset individually to train the DL model. After training, the predicted masks of the respective imaging modalities are displayed in Figure 3. In Table 1, the outcome of the performance measures utilized in the process of segmenting the medical images is given, and the ROC curves of each modality are depicted in Figure 4. The comparison between these metrics is shown in Figure 5. The U-Net architecture [31] is trained with binary cross-entropy as a loss function. Training the model with 0.0001 as a learning rate for a maximum of 50 epochs, and 8 batch sizes was utilized.

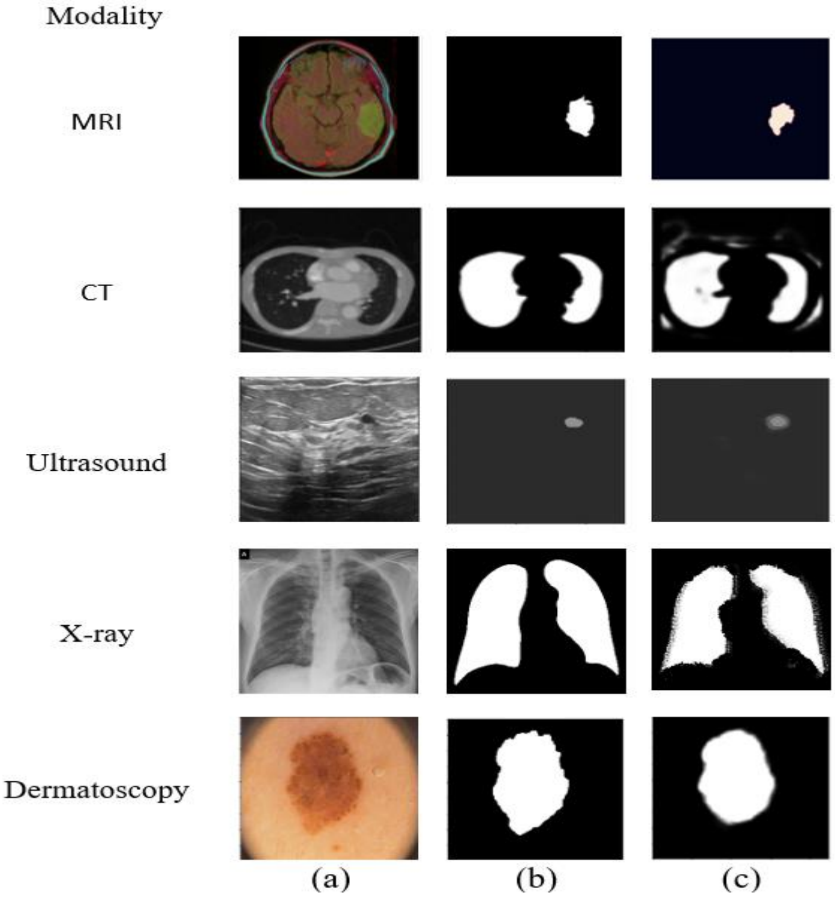

Figure 3. Segmentation of various medical imaging modalities (a) Original Image (b) Ground Truth (c) Predicted Mask

Metrics such as accuracy, consistently show noticeably high scores in any Medical Image Segmentation scenario, as a result, any assessment of segmentation performance should avoid using these metrics. In a medical context, metrics that prioritize true positive categorization alone without accounting for true negative inclusion offer a more accurate depiction of performance. For this reason, in the Medical Image Segmentation field, the DSC and IoU are widely used and prescribed [25]. Here, we segment medical images using a variety of modalities, such as brain MRI [32], liver CT [33], breast ultrasound [34], lung chest X-ray[35], and skin Dermatoscopy [36] with its respective ground truth which is available publicly.

From these experiments for the segregation of medical images, we noticed that Accuracy, Dice, IoU, and F1-Score are the suitable metrics for medical image segmentation.

Table 1. Computed segmented metrics for the chosen sample images are shown in Figure 3.

	Modality	MRI	CT	Ultrasound	X-ray	Dermatoscopy
Metrics	Accuracy	0.97	0.92	0.97	0.92	0.93
	Sensitivity	0.85	0.82	0.75	0.69	0.89
	Specificity	0.86	0.80	0.85	0.89	0.82
	Dice	0.77	0.80	0.64	0.63	0.87
	IoU	0.87	0.85	0.80	0.79	0.93
	Precision	0.84	0.83	0.86	0.82	0.87
	F1-Score	0.75	0.78	0.72	0.73	0.80
	Cohen's kappa	0.82	0.78	0.88	0.75	0.73
	Hausdorff distance	0.89	0.81	0.75	0.73	0.71

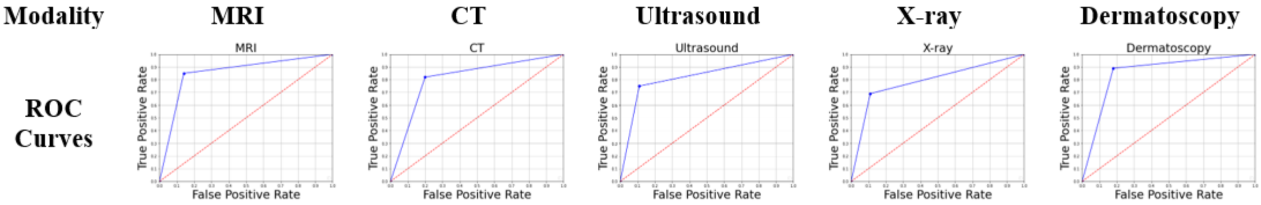

Figure 4. Representation of ROC curves for the Images in various medical imaging modalities.

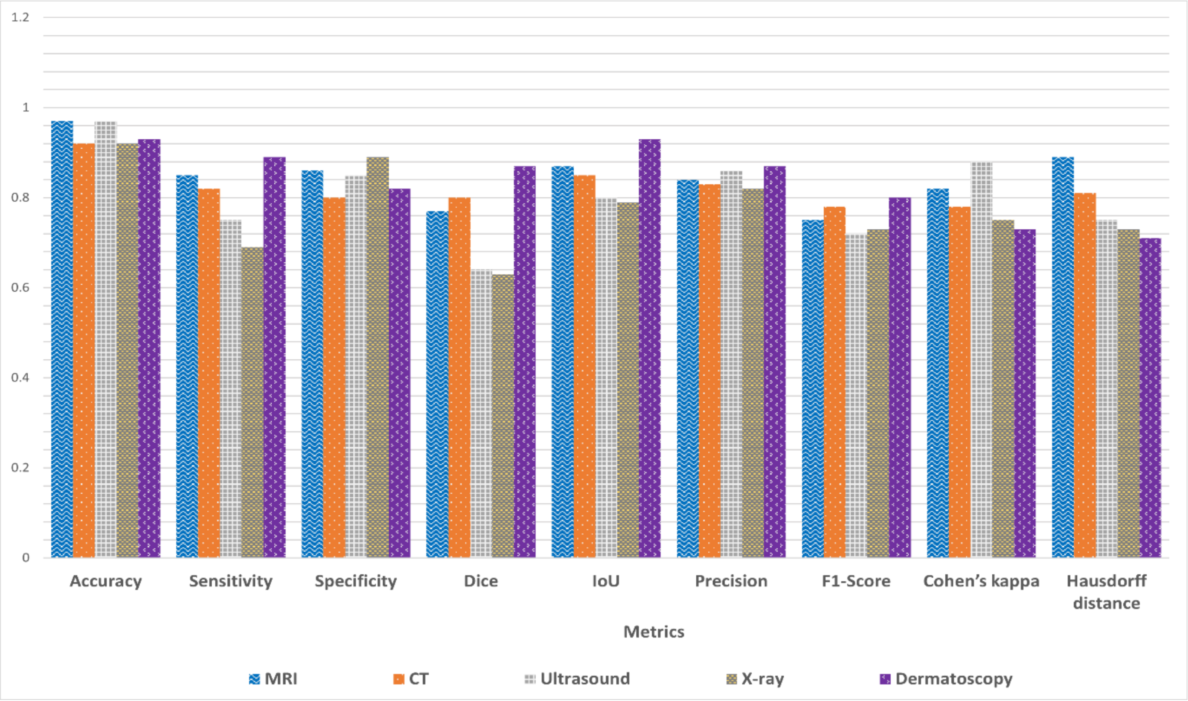

Figure 5. Comparison of metrics used for medical image segmentation.

Segmentation models can be assessed using a variety of metrics. Implementing a model in a clinical setting with only a subset could lead to unexpected outcomes by creating an inaccurate representation of the model's true performance. For this reason, it is critical to combine several indicators and exhaustively interpret the results. There is no evident reason not to include a set of metrics other than space constraints, hiding real performance. Metrics for the various modality of images should be computed independently in addition to interpreting the metrics collectively. The robustness of the model's performance should be evaluated and extra caution should be exercised in circumstances of imbalanced datasets. Many additional metrics in the segmentation of medical images, can be employed based on the study's interpretation priority and research issue. This study aims to ascertain the optimal parameters for achieving a consistent medical image segmentation approach, to optimize model performance.

5. Conclusions and Future Work

DL algorithms have been increasingly popular for medical images in recent years analysis due to their ability to boost both the precision and effectiveness of medical diagnostics, as well as treatment. To improve consistency and provide a uniform medical image segmentation evaluation process, this work concentrated on selecting the most appropriate metrics. Choosing appropriate measurements is a difficult task. This analysis concludes by choosing the metrics that analyze the qualities and requirements for the segmentation of medical images. Several medical datasets were utilized for the experiments. CNN which is based on DL, is used to segment and classify various medical datasets. To compare all potential metrics to be investigated, we employed U-Net architecture in this work.

CONFLICT OF INTEREST

There are no financial conflicts of interest that the authors have acknowledged.

References

[1] Y. Huang, “Overview of Research Progress of Digital Image Processing Technology,” in Journal of Physics: Conference Series, Institute of Physics, 2022. doi: 10.1088/1742-6596/2386/1/012034.

[2] B. Sistaninejhad, H. Rasi, and P. Nayeri, “A Review Paper about Deep Learning for Medical Image Analysis,” 2023, Hindawi Limited. doi: 10.1155/2023/7091301.

[3] G. Dhiman et al., “A Novel Machine-Learning-Based Hybrid CNN Model for Tumor Identification in Medical Image Processing,” Sustainability (Switzerland), vol. 14, no. 3, Feb. 2022, doi: 10.3390/su14031447.

[4] Maad M. Mijwil et al., “From Pixels to Diagnoses: Deep Learning’s Impact on Medical Image Processing-A Survey,” Wasit Journal of Computer and Mathematics Science, vol. 2, no. 3, pp. 9–15, Sep. 2023, doi: 10.31185/wjcms.178.

[5] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, “Medical image analysis using deep learning algorithms,” Front Public Health, vol. 11, 2023, doi: 10.3389/fpubh.2023.1273253.

[6] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” Jul. 01, 2022, Elsevier B.V. doi: 10.1016/j.media.2022.102470.

[7] Z. Amiri et al., “The Personal Health Applications of Machine Learning Techniques in the Internet of Behaviors,” Sustainability (Switzerland), vol. 15, no. 16, Aug. 2023, doi: 10.3390/su151612406.

[8] M. Puttagunta and S. Ravi, “Medical image analysis based on deep learning approach,” Multimed Tools Appl, vol. 80, no. 16, pp. 24365–24398, Jul. 2021, doi: 10.1007/s11042-021-10707-4.

[9] N. Goenka and S. Tiwari, “Deep learning for Alzheimer prediction using brain biomarkers,” Artif Intell Rev, vol. 54, no. 7, pp. 4827–4871, Oct. 2021, doi: 10.1007/s10462-021-10016-0.

[10] M. Gheisari et al., “Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey,” Sep. 01, 2023, John Wiley and Sons Inc. doi: 10.1049/cit2.12180.

[11] M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, and H. Abolghasemi, “A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood

smear images," International Journal of Intelligent Systems, vol. 37, no. 8, pp. 5113–5133, Aug. 2022, doi: 10.1002/int.22753.

[12] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artif Intell Rev, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[13] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Jun. 01, 2020, Springer. doi: 10.1007/s13748-019-00203-0.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[16] C. Szegedy et al., "Going Deeper with Convolutions," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.4842

[17] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," Sep. 01, 2022, MDPI. doi: 10.3390/app12188972.

[18] R. K. Shinde et al., "Squeeze-MNet: Precise Skin Cancer Detection Model for Low Computing IoT Devices Using Transfer Learning," Cancers (Basel), vol. 15, no. 1, Jan. 2023, doi: 10.3390/cancers15010012.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.00915

[21] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," May 01, 2016, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/TMI.2016.2553401.

[22] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, "An overview of deep learning in medical imaging," Jan. 01, 2021, Elsevier Ltd. doi: 10.1016/j.imu.2021.100723.

[23] H. Zhang and Y. Qie, "Applying Deep Learning to Medical Imaging: A Review," Sep. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/app131810521.

[24] Y. H. Nai et al., "Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset," Comput Biol Med, vol. 134, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104497.

[25] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," Dec. 01, 2022, BioMed Central Ltd. doi: 10.1186/s13104-022-06096-y.

[26] "https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database."

[27] "https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images."

[28] "https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri."

[29] "https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset."

[30] "https://www.kaggle.com/code/youssefouama/a-comparative-study-of-skin-cancer-classification/input."

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[32] "https://www.kaggle.com/code/abdallahwagih/brain-tumor-segmentation-unet-dice-coef-89-6/input."

[33] "https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images."

[34] "https://www.kaggle.com/code/vanvalkenberg/segmentation-model-for-breast-cancer/input."

[35] " https://github.com/v7labs/covid-19-xray-dataset."

[36] "https://www.kaggle.com/code/hansern/skin-cancer-segmentation/input."