Predviđanje iskoristivog raspona nijansi sive u slikama medicinske radiologije

Bakotić, Natali

Undergraduate thesis / Završni rad

2023

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Sveučilišni prijediplomski studij računarstva

Završni rad

PREDVIĐANJE ISKORISTIVOG RASPONA NIJANSI SIVE U SLIKAMA MEDICINSKE RADIOLOGIJE

Rijeka, srpanj 2023.

Natali Bakotić
0036522792

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Sveučilišni prijediplomski studij računarstva

Završni rad

PREDVIĐANJE ISKORISTIVOG RASPONA NIJANSI SIVE U SLIKAMA MEDICINSKE RADIOLOGIJE

Mentor: Prof. dr. sc. Ivan Štajduhar

Rijeka, srpanj 2023.   Natali Bakotić
                                                       0036522792

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 7. ožujka 2023.

Zavod: **Zavod za računarstvo**
Predmet: **Uvod u umjetnu inteligenciju**
Grana: **2.09.04 umjetna inteligencija**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Natali Bakotić (0036522792)**
Studij: Sveučilišni prijediplomski studij računarstva

Zadatak: **Predviđanje iskoristivog raspona nijansi sive u slikama medicinske radiologije / Prediction of usable greyscale range in medical radiology images**

Opis zadatka:

Razmotriti upotrebljivost korištenja slika medicinske radiologije za predviđanje parametara koji određuju korišteni raspon nijansi sive. Opisati motivaciju i srodnu literaturu. Analizirati i opisati zadan skup podataka iz KBC-a Rijeka. Implementirati i opisati prikladne metode pripremne obrade i modeliranja slika medicinske radiologije za zadanu namjenu. Osmisliti, opisati i implementirati eksperiment vrednovanja kombinacija implementiranih metoda. Navesti, interpretirati i komentirati rezultate. Zadati i obrazložiti generičke preporuke za buduću primjenu.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 20. ožujka 2023.

Mentor: Predsjednik povjerenstva za
                                                                                  završni ispit:

_____   _____
Prof. dr. sc. Ivan Štajduhar   Prof. dr. sc. Miroslav Joler

IZJAVA

o samostalnoj izradi Završnog rada

Izjavljujem pod punom materijalnom i moralnom odgovornošću da sam ovaj rad izradio/la samostalno te da u njemu nema kopiranih ili prepisanih dijelova teksta tuđih radova, a da nisu propisano označeni kao citati s navedenim izvorom iz kojeg su preneseni.

U Rijeci, _____07.07.2023._____

_____Natali Bakotić_____
(ime i prezime tiskanim slovima)

_____Natali Bakotić_____
(vlastoručni potpis)

SADRŽAJ

Estimating Observation Window Parameters From DICOM Images

Abstract—Digital Imaging and Communication in Medicine (DICOM) is a standard format for storing medical images along with their associated metadata. It supports files obtained by different imaging techniques and with different bit depths. However, most monitors can only display images with a depth of 8 bits, so medical images often need to be converted. To preserve as much information as possible, a windowing area of interest is selected from the entire range of pixel values. This area is defined by two parameters: window level and width, which are often missing, so the conversion cannot be done accurately. In addition, most state-of-the-art deep learning models require images in 8-bit format, so the missing information hinders wider application of artificial intelligence in clinical practise. In this manuscript, we explore the possibility of using a convolutional neural network trained on medical images and metadata to estimate the missing window parameters. The hypothesis was that semantically similar images have a similar area of interest so the window parameters can be estimated directly from the images themselves. The dataset consisted of approximately 24,700 DICOM files with different bit depths, modalities and body parts, obtained from the Clinical Hospital Centre Rijeka PACS. The performance of the predicted windowing parameters was measured by the mean squared error of the true and predicted values and by entropy (amount of preserved information). This performance was compared to windowing with true parameters and methods proposed in "Estimation of Missing Parameters for DICOM to 8-bit X-ray Image Export" by Hržić et al. We show that although it is possible for a neural network to learn windowing parameters from medical imaging data, in terms of entropy it still fails to outperform some simpler methods proposed in the aforementioned study, so we investigate why.

Index Terms—DICOM, Medical Imaging, Radiology, Artificial Intelligence, Deep Learning, Convolutional Neural Network, Mean Squared Error, Entropy

I. INTRODUCTION

In recent years, the rapid growth of computing power has enabled major breakthroughs in the field of artificial intelligence (AI), leading to its integration into our daily lives, and it is becoming an invaluable tool in clinical practice. Various AI techniques, such as fuzzy expert systems, evolutionary computation and hybrid intelligent systems have been explored, but artificial neural networks (ANNs) are most commonly used [1]. With advances in deep learning and computer vision, there is growing iinterest in the application of AI technologies in radiology, as the amount of radiological imaging data grows faster than the number of medical professionals who can interpret it [2]. Some advantages of using AI in medical applications are that it can recognise patterns and relationships in medical data that are too complex for humans and provide quantitative assessment with reproducible results, and in case of deep learning, the features to be learned by the machine learning (ML) model do not need to be computed before

training. This can, in some cases, remove the human bias of selecting features that deemed to be important and instead keep only those features that have been calculated as truly relevant [3].

An important factor that has led to greater applicability of AI in radiology is the standardisation of medical images. As medical technology advanced, the volume of medical image data with different characteristics increased exponentially, and picture archiving and communication systems (PACS) were developed in order to provide economical storage, easy retrieval and availability of medical images at different locations [4]. There are several widely used file formats for medical images [5], but this paper will focus on Digital Imaging and Communication in Medicine (DICOM) format. The DICOM file consists of an image (raw pixel data) and metadata stored in the header of the file, which provides additional information about the image, such as patient information, information about the institution the image belongs to and properties of the device the image was captured on [6]. The DICOM format supports images obtained through various imaging techniques, such as Magnetic Resonance (MR), Computed Tomography (CT), Computed Radiography (CR), Nuclear Medicine (NM) and so on [7], [8]. These techniques result in images with different bit depths, such as 10, 12 or 16 bits, but most devices can only display images with 8 bits depth. For this reason, medical images are transformed by enhancing the area of the image that contains information important to medical professionals. This is achieved by defining useful pixel intensity range using two parameters: window level and width, which determine the lowest and highest pixel intensity that will be displayed. Values within the window area are mapped to range 0-255 and linearly interpolated, while those below the window threshold are mapped to 0 (pure black) and those above to 255 (pure white). Additionally, the human eye can only distinguish between 700 and 900 shades of gray [9], which means that there would be no benefit of displaying images with bit depth greater than 10 in their true depth. Furthermore, studies have shown that the human eye can only distinguish about 30 shades of gray in images displayed on monitors [10], [11], which means that applying windowing makes it easier to see important information in the image as the difference between different shades would appear more significant. The flowchart in Fig. 1 shows the process of applying windowing to an image.

Despite technological advances and the increase in medical imaging data, only a small subset of it is actually useful for the development of ML algorithms [12], [13]. Some of the challenges are that most of this data is kept private in hospitals due to patient privacy concerns and that the data is often mislabeled or there is crucial information missing. Fur-
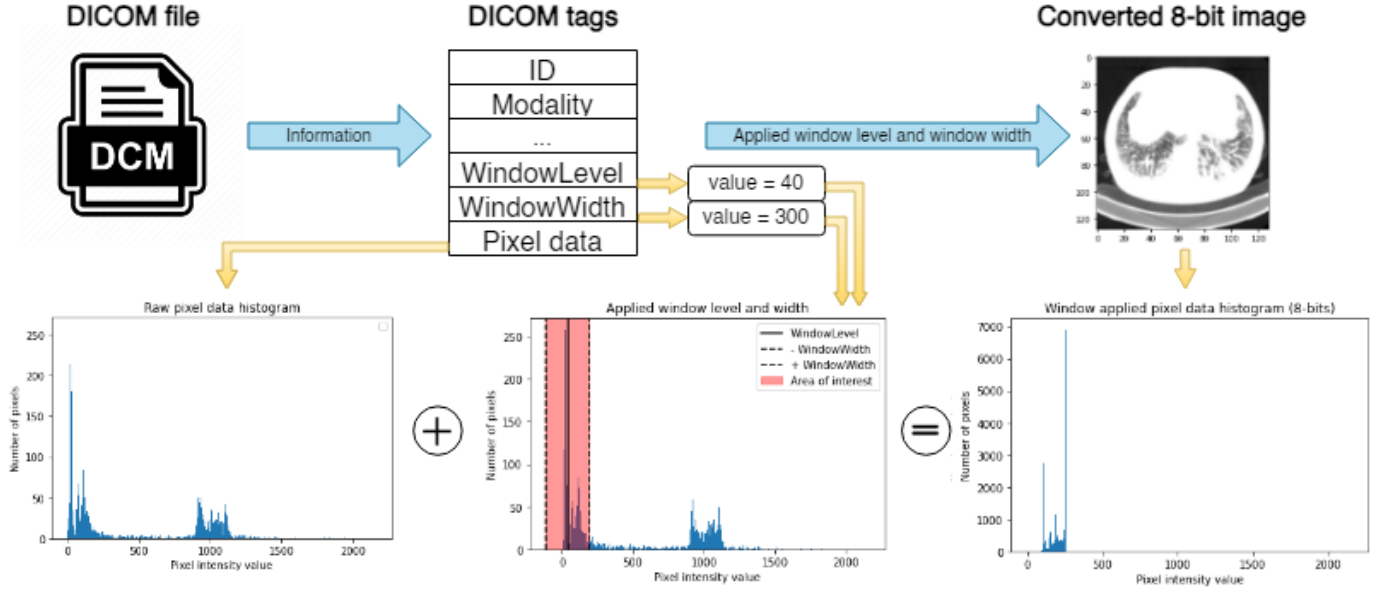
Fig. 1: Flowchart of steps taken to transform original image to 8-bits.

thermore, most publicly available ML algorithms are intended for use with common image formats (jpg, png, bmp, etc.), which typically have a depth of 8 bits. Therefore, medical image data must be preprocessed accordingly [14], which is not possible if important information such as imaging method or window parameters is missing. The need for high quality datasets is why many researchers are exploring methods for preparing available data, such as using autoencoders to reduce metadata complexity [15] and methods for filling in missing information, such as the study by Hržić et al. [16] that aims to estimate missing window level and width parameters.

While this paper has the same goal, there are some key differences between these studies. Firstly, in the study by Hržić et al. windowing parameters were chosen based on different policies that do not take true windowing parameter values into account, while this study explored the feasibility of using deep learning to predict windowing parameters. The initial assumption was that semantically similar images have similar values for window level and width, and therefore a deep neural network could learn features for predicting them. Semantic similarity refers to extraction of higher level features from images, meaning it is about interpreting the objects and relationships between them, unlike visual similarity which compares purely visual features, such as colour, shape, texture and so on [17]. Secondly, the dataset used in the study by Hržić et al. consisted solely of X-ray images of 12 or 16-bit depth, while in this study the images had various modalities and bit depths ranging from 8 to 16 bits. Finally, this study also aimed to explore whether and how much other metadata information, namely bit depth, modality and body part examined, contribute to the accuracy of window parameter predictions.

II. MATERIALS AND METHODS

A. Dataset

The utilised dataset was originally used in the study by Napravnik et al. [15]. It originates from the PACS system in the Clinical Hospital Centre (CHC) Rijeka. There are approximately 24,700 images with grayscale pixel data that were obtained in one of these six modalities – MR, CT, CR, NM, X-ray Aniography (XA) and Radio Flouroscopy (RF). They appear in similar ratios (approximately 4,000 images per modality). The metadata for these images contained the following information: *Modality* (one of the six possible values), *WindowCenter* and *WindowWidth*, *Rows* and *Columns* (original pixel ratio of the image), *BodyPartExamined* (there were 28 different values present, but many samples had this information missing), *HighBit* (bit depth of the image) and *StudyDescription*. When the images were received, they were scaled to dimensions 512 x *Columns*/512 and saved as numpy arrays, which significantly reduced the number of required preprocessing steps. The dataset was divided into train set (approximately 18000 images), validation set (2000 images) and test set (5000 images).

1) Metadata preprocessing: *Rows* and *Columns* tags were dropped as they are not relevant to the experiment and the images were already scaled. *StudyDescription* tag was also dropped since it contains natural language which is challenging for machines to process. Next, some instances of *WindowCenter* and *WindowWidth* contained two numeric values, so only the first of them was kept. Since images had different bit depths, their values were scaled to range 0-255. Lastly, tags *Modality*, *BodyPartExamined* and *HighBit* describe a distinct category an image belongs to so they were one-hot encoded.

2) Image preprocessing: The images initially had different shapes and bit depths, but most deep learning libraries require the images to have the same shape, bit depths and to be

saved in a more widely used image format. Therefore, the images were scaled to 8-bit depth, resized to 128 x 128 pixels using bilinear interpolation in the original aspect ratio, so zero padding was applied where necessary. Next, the images were saved in grayscale png format. Normally windowing would have been applied before scaling to 8-bit depth to retain as much useful information as possible, but because the idea of this study was that required information is missing, this step was omitted.

When inspecting the data, some images appeared to be corrupted, more specifically they contained only zeroes even in the original numpy array format, so they were removed from the dataset. The dataset was standardized by subtracting the mean value of the train dataset from all data and dividing it by standard deviation, as recommended by many state-of-the-art convolutional neural network (CNN) models [18]–[21], and the same values were later used for validation and test set preprocessing.

B. Convolutional neural network

Model building was carried out in two stages. In the first stage, the goal was to evaluate model performance based solely on the image input, while in the second stage, metadata features were added in order to asses whether they contribute to prediction accuracy. The influence of each metadata feature was tested separately and in different combinations. Structures of both final models can be seen in Fig. 2. As the division of train, validation and test sets can greatly affect the model performance score, an average performance of the models, measured by mean squared error metric, was determined using 10-fold cross validation, so models with performance close to their mean scores were used for the experiment.

1) Image only input CNN: Several model structures were tested with the following hyperparameters to be adjusted: input image size (64 x 64, 128 x 128 and 256 x 256), number of convolution layers (2 to 5), the number of filters per layer (32, 64, 128, 256 and 512), regularization method (none, batch normalization and L2, for which regularization factors 0.0001, 0.001, 0.01 and 0.1 were tested), the number of dense layers following the convolution part of the network (from 0 to 3), units per dense layer (from 32 to 512), learning rate (0.01, 0.001, 0.0001) and finally batch size (32, 64, 128, 256). Rectified Linear Unit (ReLU) was used as the activation function and He was used as kernel initializer, as research shows it performs better than default glorot initializer when ReLU is used as the activation function [22]. Mean squared error was used for the loss function and metrics, and Adam was used as the optimizer. Early stopping was used to find the optimal number of epochs. The best combination of hyperparameters was found with grid search method. Different models were trained and tested on the same data, and mean squared error was used as the comparison metric. The final model takes images of 128 x 128 pixels as input. Convolution part of the network is 5 layers deep, each layer having kernel size 3 and 32, 64, 128, 128 and 256 filters respectively. Each convolution layer is followed by MaxPooling layer and flattened after the last one. There are two dense layers after

convolutional part, both with 64 units. Regularization method is L2 with factor 0.01. There are two dense output layers, for *WindowCenter* and *WindowWidth*, with one unit and linear activation function. The chosen value of learning rate is 0.0001 and batch size is 256. The structure can be seen in Fig. 2a.

2) Image and metadata input CNN: The second model used the same starting structure and hyperparameters, with the addition of a second input of size 35 (vector of one-hot encoded metadata features) followed by 2 to 5 dense layers (possible number of units 32, 64, 128, 256). The output of these layers is concatenated with the output of convolution layers. Instead of broad grid search, only minor variations of the original structure were tested, and grid search was applied to the newly added dense layers. The final network structure is the same as the first model, with the only difference being that one dense layer following the convolution was removed. The number of dense layers for metadata was 4, with number of units 32, 64, 128 and 256. The structure can be seen in Fig. 2b.

C. A brief overview of scaling methods

In order to compare the performance of different scaling methods, it will be explained how each of them works.

The method that is usually employed in practice, here named **window_scaling**, uses parameters *WindowCenter* and *WindowWidth*, which are stored in tags *WindowLevel* and *WindowWidth* in the header of a DICOM file, to define an area of interest and maps it to range 0-255 (8-bit grayscale image). Value mapping is performed using the following equation:

$$PixelIntensity = \begin{cases} 0, & \text{if PixelValue} \leq \text{-WindowWidth} \\ 255, & \text{if PixelValue} \geq \text{+WindowWidth} \\ x, & \text{otherwise} \end{cases}$$
(1)

The values of *-WindowWidth*, *+WindowWidth* and *x* are calculated as:

$$-WindowWidth = WindowCenter - \frac{WindowWidth}{2}$$
(2)

$$+WindowWidth = WindowCenter + \frac{WindowWidth}{2}$$
(3)

$$x = \frac{PixelValue - WindowCenter + \frac{WindowWidth}{2}}{WindowWidth} \cdot 255$$
(4)

Methods **predicted_image_scaling** and **predicted_metadata_scaling** work the same, but the difference is that *WindowCenter* and *WindowWidth* are not taken from DICOM metadata, but predicted using a CNN. The difference between these two methods is that **predicted_image_scaling** used parameters predicted with a model with only image input, and **predicted_metadata_scaling** those with a model that takes additional metadata inputs (bit depth, modality and body part examined). The following methods, while not the
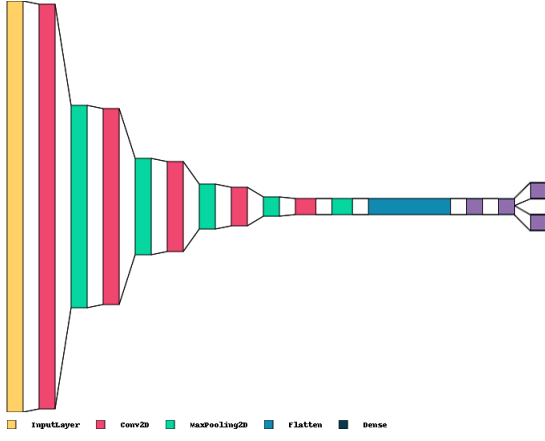
(a) Image only input CNN structure.
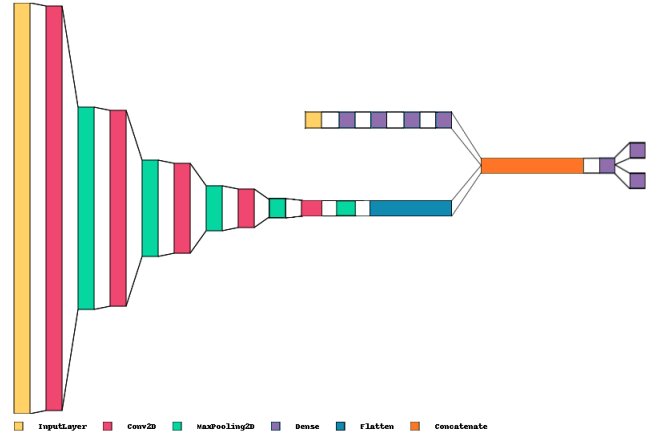
(b) Image and metadata input CNN structure.

Fig. 2: CNN structure comparison.

focus of this study, were implemented to provide additional context for evaluation:

- **max_bit_scaling** linearly scales the entire pixel intensity range of an original image to 8-bit values. More precisely, 0 is mapped to 0, the maximum value in the pixel intensity histogram is scaled to 255 and the values in between are linearly interpolated to fit the range. No windowing is applied. This is the method that was used to create the image dataset for model training.
- **min_max_scaling** sets the lower boundary of the window to the minimum non-zero pixel intensity value in a histogram that holds some data, and the upper boundary is set to the maximum pixel intensity.
- **percentile_scaling** sets the lower boundary to 10-th percentile of pixel intensity histogram, and upper boundary to 90-th percentile, as these values had the best performance in the baseline study.
- **max_peak_scaling** searches for peaks in the raw pixel intensity histogram. "Max peak" is defined as the maximum pixel intensity in a range. The histogram of pixel intensities is first denoised by eliminating all pixel intensity values that occur less than the 25-th percentile of all pixel intensity occurrences in the histogram. The lower/upper boundary of the window is set to the first/last pixel in the histogram after/before which 10 consecutive pixel intensity values are different from 0. After selecting lower and upper boundary, they are moved to maximum peak in range [*StartIndex*, *StartIndex* + *SearchSize*] for the lower boundary and [*EndIndex* - *SearchSize*, *EndIndex*] for the upper boundary, where *SearchSize* = 32. Values of denoising percentile and *SearchSize* were determined in the original experiment.

D. Evaluation

The evaluation process was carried out in two parts. First, mean squared error of true and predicted window parameters was used to evaluate the performance of different models. Ablation study was performed to determine the contribution of each metadata feature to accuracy of the model. As this study builds on the paper by Hržić et al., methods proposed in it were implemented and compared to deep learning approach. Second, pixels' local entropy evaluation, which was used in the baseline study, was employed to further evaluate the methods, as shown in Fig. 3. Pixels' local entropy images and their histograms were used to measure exactly how much information is retained after scaling images. A pixel's local entropy is a value that represents "level of complexity" of an area of an image, and it is calculated as Shannon's entropy of the observed pixel in relation to intensity values of the surrounding pixels [23]. The chosen surrounding area size was 3, like in the baseline study.

After scaling images using different methods, their pixels' local entropy images were compared to pixels' local entropies of the original images. The only difference between the baseline and this study is that two additional scaling methods were added: window scaling using predictions of a model with image only as input, and window scaling using the model that takes other metadata into account. Local pixel's entropy images were compared using the following four methods:

- **Histogram intersection**. This algorithm calculates a value that represents how much two histograms overlap. It is often used in image classification [24], [25] and therefore assumed to be an appropriate measure of image similarity.
- **Hellinger distance** is a metric used to quantify the difference between two probability distributions. It is also used in classification problems [26].
- **Bhattacharyya distance** measures dissimilarity between distributions of features based on Bhattacharyya coefficient [27], [28].
- **Mean Entropy Distance (MED)**, which was employed in the baseline study as a means to determine the difference of entropy values in relation to a pixel's position. It is calculated as mean squared error between each pixel of an original and a scaled image's local pixels' entropy image. Essentially, mean entropy distance is higher the more information a scaled image lost in different areas. This was done because histogram comparison methods
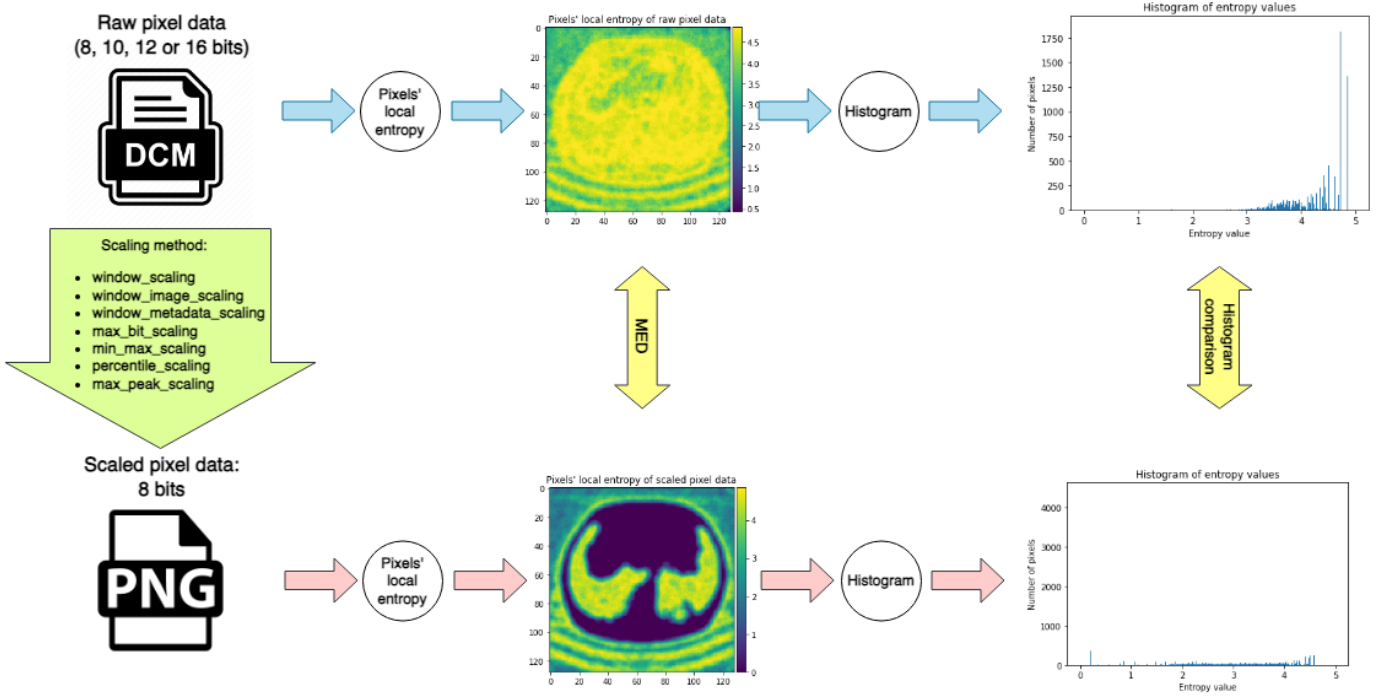
Fig. 3: Flowchart of steps taken to evaluate scaling methods.

don't take the position of the pixels' entropy values into account and only compare the number of occurrences of each value, but this information is crucial because some areas of the image contain information that is more important.

After comparing pixels' local entropy images, one-way ANOVA test [29] was used to determine if there are any statistically significant differences between the means of utilised method results. Tukey's honestly significant difference (HSD) test [30] was performed afterwards to determine the significance between each pair of groups.

III. RESULTS AND DISCUSSION

A. Mean squared error of windowing parameters and ablation study

Mean squared error analysis shows the difference between true windowing parameter values and predicted or calculated ones. The results of each method are shown in table I. It can be seen that the model with metadata input performed better than one using only the image, the one using a combination of *HighBit* and *Modality* was only slightly worse, while the other models had similar performance. It is interesting that some models with less metadata inputs performed worse than the model with only image input. Scaling methods not relying on windowing parameters performed significantly worse for this metric, especially max_peak_scaling, due to them determining windowing parameters using different policies and not learning them directly.

B. Pixels' local entropy and histogram analysis

The results of windowing method performances for pixels' local entropy evaluation are given in Fig. 4. Subfigures a), b),

c) and d) shows results of histogram intersection, Hellinger distance, Bhattacharyya distance and MED respectively. Each diagram shows minimum, maximum, median, first and third quartile of the observed metric. For histogram intersection, the goal is to obtain higher value, and the maximum value possible to obtain is 256. For the remaining metrics, lower scores are better.

Scaling methods window_scaling, max_bit_scaling, min_max_scaling, percentile_scaling and max_peak_scaling were already compared in the baseline study, so similar results were expected. The performance and results of these methods will not be discussed as they are not the focus of this paper, but there are two interesting observations worth mentioning. The first is that the similarity of max_bit_scaling and min_max_scaling performance is due to many images containing a completely white letter ("R" or "L") that specifies the side of the projection and many images have a white border around the edges, which causes min_max_scaling to behave exactly like max_bit_scaling as the entire range of values is present in most images. The second is that percentile_scaling method obtains the best scores across all metrics except MED, which is due to this method preserving information evenly distributed over the whole image, but losing information in points of interest where it is more important.

Comparing the results of predicted_image_scaling and predicted_metadata_scaling, it can be seen that predicted_image_scaling performs only slightly better than max_bit_scaling and min_max_scaling across all metrics except MED, where it performs the worst among all methods. This is because incorrect windowing parameter predictions lead to completely cropping out areas of interest. However,

TABLE I: Mean squared error (MSE) of models with different combinations of metadata input labels.

| Method	*WindowWidth* MSE	*WindowCenter* MSE
image only window crop	19.26	20.86
image, *HighBit*, *Modality*, *BodyPartExamined* window crop	16.33	19.07
image, *HighBit*, *Modality* window crop	18.20	21.12
image, *HighBit*, *BodyPartExamined* window crop	19.58	21.88
image, *Modality*, *BodyPartExamined* window crop	20.18	20.18
image, *HighBit* window crop	21.99	24.09
image, *Modality* window crop	21.56	24.17
image, *BodyPartExamined* window crop	20.32	21.79
max_bit_scaling	41.40	35.52
min_max_scaling	45.07	32.08
percentile_scaling	77.91	73.13
max_peak_scaling	94.96	93.15
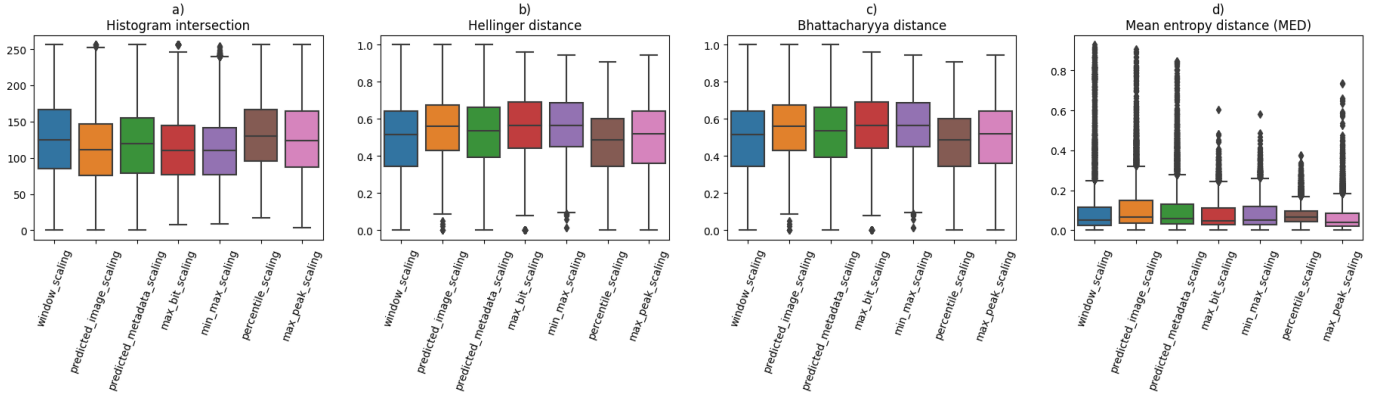
Fig. 4: Box plot of evaluation results of different methods: a) histogram intersection, b) Hellinger distance, c) Bhattacharyya distance, d) MED.

predicted_metadata_scaling appears to perform much better. In terms of histogram intersection, Hellinger and Bhattacharyya distance, it performs slightly worse than window_scaling due to prediction error, while in terms of MED it still performs worse than all methods except predicted_image_scaling and percentile_scaling.

Next, one-way ANOVA test was done to determine the statistical significance of these differences. All metrics resulted in p-value < 0.05, which means that a significant difference exists, so Tukey's HSD test was performed to determine the significance of difference between each pair. The results are shown in table II. It can be seen that in terms of histogram intersection there is no significant difference between predicted_metadata_scaling and max_bit_scaling, and between predicted_image_scaling and min_max_scaling. In terms of distance metrics, there is no significant difference between max_bit_scaling and predicted_image_scaling, min_max_scaling and predicted_image_scaling. However, the most important observation is that the difference between window_scaling and predicted_metadata_scaling is still significant, meaning that windowing parameters predicted using a CNN are still worse than max_peak_scaling in relation to window_scaling.

C. Error analysis

After comparing the methods, we looked into errors that occur in the model predictions. We noticed that the largest errors occurred for images of 16-bit depth, usually wholebody imaging or renal scans. Comparing them side by side, we found that initial scaling of those images using max_bit_scaling resulted in images containing virtually no information, as can be seen in Fig. 5.
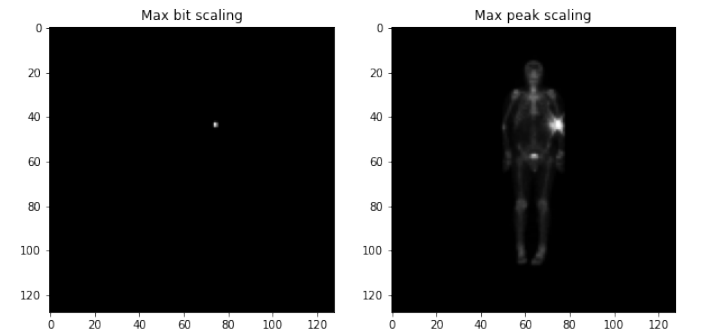
Fig. 5: Comparison of wholebody images scaled with max_bit_scaling and max_peak_scaling.

This proves the need for developing windowing methods before using medical datasets for deep learning, as naive preprocessing fails to capture all information. These findings led us to train an additional model, with the same structure as described in materials and methods, but which used images scaled with max_peak_scaling instead of max_bit_scaling since they would contain more relevant information. However,

TABLE II: Results of Tukey's HSD test. p-values without significant difference are written in bold.

Method #1	Method #2	Histogram intersection	Hellinger distance	Bhattacharyya distance	MED
max_bit_scaling	max_peak_scaling	0.0	0.0	0.0	0.0002
max_bit_scaling	min_max_scaling	**0.435**	**0.8708**	**0.8708**	**0.9537**
max_bit_scaling	percentile_scaling	0.0	0.0	0.0	**0.9997**
max_bit_scaling	predicted_image_scaling	0.0339	**0.988**	**0.988**	0.0
max_bit_scaling	predicted_metadata_scaling	**0.5944**	0.0003	0.0003	0.0
max_bit_scaling	window_scaling	0.0	0.0	0.0	0.0
max_peak_scaling	min_max_scaling	0.0	0.0	0.0	0.0
max_peak_scaling	percentile_scaling	0.0	0.0	0.0	0.001
max_peak_scaling	predicted_image_scaling	0.0	0.0	0.0	0.0
max_peak_scaling	predicted_metadata_scaling	0.0	0.0	0.0	0.0
max_peak_scaling	window_scaling	0.0094	**0.6008**	**0.6008**	0.0
min_max_scaling	percentile_scaling	0.0	0.0	0.0	**0.8013**
min_max_scaling	predicted_image_scaling	**0.925**	**0.999**	**0.999**	0.0
min_max_scaling	predicted_metadata_scaling	0.0041	0.0	0.0	0.0
min_max_scaling	window_scaling	0.0	0.0	0.0	0.0
percentile_scaling	predicted_image_scaling	0.0	0.0	0.0	0.0
percentile_scaling	predicted_metadata_scaling	0.0	0.0	0.0	0.0
percentile_scaling	window_scaling	0.0	0.0	0.0	0.0
predicted_image_scaling	predicted_metadata_scaling	0.0	0.0	0.0	0.0
predicted_image_scaling	window_scaling	0.0	0.0	0.0	**0.0734**
predicted_metadata_scaling	window_scaling	0.0	0.0	0.0	0.0438

these models did not perform any better despite changing the preprocessing method. The errors that occurred were similar, except the max_peak_scaling models showed an even bigger tendency to underestimate the parameters, which is in practice worse as it can crop out areas of interest. Further analysis showed that most incorrect predictions were made on images with larger values of windowing parameters (over 10,000, or around 70 when scaled to range 0-255), which are fewer in the dataset. Error distribution of image only and image with metadata input models are shown in Fig. 6.

IV. CONCLUSION

To summarise, the aim of this study was to utilise deep learning methods to estimate window level and width parameters for scaling images stored in DICOM format, since they are often missing, but are crucial for correct conversion to an 8-bit format that can be displayed on most monitors. On top of that, scaling methods using predicted parameters were compared to methods proposed by Hržić et al. The deep learning approach consisted of learning window parameters solely from images and from images combined with additional metadata from DICOM files, namely bit depth, body part examined and modality. The results show that window parameters predicted using the model with combined input are slightly better than those of the model learning only from images, but both models outperform windowing parameters estimated using other methods. However, they still performed significantly worse than true window parameters or max_peak_scaling (which is much simpler to apply) when it comes to preserving local pixel entropy, likely due to incorrect predictions cropping out areas of interest. The largest errors occurred for images with 16-bit depth that were rarely present in the dataset.

Nevertheless, the study showed that CNNs can learn windowing parameter, given the sufficient amount of data and model complexity, as even a fairly small model trained in a few minutes is able to achieve results better than naive scaling methods in terms of estimating exact window parameter values. Some ideas worth exploring are whether a better model architecture exists that further improves prediction accuracy, such as employing transfer learning to benefit from models trained on larger datasets, as medical imaging datasets are very small in comparison. Considering that most errors occurred for the rarely present or unique data, balancing the dataset in terms of value range in future studies might be beneficial. However, the most important research might be developing preprocessing methods, such as max_peak_scaling, that preserve more information despite missing metadata information to properly prepare data for future machine learning projects, as poor data quality leads to poor model performance.

REFERENCES

[1] A. N. Ramesh, C. Kambhampati, J. R. Monson, and P. J. Drew, "Artificial intelligence in medicine," *Annals of the Royal College of Surgeons of England*, vol. 86, no. 5, p. 334–338, 2004.
[2] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, p. 500–510, 2018.
[3] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *RadioGraphics*, vol. 37, no. 2, p. 505–515, 2017.
[4] R. H. Choplin, n. J M Boehme, and C. D. Maynard, "Picture archiving and communication systems: an overview," *RadioGraphics*, vol. 12, no. 1, pp. 127–129, 1992.
[5] M. Larobina and L. Murino, "Medical image file formats," *Journal of Digital Imaging*, vol. 27, p. 200–206, 2014.
[6] M. Mustra, K. Delac, and M. Grgic, "Overview of the dicom standard," in *2008 50th International Symposium ELMAR*, vol. 1, 2008, pp. 39–44.
[7] W. E. Brant and C. A. Helms, *Fundamentals of Diagnostic Radiology*, 3rd ed. Lippincott Williams & Wilkins, 2007.
[8] K. K. Shung, M. Smith, and B. M. Tsui, *Principles of Medical Imaging*. Academic Press, 2012.
[9] T. Kimpe and T. Tuytschaever, "Increasing the number of gray shades in medical display systems—how much is enough?" *Journal of Digital Imaging*, vol. 20, p. 422–432, 2007.
[10] E. Kreit, L. M. Mäthger, R. T. Hanlon, P. B. Dennis, R. R. Naik, E. Forsythe, and J. Heikenfeld, "Biological versus electronic adaptive coloration: how can one inform the other?" *Journal of the Royal Society Interface*, vol. 10, 2013.

Image only input model error distribution

WindowCenter predictions

WindowWidth predictions

(a) Image only input model error distribution.

Image and metadata input model error distribution

WindowCenter predictions

WindowWidth predictions

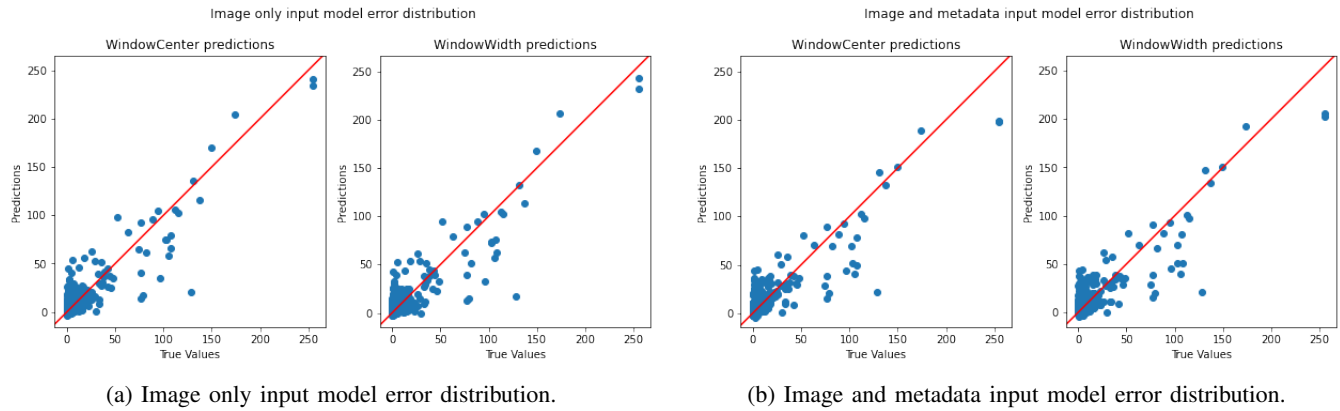(b) Image and metadata input model error distribution.

Fig. 6: Error distribution of window parameter predictions. The values were scaled to range 0-255. The red line represents matching predictions and true values.

[11] D. M. Glover, W. J. Jenkins, and S. C. Doney, *Modeling Methods for Marine Science*. Cambridge University Press, 2011.

[12] H. Harvey and B. Glockers, *A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology*, E. R. Ranschaert, S. Morozov, and P. R. Algra, Eds. Springer International Publishing, 2019.

[13] P. M. A. van Ooijen, *Quality and Curation of Medical Images and Data*, E. R. Ranschaert, S. Morozov, and P. R. Algra, Eds. Springer International Publishing, 2019.

[14] S. Masoudi, S. A. A. Harmon, S. Mehralivand, S. M. Walker, H. Raviprakash, U. Bagci, P. L. Choyke, and B. Turkbey, "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research," *Journal of Medical Imaging*, vol. 8, no. 1, 2021.

[15] M. Napravnik, R. Baždarić, D. Miletić, F. Hržić, S. Tschauner, M. Mamula, and I. Štajduhar, "Using autoencoders to reduce dimensionality of dicom metadata," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 1–6.

[16] F. Hržić, M. Napravnik, R. Baždarić, I. Štajduhar, M. Mamula, D. Miletić, and S. Tschauner, "Estimation of missing parameters for dicom to 8-bit x-ray image export," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 1–6.

[17] M. Antunes, "Semantic vision agent for robotics," Ph.D. dissertation, University of Aveiro, 01 2011.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, 2017.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[22] ——, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," pp. 1026–1034, 2015.

[23] Y. Wu, Y. Zhou, G. Saveriades, S. Agaian, J. P. Noonan, and P. Natarajan, "Local shannon entropy measure with statistical tests for image randomness," *Information Sciences*, vol. 222, pp. 323–342, 2013, including Special Section on New Trends in Ambient Intelligence and Bio-inspired Systems.

[24] E. Cheng, N. Xie, H. Ling, P. R. Bakic, A. D. Maidment, and V. Megalooikonomou, "Mammographic image classification using histogram intersection," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010, pp. 197–200.

[25] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, vol. 3, 2003, pp. III–513.

[26] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the hellinger distance," *Information Sciences*, vol. 218, pp. 146–164, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025512004069

[27] K. G. Derpanis, "The bhattacharyya measure," *Mendeley Computer*, vol. 1, no. 4, pp. 1990–1992, 2008.

[28] F. J. Aherne, N. A. Thacker, and P. I. Rockett, "The bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.

[29] K. T. Kyun, "Understanding one-way anova using conceptual figures," *kja*, vol. 70, no. 1, pp. 22–26, 2017. [Online]. Available: http://www.e-sciencecentral.org/articles/?scid=1156679

[30] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.

PROŠIRENI SAŽETAK NA HRVATSKOM JEZIKU

Uvod

U posljednjih nekoliko godina, umjetna inteligencija (UI) se sve češće koristi u medicinskoj radiologiji. Razlog tome je povećanje obima medicinskih podataka u digitalnom obliku i činjenice da UI modeli imaju sposobnost naučiti kompleksne, ljudima neuočljive uzorke i veze među njima.

Jedan od razloga za sve češću primjenu UI je standardizacija medicinskih slika. Jedan od najkorištenijih formata za medicinske slike je Digitalne slike i komunikacija u medicini (DICOM), koji osim same slike u zaglavlju datoteke ima spremljene i metapodatke s dodatnim informacijama o slici. DICOM format podržava slike dobivene raznim tehnikama i u različitim dubinama boja. Međutim, većina zaslona može prikazati samo slike dubine 8 bitova, pa medicinske slike moraju biti prilagođene za takav prikaz. To se najčešće provodi koristeći dva parametra, razina i širina prozora (eng. *window level and width*), koji su pohranjeni u metapodacima. Postupak prilagodbe se provodi tako da se pomoću razine i širine prozora odredi početak i kraj korisnog raspona nijansi sive, a sve izvan tog područja se prikazuje kao crno ako su vrijednosti manje od početka raspona, odnosno kao bijelo ako su veće od kraja raspona. Vrijednosti unutar korisnog raspona preslikavaju se u raspon 0-255 i linearno interpoliraju, što dovodi to znatno jasnijeg prikaza slike zbog smanjenog raspona koji se treba prikazati.

Osim toga, većina suvremenih modela za duboko učenje zahtjeva da slike budu u nekom od češće korištenih formata, što znači da nedostatak informacija za prilagodbu slika sprječava širu primjenu UI u medicini. Zbog toga se istražuju brojne tehnike za obradu medicinskih podataka. U ovom radu ispitana je mogućnost korištenja konvolucijske neuronske mreže trenirane na slikama medicinske radiologije za predviđanje parametara koji nedostaju. Izvorna hipoteza je da semantički slične slike (tj. slike koje sadrže slične objekte), imaju sličan iskoristivi raspon boja, te stoga neuronska mreža taj raspon može naučiti iz samih slika. Osnova za ovo istraživanje je članak *„Estimation of Missing Parameters for DICOM to 8-bit X-ray Image Export"*, pa su metode predložene u njemu korištene za usporedbu.

Materijali i metode

Skup podataka sastojao se od približno 24,700 DICOM slika s različitim dubinama boja sive, tehnikama dobivanja slike i dijelovima tijela, a dobiven je iz Kliničkog bolničkog centra (KBC) Rijeka. Podaci su bili podijeljeni u trening, testni i validacijski skup.

Metapodaci su obrađeni tako da se izbacilo tekstualno polje opis studije (*StudyDescription*) i polja o izvornim dimenzijama slike (*Rows, Columns*), a metapodaci dubina boja, modalitet i proučeni dio tijela (*HighBit, Modality, BodyPartExamined*), koji opisuju kategoriju slike, su *one-hot*

kodirani. Vrijednosti razine i širine prozora (*WindowCenter*, *WindowWidth*) skalirani su u raspon 0-255. Slike su obrađene transformacijom u 8-bitni format s dimenzijama 128 x 128 piksela u izvornom omjeru pa je crna podstava dodana gdje je bilo potrebno. Za duboko učenje su standardizirane oduzimanjem srednje vrijednosti i dijeljenjem sa standardnom devijacijom trening skupa.

Izrađene su dvije konvolucijske neuronske mreže, gdje je jedna od njih učila parametre samo iz slika, a druga je koristila i metapodatke koji opisuju kategoriju slike. Doprinos svake kategorije metapodataka ispitan je u ablacijskoj studiji. Obje mreže na izlazu imaju dvije vrijednosti, razinu i širinu prozora, skaliranu u raspon 0-255. Srednja uspješnost modela utvrđena je desetostrukom unakrsnom provjerom, a za mjerilo točnosti je korištena srednja kvadratna pogreška između stvarnih i predviđenih vrijednosti.

Evaluacija se sastojala od provjere srednje kvadratne greške između stvarnih i predviđenih vrijednosti te od usporedbe entropije (količine sačuvanih informacija) i histograma izvornih slika i slika transformiranih različitim metodama. Metode iz članka „*Estimation of Missing Parameters for DICOM to 8-bit X-ray Image Export*" također su implementirane i njihovi su rezultati uspoređeni s predviđenim vrijednostima parametara.

Rezultati i rasprava

Analiza srednje kvadratne pogreške pokazala je da su vrijednosti predviđenih parametara znatno bliže stvarnim vrijednostima pohranjenim u metapodacima. Ablacijska studija pokazuje da je model s dodanim metapodacima nešto bolji od onoga koji je učio samo iz slika. Međutim, analiza entropije i usporedba histograma pokazale su da kada model griješi, gube se gotovo sve korisne informacije te su u tim slučajevima neke metode iz članka *Estimation of Missing Parameters for DICOM to 8-bit X-ray Image Export*" znatno uspješnije u očuvanju informacija.

Analizom grešaka utvrđeno je da korištenje naivnih metoda za prilagodbu slika neke od njih transformira tako da se gube gotovo sve informacije, pa modeli iz njih ne mogu učiti. Za takve je slike potrebno primijeniti naprednije metode pripremne obrade. Nadalje, utvrđeno je da modeli češće podcjenjuju vrijednosti parametara, što je u primjeni lošije jer se tako može promašiti veliki dio korisnog raspona. Najveća greške učinjene su na 16-bitnim slikama, obično slike cijelog tijela i skeniranje bubrega, s vrijednostima parametara iznad 10,000 (skalirano u raspon 0-255 oko 70), koje su znatno rjeđe u skupu podataka.

Zaključak

Konačno, eksperimentom je utvrđeno da neuronske mreže uistinu mogu naučiti vrijednosti za određivanje korisnog raspona nijansi sive, ali za određene kategorije slika znatno griješe. Budući

da su to slike velikih dubina boja koje se teško mogu pretvoriti u 8-bitni oblik bez pripadajućih parametara o korisnom rasponu, u budućim istraživanjima potrebno je istražiti metode za uspješniju transformaciju takvih slika kako bi bile korisnije za duboko učenje. Nadalje, budući da je u ovom radu korištena vlastita struktura neuronske mreže, u budućim radovima bilo bi zanimljivo proučiti može li korištenje složenijih modela, npr. primjenom prijenosnog učenja, povećati točnost predviđenih parametara. Razlog tomu je što su modeli za prijenosno učenje trenirani na jako velikim skupovima podataka (milijuni slika), a skupovi podataka slika medicinske radiologije su u usporedbi jako maleni. U tome je slučaju potrebno istražiti potrebni postupak za prilagodbu slika kako bi odgovarale obliku koji modeli za prijenosno učenje podržavaju.
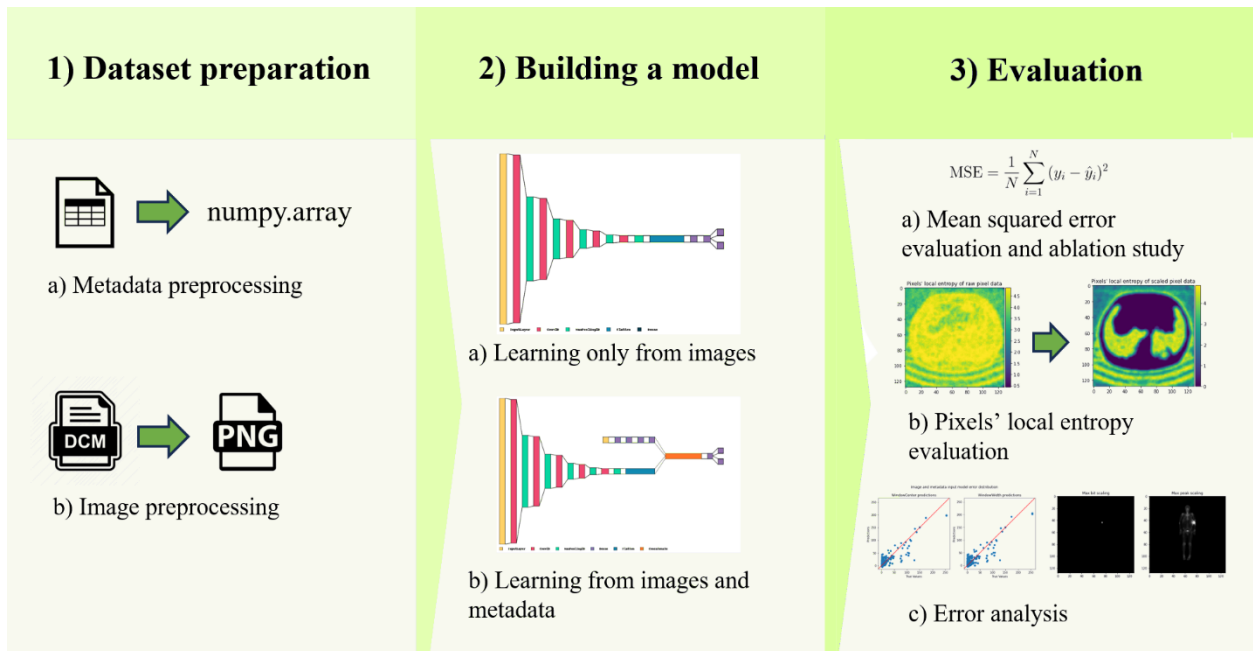
KLJUČNE RIJEČI

DICOM, medicinske slike, radiologija, umjetna inteligencija, duboko učenje, konvolucijska neuronska mreža, srednja kvadratna pogreška, entropija

KEYWORDS

DICOM, Medical Imaging, Radiology, Artificial Intelligence, Deep Learning, Convolutional Neural Network, Mean Squared Error, Entropy

PRILOZI

Grafički sažetak

GitHub repozitorij

https://github.com/nbakotic/zavrsni