Detection of Landmarks in X-Ray Images Through Deep Learning

Mauro Fernandes[1], Vitor Filipe[1,2], António Sousa[1,2], and Lio Gonçalves[1,2(✉)]

[1] School of Science and Technology University of Trás-os-Montes and Alto Douro (UTAD), Vila Real 5000-811, Portugal
al31090@utad.eu, {vfilipe,amrs,lgoncalv}@utad.pt
[2] INESC Technology and Science (INESC TEC), Porto 4200-465, Portugal

Abstract. This paper presents a study on the automated detection of landmarks in medical x-ray images using deep learning techniques. In this work we developed two neural networks based on semantic segmentation to automatically detect landmarks in x-ray images, using a dataset of 200 encephalogram images: the UNet architecture and the FPN architecture. The UNet and FPN architectures are compared and it can be concluded that the FPN model, with IoU=0.91, is more robust and accurate in predicting landmarks. The study also had the goal of direct application in a medical context of diagnosing the models and their predictions. Our research team also developed a metric analysis, based on the encephalograms in the dataset, on the type of Mandibular Occlusion of the patients, thus allowing a fast and accurate response in the identification and classification of a diagnosis. The paper highlights the potential of deep learning for automating the detection of anatomical landmarks in medical imaging, which can save time, improve diagnostic accuracy, and facilitate treatment planning. We hope to develop a universal model in the future, capable of evaluating any type of metric using image segmentation.

Keywords: Automated Landmark Detection · Deep Learning · UNet Architecture · FPN Architecture

1 Introduction

The automatic detection of reference points in medical x-ray images has assumed, in the recent past, an increasing importance at the diagnostic level, being considered as essential in a fast, accurate and effective diagnosis so that the treatment of a patient, human or non-human, is done in time in order to increase the success rate of the same.

The fact that it is an automatic but reliable detection increases the accuracy and effectiveness of the diagnosis - it reduces the time needed to analyze the images, so that each specialist can have his or her own opinion on a given image, and minimizing any potential human error. Furthermore, being accurate

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 264–279, 2024.
https://doi.org/10.1007/978-3-031-60665-6_20

and done automatically, it allows accurate and quantitative measurements to be made, providing highly objective information as an aid to diagnosis. These metrics can include measurements of distance, angles, and even proportions, greatly facilitating timely treatment.

In addition, the automatic detection of landmarks is also especially important in situations where the progression of diseases is being followed and monitored, and the effectiveness of treatments is being assessed. The primary benefit of this diagnostic technique is undoubtedly the standardization of medical image analysis. It ensures that within the medical community, where different specialists may typically provide varying interpretations of images and their conditions, consistent results are obtained. This standardization enhances the reliability and uniformity of the analysis, leading to more accurate diagnosis. By doing this, the ability of artificial intelligence and machine learning to train increasingly better models to detect these points of interest is also increased, thus boosting the automation of clinical processes and improving patient care by allowing the time needed for the interpretation of exams to be reduced and thus allowing health professionals to focus on more complex tasks and direct interaction with patients.

This project will apply the trained models to a specific case - the Mandibular Occlusion, which relates the upper tooth structure to the lower tooth structure. This, as we will refer to later, has three types of classification - Class I (Normal), II (Prominent Maxilla) and III (Protruding Mandible), which is obtained by analyzing the YEN angle, which is the angle formed by the Sella point, the midpoint of the anterior maxilla and the center of a circumference inside the Gnatio.

Throughout this paper we will talk a little about the state of the art in this area, then explain how we carried out our work and end with a presentation of the results and their discussion, along with the conclusions we were able to draw.

2 Related Work

The identification of landmarks in X-ray images through Deep Learning has been an essential task for several clinical procedures, such as alignment analysis, distance measurement and surgery planning. In recent years, the use of Deep Learning techniques has driven significant advances in this area, providing accurate and automated results.

However, and according to [1], the exclusive use of Deep Learning for automatic landmark detection in medical radiography images, despite having a relatively high accuracy also has a rather high risk of bias. It is therefore necessary to increase robustness and to develop new techniques that use Deep Learning as a basis to increase the reliability of studies involving this method.

One of the notable advances in this field is related to the detection of anatomical points in dental radiographs [2]. Using convolutional neural networks (CNNs), it has been possible to train models capable of accurately identifying landmarks such as the root apex, cusps, and reference lines. These automated models have shown results comparable and, in some cases, even superior to traditional methods, saving time and resources for dental professionals. Another approach to the

problem proposed the use of convolutional neural networks combined with recurrent neural networks (RNNs) to identify specific anatomical landmarks, such as the tooth root and periodontal structures. This approach showed promising results, outperforming traditional manual identification methods, and providing a detailed analysis of tooth morphology.

Another proposal presents a novel learning framework for detecting anatomical landmarks in medical images, including panoramic radiographs [8]. This work uses deep learning and reinforcement learning to optimize the target accuracy in detecting multiple landmarks. The proposed method is evaluated on two datasets, one of prenatal ultrasound and another of cephalometric x-rays, demonstrating improved training stability and enhanced localization accuracy. The promising results of this study suggest that the framework can be effectively applied to landmark detection in panoramic radiographs, providing a valuable tool for health professionals in diagnosing and planning dental treatments.

In another study, an attention mechanism that incorporates multidimensional information and separates spatial dimensions is proposed [7]. This novel method is evaluated on a dataset of pelvic radiographs and demonstrates excellent performance compared to other landmark detection models. The results obtained indicate the effectiveness of this attention mechanism in accurately identifying anatomical landmarks in the pelvic region, providing a valuable tool for healthcare professionals in this field, making this automated approach one that promises to facilitate and expedite diagnosis and treatment planning, improving the efficiency and quality of patient care.

Another notable advance is focused on X-ray imaging of the spine, in which a new specialized landmark detection network is presented [6]. The proposed network consists of two stages and uses techniques such as random spinal slice augmentation and CoordConv to improve detection accuracy. These techniques allow to identify the centers of lumbar vertebrae and corners of vertebrae more accurately in X-ray images. The results obtained demonstrate high accuracy in detecting these landmarks, which can assist radiologists in analyzing lumbar x-ray images, providing important information for diagnosis and treatment planning related to spinal conditions.

Also, in the realm of spine evaluation, another study argued that the balance of the human spine depends on the accurate measurement of sagittal radiographic parameters [3]. In this study, deep learning models capable of automatically locating anatomical landmarks and generating radiographic parameters from lateral radiographs of the spine were developed. Based on a large number of annotated images, the models achieved high accuracy in locating landmarks in different areas of the spine. Moreover, the radiographic parameters predicted by the models showed a significant correlation with the actual values. Comparing the performance of the models with human intelligence, it was deduced that the deep learning model achieved results comparable to those of physicians on several parameters. This automated approach provides an accurate analysis of spinal alignment, assisting medical professionals in the diagnosis and treatment of spinal diseases.

Another application demonstrates that the use of deep learning on chest X-Ray has shown promising results on several tasks [4]. Image-level classification and regression enable the detection and diagnosis of lung diseases, such as pneumonia and lung cancer, with increasing accuracy. Segmentation aids in the precise delineation of anatomical structures, enabling the identification and measurement of lesions with greater accuracy. In addition, landmark localization and synthetic image generation contribute to the improvement of surgical planning techniques and procedure simulation. Domain adaptation enables knowledge transfer between different datasets, extending the practical applications of deep learning models. However, it is important to note that the use of public datasets has limitations, as they do not always reflect the diversity of clinical cases encountered in practice. Therefore, future research is needed to fill existing gaps and develop clinically useful systems that can handle the complexity and variability of chest radiography.

A different but interesting approach was debated in [5], where they aimed to develop an automated calibration system to make linear measurements in lateral cephalometric radiographs more efficient. The system was based on deep learning algorithms and previous medical knowledge of a stable structure, the anterior cranial base (Sella-Nasion). A two-step cascaded convolutional neural network was constructed based on 2860 cephalograms to locate the Sella, Nasion and 2 ruler points in regions of interest. The accuracy of automated landmark localization, ruler length prediction, and linear measurement based on automated calibration was evaluated with statistical analysis, with this high accuracy attributed to the inclusion of diverse training data and the application of prior medical knowledge of anatomically stable structures.

In summary, the use of deep learning techniques has revolutionized landmark identification in radiographies. Automated anatomical landmark detection in various radiographic modalities, such as dental, panoramic, cephalometric, thoracic, and orthopedic, has shown accurate and efficient results. These automated approaches can save time, improve diagnostic accuracy, and facilitate treatment planning, benefiting both healthcare professionals and patients.

3 Development

The main goal of this study was to develop a neural network that could automatically detect landmarks in X-ray images through Deep Learning. Two approaches were then developed in order to create two models capable of successfully predicting landmarks in a new untrained X-ray image.

3.1 Dataset

Our dataset consisted of 200 completely random encephalogram images from a pack of 400 images available on Kaggle, with all of them being right lateral x-ray images of the skull of different people with different anatomies, as depicted in Fig. 1. The aim of our work, by using only half of the available dataset, was

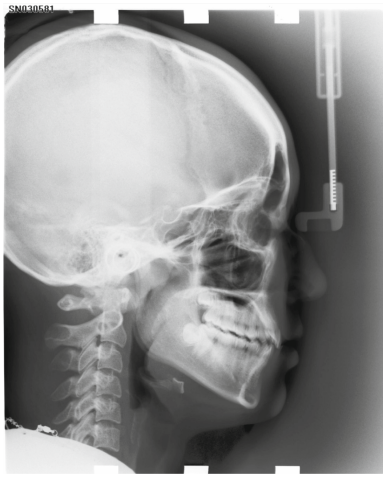

Fig. 1. Encephalogram.

to try to develop a model that would achieve reliable results even in a context where there wasn't a very large dataset for training. The original images were all 1935×2400 in size and .jpg format. They were divided, for training, validation and testing purposes, into a 60/20/20% ratio (120 images for training, 40 images for validation and 40 images for testing).

Before starting the pre-processing of the 200 images, we marked 6 landmarks on all of the images as shown in Fig. 2. The points in question were the S point (Sella), M point (Midpoint of the Anterior Maxilla), the G point (Center of a circle inside the Gnatio point), the N point (Tip of the Nose), SL (Superior Lip) and IL (Inferior Lip). The choice of some of these points was not random, but rather to later calculate the YEN angle to determine the type of mandibular occlusion present in each encephalogram, as we will discuss in the results section. The resulting files, in .json format, were used to collect the coordinates of each of the points in a .csv file, using a script that performs the extraction of the points' coordinates in each of the images.

3.2 Pre-Processing

After extracting the coordinates and grouping them into .csv files for training, validation and testing purposes, a preprocessing is performed on the images in order to facilitate the further processing of the model.

Firstly the original images are loaded and transformed to grayscale and pixels are normalized. The coordinates of the landmarks are used to define rectangular regions in the mask, with dimension 20×20, defining the pixels corresponding to that region as having the value "1". After creating the masks, we resized images and masks to the model's input size, 256×256.

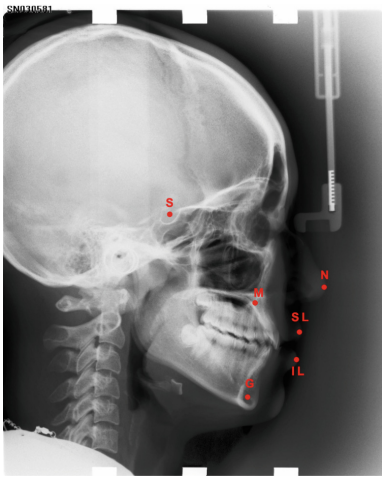

Fig. 2. Marking Landmarks in LabelMe.

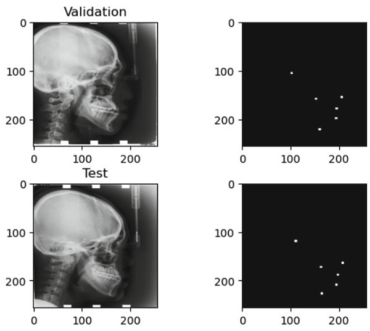

Fig. 3. Verification of the Pre-Processing.

Finally, vertical stacking of the training and validation images and masks is performed in order to create appropriate input matrices for the models.

To verify if the pre-processing was done well, the first validation image and mask is previewed, as well as the first image and mask from the testing set, to ensure that it conforms to the processing done to the image and mask sets, as represented in Fig. 3.

3.3 UNet Architecture

The first architecture we developed throughout this project was an architecture called "UNet" - a U-shaped architecture with a downsampling path and an upsampling path, with skip connections between the encoding and decoding paths to allow for fine detail reconstruction, as Fig. 4. This architecture presents fundamental characteristics for a project with the dimension of this study, due

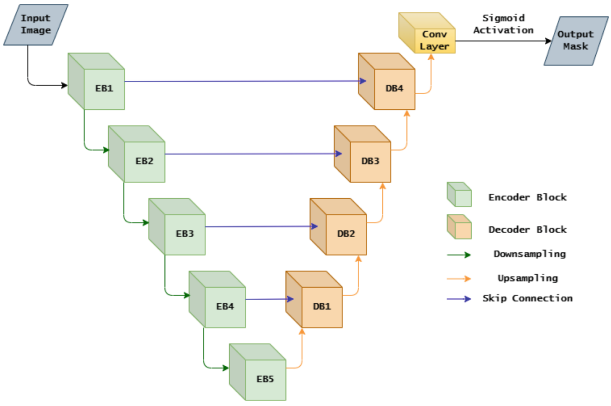

Fig. 4. Proposed UNet architecture.

to its ease of training models with small datasets. In addition, and as mentioned above, the reconstruction of fine details with skip connections and also the possibility of reducing overfitting during training with the introduction of dropout layers. It is an architecture widely used in image segmentation tasks and, with special relevance for our study, in medical image segmentation, thanks to its proven flexibility and performance.

The downsampling path is characterized as the part of the model responsible for reducing the spatial dimensions of the input images, as well as extracting fine details. In our case, this path starts by defining a block (which we will refer to as Block 1). In this block, we have the following layers:

- Two 2D convolution layers with a ReLU activation function and a "he-normal" weight initialization with a Batch Normalization layer in between.
- A dropout layer.
- A Max Pooling layer

Blocks 2,3,4 and 5 follow, with the structure very similar to Block 1. However, the number of filters doubles with each new block in order to capture more complex features as the network goes deeper.

However, there are some differences in certain blocks, relative to Block 1, that are important to note:

- In block 4 there is no dropout layer added, in order to extract features critical to the network's ability to train and learn.
- Blocks 4 and 5 do not have the Max Pooling layer.

The output of each of the blocks is fed to the input of the next block, and each of these outputs is also characterized by having a skip connection to the upsampling path.

The upsampling path is characterized by being the part of the model responsible for reconstructing the spatial resolution of the images from the features

obtained in the downsampling path - features obtained by passing between the different blocks and by skip connections.

In our case, the upsampling path starts with a block, which we will call block 6:

- A transposed convolution layer.
- This is followed by a concatenation with the output of a coding block that received the information from the skip connections, allowing the fusion of low resolution information with fine details.
- Two 2D convolution layers with a ReLU activation function and a "he-normal" weight initialization with a Batch Normalization layer in between.

Blocks 7, 8 and 9 are exactly the same as block 6, with a difference in that each layer of transposed convolution and concatenation is followed by convolution followed by BatchNormalization. In the specific case of block 9, a flag is set in order to control whether there is an additional layer of convolution and Batch-Normalization - the variable "output" is optional, and if it is "True", we have 3 layers followed by convolution interleaved with BatchNormalization, and if it is "False", this additional layer is not included. This last step assumes special relevance because it is a step that can affect the performance and generalization ability of the model, being possible to adjust, by experimentation, according to the objective of the work.

At the end of all this process, we then have the addition of a final convolution layer with a sigmoidal activation (this choice falls on the fact that it is a classification function that returns a probability of each pixel belonging, or not, to the class of interest).

3.4 FPN Architecture

The second architecture developed throughout this project was an architecture called FPN - Feature Pyramid Network, chosen because it is an architecture widely used for semantic image segmentation, mainly thanks to its ability to handle objects at different scales. As with UNet, this type of architecture also has an encoding path, but then builds a FPN on the decoder path, as Fig. 5 shows.

This architecture is based on the same block system used in UNet. The encoding blocks all have the same type and the same number of layers and are called "ConvBlock". These blocks have the same type and the same number of layers:

- Two 2D convolution layers with a ReLU activation function.
- In between the convolution layers we have two BatchNormalization layers.

The encoder is the part of the model responsible for extracting high quality features at different scales from the input image. The main part of the encoding path is composed of 4 blocks (Block 1, 2, 3 and 4) with an increasing number

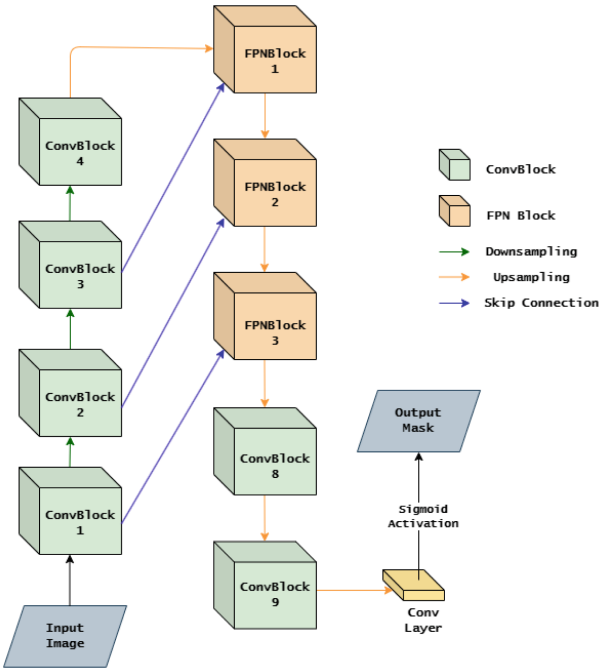

Fig. 5. Proposed FPN architecture.

of filters, where in each block a MaxPooling layer is added in order to halve the spatial dimension of the features.

After Block 4, the FPN is then built on the decoding path, responsible for refining the segmentation with the features obtained during the encoding process, while reconstructing the spatial resolution of the segmented image. The FPN has 3 blocks, with each of them being organized as follows and nominated "FPNBlock":

- An initial layer with a 2×2 stride transposed convolution, called the "up" layer.
- A "skip connection" is added between the initial layer of each block and a convolutional layer of the higher spatial resolution encoder, and a 1×1 filter and ReLU activation function is applied to this layer to combine the high-level information from the "up" layer with the information from the layer that has the higher spatial resolution.
- The outputs from the "up" layer and the "skip connection" layer are added element by element in order to combine the information from the two layers and thus allow the network to use both the high-level and lower-level features together.
- Finally, the result of this addition goes through the "ConvBlock" block, thus refining the combined features, capturing more details and improving the quality of the segmentation.

As mencioned before, this operation is performed three times (on Blocks 5,6 and 7), until two additional "ConvBlock" are added to the decoder in order to further refine the features and obtain finer details (Blocks 8 and 9). The output layer is then created with a 1×1 convolution in order to generate a segmentation map, where sigmoidal activation was used in order to produce a classification function that returns the probability of a pixel belonging, or not, to the class of interest, as happened in the UNet architecture.

3.5 Compiling and Training the Models

Before proceeding to the explanation of model compilation and training, it is first necessary to talk about the Dice Coefficient. This metric is usually used to evaluate the degree of overlap (or similarity) between two binary masks - in this specific case, between the segmentation masks generated by the model and the reference masks (generated by the true labels resulting from marking the coordinates of the points and, later, the development of a region around it in the pre-processing of the images). The Dice Coefficient measures the similarity between the predicted segmentarion region and its corresponding ground truth, with its formula described in (1), where we have set the epsilon value to be $1 * 10^{-7}$.

$$\text{Dice Coeff } = \frac{2 * \sum (y_{\text{true}} * y_{\text{pred}}) + \varepsilon}{\sum y_{\text{true}} + \sum y_{\text{pred}} + \varepsilon} \tag{1}$$

For training purposes, we will use the complementary function of the Dice Coefficient, called Dice Loss, as described in (2), where we calculate the difference between the perfect overlap, 1, and the evaluation given by the Dice Coefficient.

$$\text{Dice Loss } = 1 - DiceCoeff \tag{2}$$

Another metric we used in order to better evaluate the robustness and effectiveness of our model was the Root Mean Squared Error (RMSE). This metric is usually used to evaluate the difference between predicted values by a model (on our study case, the landmarks location) and the true values (those marked in LabelMe). It is calculated as the formula bellow shows (3), and it measures the average magnitude of the prediction errors. In terms of results - a value as close to zero indicates that there is almost no error between true labels and predicted labels, while a big value indicates that there is a discrepancy between the true and predicted labels. In the formula (3), the "n" value stands for the number of samples, while the ytrue and ypred are, respectively, the real labels and the predicted ones.

$$RMSE = \sqrt{\frac{\sum (y_{\text{pred}} - y_{\text{true}})^2}{n}} \tag{3}$$

The models were compiled taking into account the same parameters for the purpose of getting a better and trustful comparison between them:

- Adam optimizer with a learning rate of 0.001 was used.
- The loss function, as mentioned earlier, was set to be the complementary function of the Dice Coefficient.
- As metrics for the evaluation of the model both the Dice Coefficient and the RMSE were used.

The training of both models was then performed, via Google Collaboration, using their T4 GPU, 12GB RAM and 78GB disk configuration. As a definition of the training, we have the following:

- The vertically stacked matrices of the training images and training masks are used as training data.
- The batch size during training was set to be 32.
- The number of training epochs was set to be 200.
- An additional parameter, consisting of the vertically stacked matrices of validation images and validation masks, is added in order to check the performance of the model against this validation set during training.
- A callback function called "ReduceLROnPlateau" is defined, in order to dynamically adjust the learning rate during training. It was established that the variable in question would be the Dice Loss of the validation set, with a "patience" parameter of 10 and a "factor" parameter of 0.5 - after 10 consecutive epochs with no improvement in the validation Dice Loss, the learning rate is halved.

In terms of training time, Unet's architecture took 38 min to train the model, while FPN took only 28 min to complete the training.

To avoid future repetitions of the training, both models were saved at the end of the training, thus allowing them to be used in a much faster way in the future.

4 Results and Discussion

4.1 Training Metrics

$$IoU = \frac{\text{Intersection}}{\text{Union}} \tag{4}$$

After the training process of both models is done, we can then compare the results of the metrics defined as the evaluators of the training - the Dice Coefficient and the RMSE. Throughout the 200 training epochs, the Dice Coefficient and RMSE values were recorded, as documented in Table 1. We then made a measurement to find out how the model performed in the testing set using the Intersection over Union (IoU) metric, as shown in the formula (4). This is a metric that can also be known as Jaccard Index and represents the overlap between two sets - in this specific case it was calculated taking into account the overlap between the true mask and the predicted mask by each of the models, and the results are as recorded in the Table 2.

Table 1. Evaluation of Training Metrics

Model	Dice Coefficient		RMSE		Training Time
	Training	Validation	Training	Validation	(min.)
UNet	0.8426	0.7054	0.0233	0.0409	38
FPN	0.8900	0.7021	0.0198	0.0338	28

Table 2. Evaluation of Testing Set Metrics

Model	IoU
UNet	0.8342
FPN	0.9117

By analysing all the data collected, it is pretty much clear that these metrics indicate that the FPN model is more robust than the UNet model.

4.2 Predictions

The final tests were then done to see if the models could accurately detect landmarks on images that had not been trained - the images belonging to our testing set, with 40 images that were completely random and different from the training and validation sets.

To do this testing we ran the images one by one, trying to have the model of each architecture predict where the landmarks would be, thus creating a mask with those landmarks. Then we manipulated the dimensions of the test images, using the "squeeze" function to remove any unitary dimensions, doing the same with the mask generated by the models. Next, we made a figure with 4 subplots - the original test image, the actual mask created by marking the points in LabelMe, the original test image with the predicted mask applied, and finally the predicted mask, as shown in Figs. 6 and 7, both being representatives of the same test image for comparison reasons. As can be seen here, both models are extremely accurate in terms of mask definition.

4.3 Application in YEN Angle Calculation

As mentioned before, our study was not only about the automatic identification of landmarks on x-rays using Deep Learning but also about trying to use precise metrics that would allow for a quick and effective diagnostic situation. In our project, we chose to mark 6 points in each image, but will only use 3 for the calculations needed for the next step - the S point (Sella), the M point (Midpoint of the Anterior Maxilla) and the G point (Center of a circle inside the Gnatio point). These three points, when connected by vectors, form an angle between them called the YEN angle.

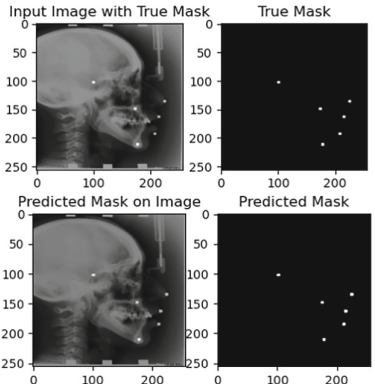

Fig. 6. Prediction made by the UNet Model.

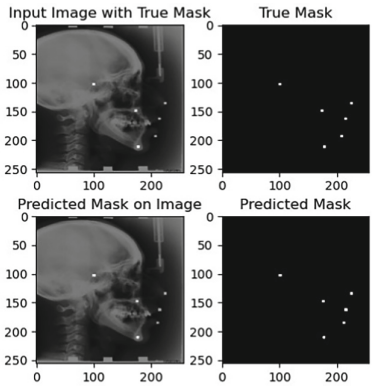

Fig. 7. Prediction made by the FPN Model.

This YEN angle is a direct metric in the evaluation of a patient's Mandibular Occlusion. It is a metric characterized by having 3 different classifications, which will now be explained:

- A YEN angle with amplitude between 117 and 123° demonstrates a Class I type of occlusion, which is the so-called normal/ideal occlusion.
- A YEN angle with amplitude below 117° is representative of a Class II type of Occlusion, which indicates a protruding maxilla.
- A YEN angle with amplitude above 123° points to a Class III type of Occlusion, which indicates that the patient has a prominent mandible.

Since our code created an ROI around the pointed landmark and the predicted mask returns small ROIs around each landmark, we calculate the centroid of each of these ROIs in order to do the YEN angle calculation. To do this step, we first converted the predicted binary mask into a label image. Next, we extracted the properties of the regions present in the label image using the

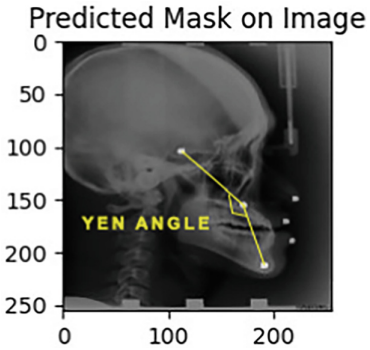

Fig. 8. YEN Angle on Predicted Mask.

"skimage" module, thus removing the information of interest for our problem - the centroid coordinates. To conclude this step, and in preparation for the next step, these centroids are converted into numpy arrays.

To calculate the angle YEN, we defined a function responsible for calculating the vectors of the segments "S-M" and "G-M" by subtracting the coordinates. Then we normalized each of these vectors by dividing them by their length, and finally used the mathematical expression in (5) in order to calculate the angle.

$$YEN = \cos^{-1}\left(\frac{v_{SM}.v_{GM}}{\|v_{SM}\| * \|v_{GM}\|}\right), \tag{5}$$

with $\mathrm{v}_{SM} = (x_{SM}, y_{SM})$ and $\mathrm{v}_{GM} = (x_{GM}, y_{GM})$.

By placing the 3 points (defined as S, M and G) we were then able to immediately get the value of the YEN angle of the person in question. In the test image of the FPN model, 8, the feedback was that the patient had a YEN angle of 150°. As mentioned earlier, it was then clear that the patient had prominent mandible.

4.4 Results Discussion

By analysing all the data we collected in our study, it is clear now that our FPN architecture produces a better and more robust model, capable of accurately predicting landmarks in medical images. The capability of this architecture to handle objects at different scales, alongside with feature sharing between different resolutions, makes the FPN model a success. However, slightly less robust but equally efficient, our UNet model proved itself to be capable of acurately predicting landmarks in medical images as well.

However, an issue arose throughout the process in which we developed our project - the issue of marking the landmarks. The fact that this marking is done by non-health professionals can lead to erroneous markings that will affect the training of the model and, consequently, the diagnosis.

We are left with the firm belief that with a professionally prepared dataset and prepared in such a way that the clinical case under study is well represented proportionally (i.e. without class imbalance) the model can take on even greater robustness.

5 Conclusion

The purpose of this project was to develop an architecture, based on Deep Learning, that would allow the automatic identification of landmarks in X-ray images.

Two architectures were developed, a UNet and a FPN, where it was verified that the FPN presented better metrics and, subsequently, more reliable results for the medical application of the work developed.

It was noted that the FPN model is faster in terms of training, without its speed disallowing the return of accurate landmarks, where it shone as extremely accurate and efficient. The UNet model, on the other hand, trained slowly, but still managed to detect the landmarks with high accuracy. These informations lead us to conclude that the FPN model, being faster and more accurate, is the better of the two models, showing a very high robustness.

This study also had as an additional goal the direct application in a medical context of diagnosing the models and their predictions. It was possible to develop a metric analysis, based on the encephalograms in the dataset, on the type of Mandibular Occlusion of the patients, thus allowing a fast and accurate response in the identification and classification of a diagnosis.

It is clear that these types of models are of particular clinical importance. Fast, accurate, and objective analysis of metrics can greatly facilitate the diagnostic process, enabling faster access to care/treatment, more personalized care for each patient, and allowing health professionals to have more time for even greater care for their patients.

The accuracy that has been obtained, in a diagnostic context, shows that a model with this architecture can easily calculate any type of metric that is based on landmarks - from hip dysplasia to scoliosis to osteoarthritis, various diseases can have their treatment and diagnosis made easier using tools of this type.

We hope that with the development of this project and study, we have a solid foundation to develop a universal model in the future, capable of evaluating any type of metric using image segmentation, something as simple as taking an x-ray and inserting the image for the model to predict the necessary landmarks, thus facilitating the entire process of diagnosis and subsequent treatment/follow-up.

References

1. Schwendicke, F., et al.: Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clin. Oral Invest. **25**, 4299–4309 (2021)
2. Reddy, P., Kanakatte, A., Gubbi, J., Poduval, M., Ghose, A., Purushothaman, B.: Anatomical landmark detection using deep appearance-context network. In: 2021 43rd Annual International Conference Of The IEEE Engineering In Medicine & Biology Society (EMBC), pp. 3569–3572 (2021)

3. Yeh, Y., Weng, C., Huang, Y., Fu, C., Tsai, T., Yeh, C.: Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs. Sci. Rep. **11**, 7618 (2021)
4. Çallı, E., Sogancioglu, E., Ginneken, B., Leeuwen, K., Murphy, K.: Deep learning for chest X-ray analysis: a survey. Med. Image Anal. **72**, 102125 (2021)
5. Jiang, F., et al.: Automated calibration system for length measurement of lateral cephalometry based on deep learning. Phys. Med. Biol. **67**, 225016 (2022)
6. An, C., Lee, J., Jang, J., Choi, H.: Part affinity fields and CoordConv for detecting landmarks of lumbar vertebrae and sacrum in X-ray images. Sensors. **22**, 8628 (2022)
7. Pei, Y., et al.: Learning-based landmark detection in pelvis x-rays with attention mechanism: data from the osteoarthritis initiative. Biomed. Phys. Eng. Express **9**, 025001 (2023)
8. Zhou, G., et al.: Learn fine-grained adaptive loss for multiple anatomical landmark detection in medical images. IEEE J. Biomed. Health Inform. **25**, 3854–3864 (2021)