Transfer Learning for an Automated Detection System of Fractures in Patients with Maxillofacial Trauma

applied sciences

MDPI

Article

Transfer Learning for an Automated Detection System of Fractures in Patients with Maxillofacial Trauma

Maria Amodeo [1], Vincenzo Abbate [2], Pasquale Arpaia [3,*], Renato Cuocolo [3,4], Giovanni Dell'Aversana Orabona [2], Monica Murero [5,6], Marco Parvis [1], Roberto Prevete [7] and Lorenzo Ugga [8]

1 Department of Electronics and Telecommunications (DET), Polytechnic University of Turin, 10129 Turin, Italy; maria.amodeo@polito.it (M.A.); marco.parvis@polito.it (M.P.)
2 Department of Neurosciences, Reproductive and Odontostomatological Science, University of Naples Federico II, 80131 Naples, Italy; vincenzo.abbate@unina.it (V.A.); dellaversana@unina.it (G.D.O.)
3 Interdepartmental Research Center on Management and Innovation in Healthcare—CIRMIS, University of Naples Federico II, Via Pansini 5, 80138 Naples, Italy; renato.cuocolo@unina.it
4 Department of Clinical Medicine and Surgery, University of Naples Federico II, 80131 Naples, Italy
5 Department of Social Sciences, University of Naples Federico II, 80131 Naples, Italy; monica.murero@unina.it
6 Distributed Artificial Intelligence Lab, Technische Universität, 10587 Berlin, Germany
7 Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, 80100 Naples, Italy; rprevete@unina.it
8 Department of Advanced Biomedical Sciences, University of Naples Federico II, 80131 Naples, Italy; lorenzo.ugga@unina.it
* Correspondence: pasquale.arpaia@unina.it

Citation: Amodeo, M.; Abbate, V.; Arpaia, P.; Cuocolo, R.; Dell'Aversana Orabona, G.; Murero, M.; Parvis, M.; Prevete, R.; Ugga, L. Transfer Learning for an Automated Detection System of Fractures in Patients with Maxillofacial Trauma. *Appl. Sci.* **2021**, *11*, 6293. https://doi.org/10.3390/app11146293

Academic Editor: Qi-Huang Zheng

Received: 26 May 2021
Accepted: 2 July 2021
Published: 7 July 2021

Abstract: An original maxillofacial fracture detection system (MFDS), based on convolutional neural networks and transfer learning, is proposed to detect traumatic fractures in patients. A convolutional neural network pre-trained on non-medical images was re-trained and fine-tuned using computed tomography (CT) scans to produce a model for the classification of future CTs as either "fracture" or "noFracture". The model was trained on a total of 148 CTs (120 patients labeled with "fracture" and 28 patients labeled with "noFracture"). The validation dataset, used for statistical analysis, was characterized by 30 patients (5 with "noFracture" and 25 with "fracture"). An additional 30 CT scans, comprising 25 "fracture" and 5 "noFracture" images, were used as the test dataset for final testing. Tests were carried out both by considering the single slices and by grouping the slices for patients. A patient was categorized as fractured if two consecutive slices were classified with a fracture probability higher than 0.99. The patients' results show that the model accuracy in classifying the maxillofacial fractures is 80%. Even if the MFDS model cannot replace the radiologist's work, it can provide valuable assistive support, reducing the risk of human error, preventing patient harm by minimizing diagnostic delays, and reducing the incongruous burden of hospitalization.

Keywords: convolutional neural network; transfer learning; maxillofacial fractures; computed tomography images; radiography

1. Introduction

In recent years, the number of requests for computed tomography (CT), magnetic resonance imaging (MRI), and, in general, radiology services has grown dramatically [1]. Nevertheless, there is a lack of radiologists due to recruitment challenges and many retirements. In this scenario, artificial intelligence (AI) can help radiologists in the time-consuming and challenging medical image analysis task. In any case, the AI-based tools do not replace medical staff, but assistive technologies prioritize, confirm, or validate radiologists' decisions and doubts.

Deep learning, a branch of AI, has recently made substantial progress in analyzing images with a consequent better representation and interpretation of complex data. In

particular, various works [2–6] deal with deep learning in orthopedic traumatology. However, the number of studies regarding deep learning on CT scans for fracture detection is low. Furthermore, building and training a neural architecture from scratch requires a huge amount of data. Image classification networks are trained on billions of data in the literature, using multiple servers running for several weeks [7]. This procedure is not feasible for most medical researchers. One way to overcome this obstacle is to use the so-called transfer learning. This process consists of adopting the highly refined characteristics of convolutional neural networks trained on millions of data and using them as a starting point for a new model. For example, to verify the extent of fracture detection on wrist radiographs, Kim and MacKinnon [8] focus on transfer learning from a deep convolutional neural network (CNNs), pre-trained on non-medical images. Using the inception V3 CNN [9], they obtained an area under the receiver operating characteristic curve (AUC-ROC) of 0.95 on the test dataset. This result shows that a CNN pre-trained on non-medical images can be used for medical radiographs successfully. Another study was carried out by Chung et al. [10], based on a CNN to detect and classify proximal humerus fractures using plain anteroposterior shoulder radiographs. The deep neural network showed a similar performance to that of shoulder-specialized orthopedic surgeons, but better than that of the general physicians and the non-shoulder specialized orthopedic surgeons. This result denotes the possibility to diagnose fractures accurately by using plain radiographs automatically. Another study in this field was carried out by Tomita et al. [11], where they focused on detecting osteoporotic vertebral fractures on CT exams. Their system consisted of two blocks: (i) a CNN to extract radiological features from CTs; and (ii) a recurrent neural network (RNN) module to aggregate the previously extracted elements for the final diagnosis. The performance of the proposed system matched the ability of radiologist practitioners. Thus, the system could be used for screening and prioritizing potential fracture cases.

Therefore, although several authors have already described some AI applications in the orthopedic field, the possibility to detect maxillofacial fractures in 3D images (CT scans) of injured patients using artificial neural networks, and in particular a transfer learning approach, has not been explored yet [12–15]. This area's anatomical complexity and the specificity of this type of fracture make radiological diagnosis very often complex with a consistent risk of incongruous hospitalizations. A fracture detection system based on AI able to detect the presence of maxillofacial fractures would be of great use in clinical practice by reducing the costs of treatment and discomfort for patients.

This research aims to develop a fracture detection system, based on the transfer learning approach, able to predict the presence of maxillofacial fractures. The inputs for this system are the CT images of a patient after a trauma. The output of the system indicates the existence or not of a fracture. The block diagram of the system is shown in Figure 1.

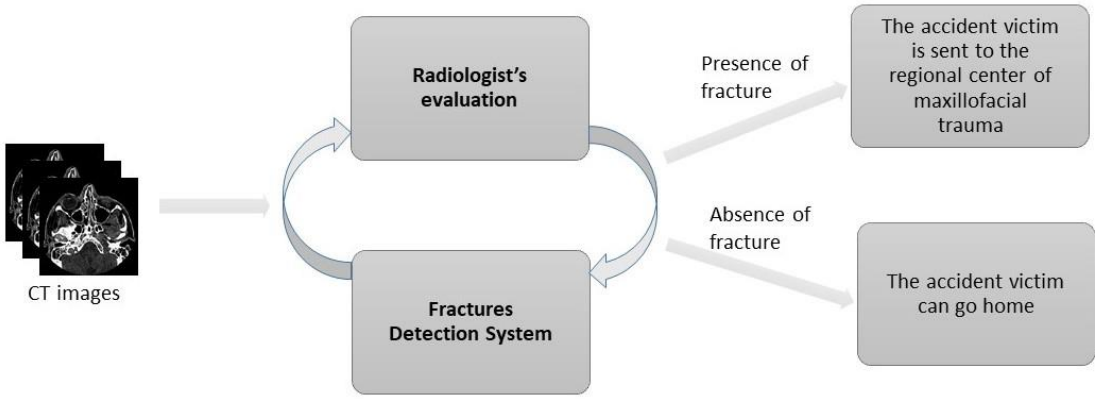

Figure 1. Block diagram of the system for patients with maxillofacial trauma. The fracture detection system assists the radiologist in evaluating the CT images of an injured patient.

The paper is organized as follows. In Section 2, the material and methods are presented, including the description of the dataset and the architecture used. In Section 3, the results are presented in terms of slices and patients. In Section 4, we discuss the results achieved, while in Section 5 the conclusions of the study are presented.

2. Materials and Methods

2.1. Dataset

This retrospective study uses images from CT exams after anonymizing patient personal data. The study was approved by the Ethics Committee of "Federico II" University, Naples, Italy (approval number 81/20). The CT scans were obtained from the internal database of the U.O.C. of Maxillofacial Surgery of the University Hospital "Federico II", which collects examinations conducted from 2000 to 2020. We performed CT investigations of the facial mass on different devices (TC 16–64 slice) with thickness volumetric acquisition (0.5–2 mm) and variable in-plane resolution (0.5 × 0.5–1 × 1 mm). For the analysis, we selected only the images we obtained with the bone reconstruction algorithm. Two radiologists (R.C., L.U.) consensually examined, interpreted, and classified each CT image according to fracture rhymes' presence/absence. We also included control CT investigations from patients with the non-traumatic facial mass disorder.

The number of CT scans corresponds to the number of patients (a CT scan for each patient). The total dataset consisted of 208 patients: 170 patients (11,260 slices of CT scans) labeled as with "fracture" and 38 patients (49,762 slices of CT scans) labeled as "noFracture". The total dataset was divided into training, validation, and test datasets. In particular, the training dataset consisted of 148 CT images (120 patients labeled as with "fracture" and 28 patients labeled as with "noFracture"). The validation dataset, used for statistical analysis, was characterized by 30 patients (5 with "noFracture" and 25 with "fracture"), and an additional 30 CT scans, comprising 25 "fracture" and 5 "noFracture" images, were used as a test dataset for final testing. It is worth noting that the total dataset was imbalanced on a patient level with the majority being fractured patients; while on a slice level, the dataset is imbalanced in favor of the slices labeled as "noFracture". Therefore, the dataset overall is not as imbalanced in favor of "fracture" images as can be assumed by only evaluating the patient-level data.

2.2. Experimental Setup Description

The system implementation was carried out through a predictive algorithm written in Python v.3.7.6 (available for different Operating Systems) [16], using PyTorch v.1.4.0 [17] and Fastai v.1.0.60 [18]. We used scikit-learn v.0.22.1 [19] for the neural architecture and Pydicom v.1.4.2 [20] to treat CT images in Dicom format. The implementation of the system is schematized in Figure 2.

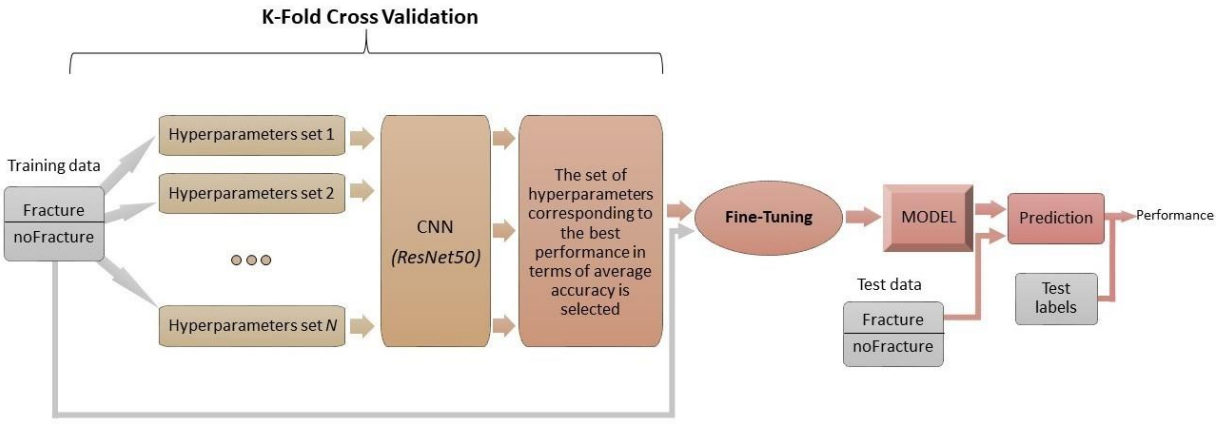

Figure 2. Block diagram of the system's implementation for detecting fractures in patients with maxillofacial trauma.

All the steps can be summarized as follows:

1. K-fold cross validation to identify the hyperparameters (learning rate, weight decay, and drop out) that allow the network to have the highest performance in terms of accuracy;
2. Fine-tuning of the network with the hyperparameters chosen in the previous step:
   2.1 Training only of the last layer;
   2.2 Unfreezing and training the whole model;
3. Evaluation of the network's performance.

All the steps are described in detail in the next paragraphs.

2.2.1. K-Fold Cross Validation

For the implementation of the architecture shown in Figure 2, the first step consists of defining the training dataset for the k-fold cross validation, comprising two classes: "fracture" and "noFracture". In particular, to keep the two classes balanced and reduce the computational times, we considered a reduced dataset, which is a subset of the total dataset described in Section 2.1. In particular, the training dataset used for the k-fold cross validation consists of 359 slices with fracture, belonging to 57 different patients, and 362 slices without fracture, belonging to 59 additional patients. In order to avoid class imbalance in patient-level, from some patients with fractures, we selected only a subset of the "noFracture" slices. Therefore, these patients will become patients with "noFracture" in this phase.

In our case study, we adopted the transfer learning technique to reduce the development burden of the CNN. The pre-trained architecture we used was ResNet50. ResNet is the deep convolutional neural network that won the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [21]. ResNet architecture has many variants: the difference between them is not only a different number of layers, but also a novel architecture, such as ResNeXt [22], or densely connected CNN [23]. ResNet50 is trained on more than a million images from the ImageNet database [24]. The network is 50 layers deep and can classify images into 1000 object categories, such as pizza, umbrella, castle, and many animals (tiger, camel, frog, etc.). As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.

The architecture of ResNet50 has 4 stages:

1. Initial convolution (kernel size of 7×7) and max-pooling (kernel size of 3×3);
2. Nine convolutional layers: kernel size of 1×1 and 64 different kernels, followed by kernel size of 3×3 and 64 different kernels, followed by kernel size of 1×1 and 256 different kernels. These three layers are repeated 3 times;
3. Twelve convolutional layers: kernel size of 1×1 and 128 different kernels, followed by kernel size of 3×3 and 128 different kernels, followed by kernel size of 1×1 and 512 different kernels. These three layers are repeated 4 times;
4. Eighteen convolutional layers: kernel size of 1×1 and 256 different kernels, followed by kernel size of 3×3 and 256 different kernels, followed by kernel size of 1×1 and 1024 different kernels. These three layers are repeated 6 times;
5. Nine convolutional layers: kernel size of 1×1 and 512 different kernels, followed by kernel size of 3×3 and 512 different kernels, followed by kernel size of 1×1 and 2048 different kernels. These three layers are repeated 3 times;
6. Average pooling layer followed by a fully connected layer with 1000 neurons and a softmax function at the end.

In order to choose the most suitable set of hyperparameters for our case, we used the stratified k-fold cross validation [25] with k = 5. The hyperparameters of interest were the following: learning rate, weight decay, and dropout; we chose them in the following ranges (0.000001; 0.005), (0.0001; 0.0005), (0.1; 0.5). We set the batch size at 50. Specifically, 20 combinations (N = 20 in Figure 2) of the hyperparameters were tested. We used a random search for hyperparameters' optimization. We also chose to adopt a random search

compared to a grid search. When there are many hyperparameters, as in our case, the first is more effective from the computational time point of view, while maintaining good performance [26]. Figure 3 describes the procedure of the k-fold cross-validation.

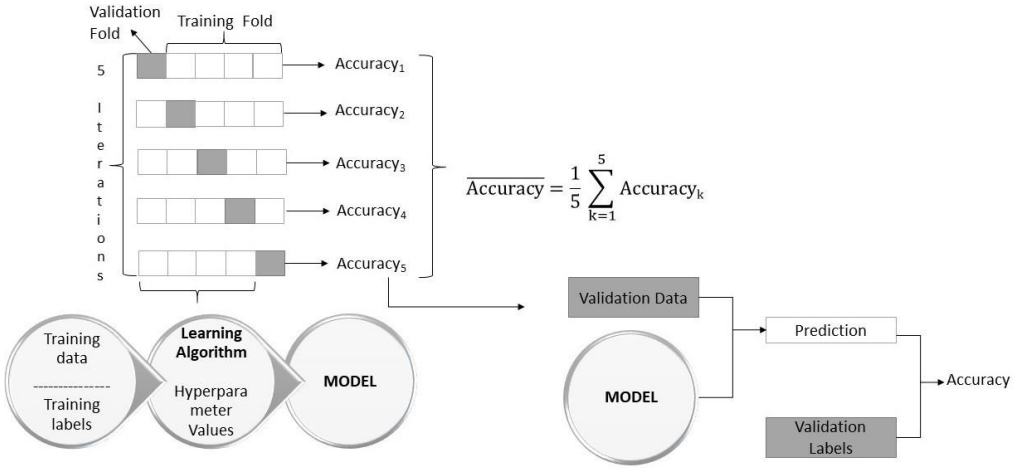

Figure 3. Five-fold cross validation procedure scheme.

Early stopping criteria can be used during the training as a trade-off between generalization ability and computational costs. In our case, we used as early stopping the following criteria: if after three attempts the accuracy does not improve by at least 0.01, the training cycle ends. The number of epochs set for each fold was 6. The images were normalized according to the ImageNet format and resized from 512 × 512 to 224 × 224 pixels. An example of Dicom images is shown in Figure 4.

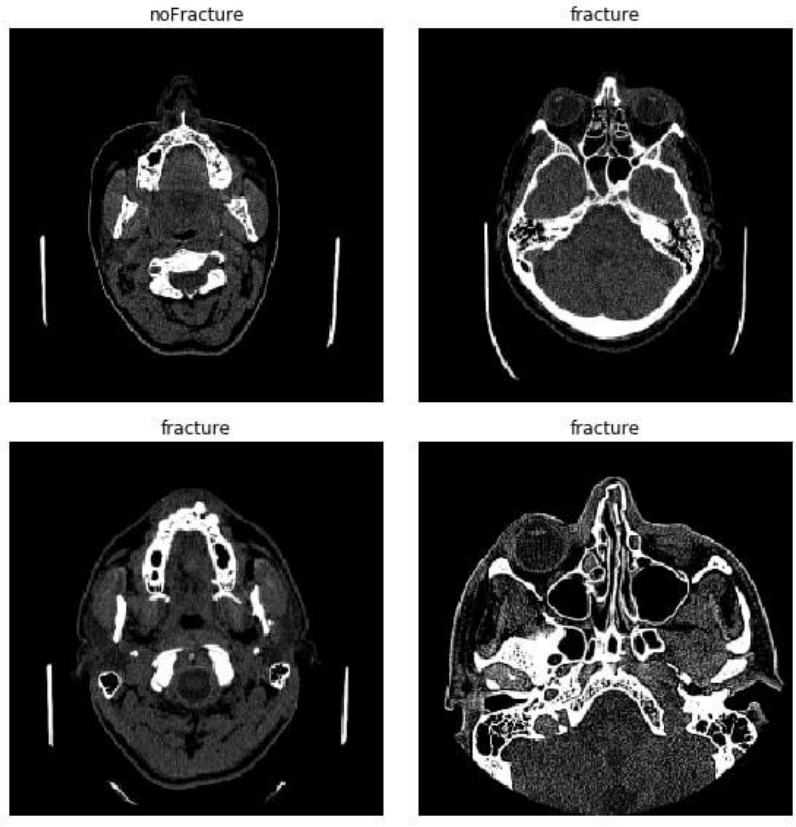

Figure 4. Example of Dicom images (8 × 8 inches) for both classes (fracture and noFracture).

After carrying out the tests for the 20 configurations, we chose the set of hyperparameters that guaranteed the network to have the highest average accuracy (0.86) and the smallest standard deviation that is the index of little variability (0.05). This set has the following hyperparameters: learning rate of 0.005, weight decay of 0.0005, and drop out of 0.5.

2.2.2. Fine-Tuning of the CNN

Pre-trained networks can be exploited to recognize classes the system is not (initially) trained on, thanks to the fine-tuning process.

The convolutional layers had already learned discriminative filters. After choosing the hyperparameters, described in the previous section (Section 2.2.1), we replaced the final set of fully connected layers of the pre-trained CNN. We introduced a new set of fully-connected layers using random weights. By doing so, the fully connected layers could act entirely randomly. If the gradient backpropagates from these random values and the whole network, the pre-trained network's powerful features risked being destroyed. To avoid this problem, we re-trained the CNN performing the following steps (Figure 5):

1. Training of the last layer: we started with the pre-trained model's weights (pre-trained on ImageNet), freezing all layers in the network's body except the last layer. In this step, we trained only the last layer.
2. Unfreezing and training the whole model: in this step, after the last layer had started to learn patterns of our medical dataset, we unfroze all the weights and trained the entire model with a very small learning rate. We wanted to avoid altering the convolutional filters dramatically.

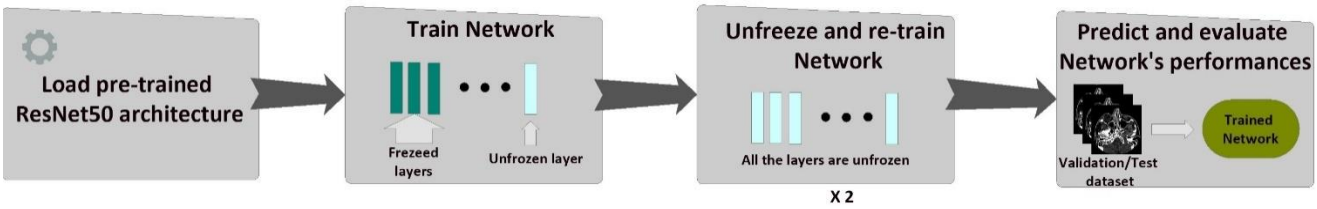

Figure 5. ResNet50 was used as a pre-trained network and, after loading the network, the fine-tuning process was started. We froze all the layers in the network except the fully-connected layers, useful for capturing high-level features on the current dataset. After the fully-connected layers have had a chance to learn patterns from our dataset, we then unfroze all the architecture layers; even the convolutional layers that had initially learned discriminative filters. We allowed each layer to be fine-tuned by performing two training steps and using differential learning rates.

For the fine-tuning of the network, we used the total dataset described in Section 2.1. In particular, the training dataset consisted of 8023 slices labeled as "fracture" and 34,962 labeled as "noFracture", for a total of 148 patients. The training and validation datasets used in the k-fold cross validation were a subset of this total training dataset. Since the two classes were no longer balanced, we used the CrossEntropyLoss as loss function with different weights for the "fracture" and "noFracture" classes (w_f and w_{nf}, respectively):

$$[w_f, w_{nf}] = [\frac{|\text{noFracture}|}{|\text{fracture}|}, \frac{|\text{noFracture}|}{|\text{fracture}|}] = [\frac{34,962}{8023}, \frac{34,962}{34,962}] = [4.36, 1.0] \quad (1)$$

The validation dataset, used for the error evaluation, consisted of 1660 slices labeled as "fracture" and 7910 labeled as "noFracture", for a total of 30 patients.

During the second step, we performed an additional fine-tuning, re-training the model twice by changing the learning rate to improve the model's performance. Before each re-train of the model, we loaded the network's weights that gave us the best performance in terms of accuracy. In particular, we used the learning rate finder [27,28] of the Fastai library to choose the learning rate at each step. Since some features remain unchanged (such as the edges and the corners of an image learned in the first layers of the network),

we applied the concept of differential learning rates implemented by the Fastai library. Using this approach, we could assign different learning rates to the various layers of our network. In particular, we passed a slice function inside the fit method and: (a) assigned a lower learning rate to the first layer, (b) assigned a higher learning rate to the last layer, and (c) distributed the values for the learning rate among all the other layers in between.

3. Results

The results presented in this section are obtained on the validation dataset and on the test dataset that consists of 1577 slices labeled as "fracture" and 6890 slices labeled as "noFracture", for a total of 30 patients. The partition of the dataset into training, validation, and test dataset was done randomly at level-patient, this means that all the slices for a single patient were considered in one of the three sets (training, validation, and test). Nevertheless, the validation and test datasets were not similar to each other. First, the CT scans were performed on different devices and, therefore, we have substantial differences among them; then, the fracture can affect any part of the splanchnocranium and, since the latter is a very large and complex region, the CT images can be very different from each other. The confusion matrix of the validation and test datasets is shown in Figure 6a,b, respectively; the AUC-ROC for both validation and test datasets is shown in Figure 6c,d, respectively.

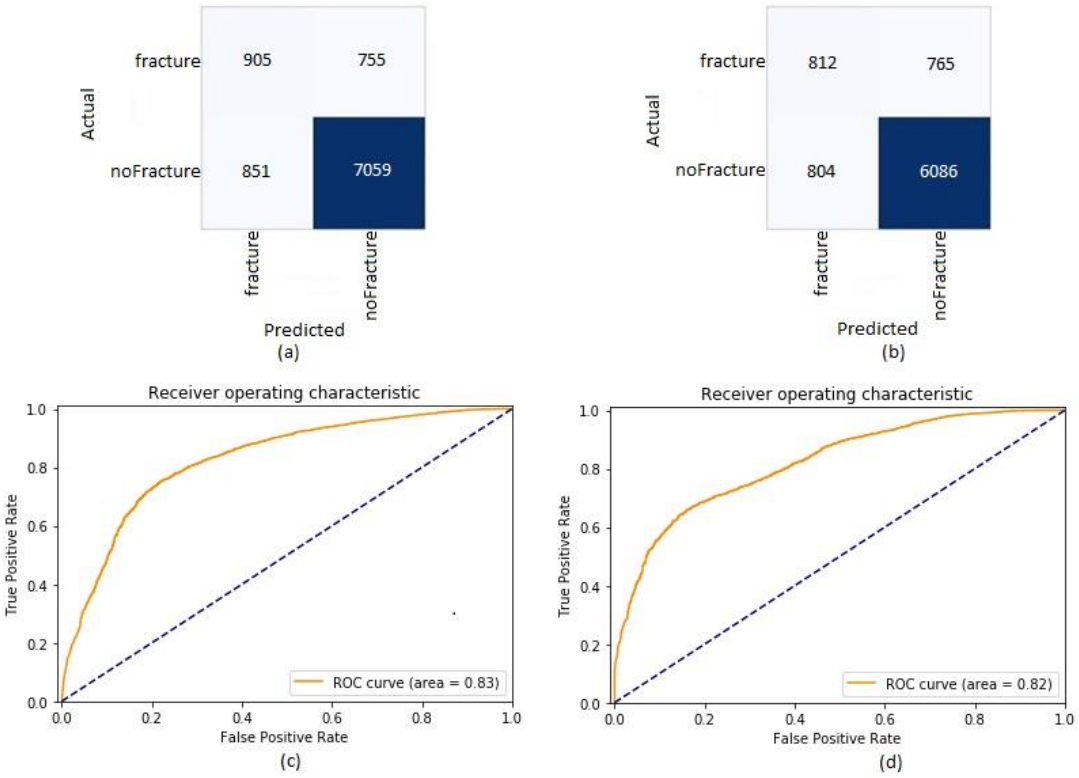

Figure 6. Results in terms of the confusion matrix for the validation (**a**) and test (**b**) datasets and ROC curve for the validation (**c**) and test (**d**) datasets. The corresponding AUC for the validation dataset is 0.83 (0.82, 0.84), while for the test dataset is 0.82 (0.81, 0.83). The 95% confidence intervals for the values of the AUC were calculated with the analytic method of Hanley and McNeil [29], such as described in Ref. [30].

For the evaluation of the performance, we considered the following metrics:

- Accuracy $= \frac{TP+TN}{P+N}$
- Recall (or sensitivity) $= \frac{TP}{TP+FN}$
- Precision (or positive predictive value) $= \frac{TP}{TP+FP}$

The corresponding values for the validation and test datasets are shown in Table 1.

Table 1. Accuracy, recall (sensitivity), and precision (positive predictive value) with the exact (Clopper–Pearson) 95% confidence intervals for the validation and test dataset.

Metric	Validation Dataset	Test Dataset
Accuracy	0.83 (0.82, 0.84)	0.81 (0.81, 0.82)
Recall	0.55 (0.52, 0.57)	0.51 (0.49, 0.54)
Precision	0.52 (0.49, 0.54)	0.50 (0.48, 0.53)

The actual width of the confidence interval is the same for both recall and precision in both validation and test datasets, while it is much smaller for the accuracy in both datasets.

In order to make a prediction in terms of a patient's injury rather than single slices, we performed an evaluation of the neural network. To this aim, the slices were grouped by referring to a single patient according to the following assumption: if two consecutive slices, belonging to the same patient, are classified as "fracture" by the CNN with a probability greater than 0.99, then classify the patient as a patient with a fracture. The confusion matrix we obtained for the test dataset is shown in Figure 7.

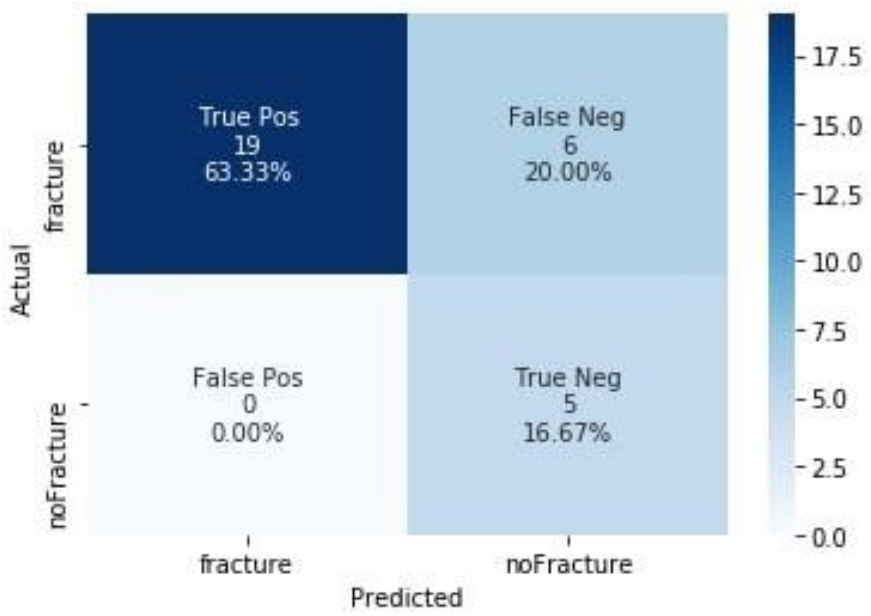

Figure 7. Confusion matrix for the test dataset in terms of patients' fractures.

The measures of diagnostic accuracy (accuracy, recall (sensitivity), and precision (positive predictive value)) with 95% confidence intervals for the test dataset in terms of patients are shown in Table 2.

Table 2. Accuracy, recall (sensitivity), and precision (positive predictive value) with the exact (Clopper–Pearson) 95% confidence intervals for the test dataset in terms of patients.

Metric	Test Dataset
Accuracy	0.80 (0.61, 0.92)
Recall	0.76 (0.55, 0.91)
Precision	1.0 (0.82, 1.00)

4. Discussion

4.1. Statement of Principal Findings

The proposed approach shows the feasibility of using transfer learning techniques to detect maxillofacial fractures in CT images effectively. The results achieved by using the

validation and test datasets are of the same order of magnitude. Our trained ResNet50 neural network can distinguish between the fractured and normal bone in CT scans of injured patients with a relatively high accuracy (80%). This result is particularly promising, given the anatomical complexity and thinness of bones in the splanchnocranium, and proves that transfer learning from CNN, pre-trained on non-medical images, can be efficiently applied to the problem of maxillofacial fracture detection on CT images.

4.2. Strengths and Weaknesses of the Study

Although a computer-aided decision system with an AUC of 0.83 (0.82, 0.84) cannot replace human interpretation, this accuracy level may be very useful in assisting radiologists with prompt a diagnosis and treatment. An automated detection system based on our proposed model has the advantages of analyzing the CT image's entire region with equal importance. This reflects in reducing the human errors related to missed readings on the whole region of the 3D image. Furthermore, small fractures are often hardly visible on CT images, and require multiple checks by the radiologists: an automated detection system can also be useful in this context.

4.3. Strengths and Weaknesses in Relation to Other Studies, Discussing Particularly Any Differences in Results

Although several authors have already investigated AI applications in the orthopedic field, the possibility to detect maxillofacial fractures in 3D images of injured patients using deep learning algorithms has not been explored yet. Even if in other studies we can find better results, for example, in terms of AUC-ROC (0.95 [8]), it must be taken into account the complexity of the region of interest, such as the splanchnocranium and the enormous variability of the fracture types that may be present in this anatomically complex district. It is important to remark that the algorithm should be intended as an aid to the radiologist in recognizing facial fractures, more as a second opinion, rather than an independent one.

4.4. Meaning of the Study: Possible Mechanisms and Implications for Clinicians or Policymakers

The assessment of CT images in trauma patients is fundamental to select the appropriate treatment and direct them towards highly specialized units if necessary. When a patient's trauma occurs in an anatomically complex district such as the splanchnocranium, two main difficulties arise from the current clinical practice. The first one is the possible failure to recognize the presence of a bone fracture, and the second is the incorrect classification of normal anatomical structures (i.e., sutures, vascular, and nerve channels) as traumatic injuries. These diagnostic difficulties frequently translate into increased costs for the health system and a burden for the patient due to unnecessary hospitalizations in specialized clinical wards. For example, once the need for urgent treatment is excluded in a craniofacial district trauma, patients are transferred from the emergency room (of first access) to the closest regional reference center specialized in maxillofacial trauma. Here, the clinical case reassessment in specialist settings frequently (about 20% of cases) highlights the incongruity of hospitalization and often the absence of indications for surgical treatment. These patients require only home medical therapy. Although there are several AI applications in the literature of the orthopedic field, they remain still unexplored in the maxillofacial district. An AI-based radiological diagnosis system would allow diagnostic errors to be minimized by providing the radiologist with a support tool to guide therapeutic choices. However, an innovative AI-based radiological system should not replace the radiologist's work but become a valuable assistive technology to reduce medical error risks, unnecessary transportation, hospitalization, and socio-economic burden for society and the public health governance [31].

4.5. Unanswered Questions and Future Research

Future studies can focus on automated fracture detection with tiny fractures, improving the algorithm to detect, for example, the corners of fractured bones to improve the detection sensitivity of the system. Furthermore, to enhance the network's performance,

a stage of preprocessing of the CT images could be introduced to remove the region of no interest for the prediction. Another interesting approach could be the investigation of the combination of deep learning models with radiomics [32]. In fact, radiomics [33] is a method for extracting a large amount of advanced quantitative imaging features from radiographic medical images obtained with computed tomography, using data-characterization algorithms. Radiomic data could be integrated into predictive models to hedge against the risk of overfitting the deep learning approach. Another possibility is to use a local feature detector as the speeded-up robust features (SURF) to improve the system's performance. In their work [34], the authors propose a computer-assisted method for automated classification and detection of calcaneus fracture locations in CT images using a deep learning algorithm. In particular, they compared two types of CNNs, a Residual network (ResNet) and a visual geometry group (VGG). Furthermore, the speeded-up robust features (SURF) method was used to determine the exact location and the type of fracture in calcaneal CT scans.

5. Conclusions

This study represents a proof of concept for using transfer learning from CNN, pretrained on non-medical images, for maxillofacial fracture detection on CT images. In the literature, the use of transfer learning applied to CT scans to detect maxillofacial fractures of injured patients has not yet been explored. Our system proved to be capable of predicting maxillofacial fractures in patients with an accuracy of 80%. MFDC can become a valuable technology in assisting radiologists with prompt diagnosis and treatment that could reduce medical error risks and prevent patient harm and stress by minimizing maxillofacial trauma's diagnostic delays. An AI-based system assisting radiological investigation in non-specialized clinical wards can reduce incongruous hospitalization's socio-economic burden for the patient, society, and health system.

Author Contributions: Conceptualization, V.A., P.A., R.C., G.D.O., R.P. and L.U.; methodology, M.A. and R.P.; software, M.A.; validation, M.A.; investigation, M.A.; data collection, V.A.; data curation, R.C. and L.U.; writing—original draft preparation, M.A.; writing—review and editing, V.A., P.A., R.C., G.D.O., M.M., M.P. and R.P.; visualization, L.U.; supervision, P.A.; project administration, P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of "Federico II" University, Naples, Italy.

Informed Consent Statement: Informed consent was waived because of the retrospective nature of the study and the analysis used anonymous clinical data.

Data Availability Statement: The data presented in this study are not available due to privacy restrictions.

Acknowledgments: This work was carried out as part of the "ICT for Health" project, which was financially supported by the Italian Ministry of Education, University and Research (MIUR), under the initiative "Departments of Excellence" (Italian Budget Law no. 232/2016), through an excellence grant awarded to the Department of Information Technology and Electrical Engineering of the University of Naples Federico II, Naples, Italy).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Kalmet, P.H.S.; Sanduleanu, S.; Primakov, S.; Wu, G.; Jochems, A.; Refaee, T.; Ibrahim, A.; Hulst, L.V.; Lambin, P.; Poeze, M. Deep learning in fracture detection: A narrative review. *Acta Orthop.* **2020**, *91*, 215–220. [CrossRef]
2. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
3. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
4. Lee, J.-G.; Jun, S.; Cho, Y.-W.; Lee, H.; Kim, G.B.; Seo, J.B.; Kim, N. Deep Learning in Medical Imaging: General Overview. *Korean J. Radiol.* **2017**, *18*, 570–584. [CrossRef] [PubMed]
5. Olczak, J.; Fahlberg, N.; Maki, A.; Razavian, A.S.; Jilert, A.; Stark, A.; Sköldenberg, O.; Gordon, M. Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? *Acta Orthop.* **2017**, *88*, 581–586. [CrossRef] [PubMed]
6. Tang, A.; Tam, R.; Cadrin-Chênevert, A.; Guest, W.; Chong, J.; Barfett, J.; Chepelev, L.; Cairns, R.; Mitchell, J.R.; Cicero, M.D.; et al. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J.* **2018**, *69*, 120–135. [CrossRef] [PubMed]
7. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. Kim, H.D.; MacKinnon, T. Artificial intelligence in fracture detection: Transfer learning from deep convolutional neu-ral networks. *Clin. Radiol.* **2018**, *73*, 439–445. [CrossRef] [PubMed]
9. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, Las Vegas, NV, USA, 27–30 June 2016.
10. Chung, S.W.; Han, S.S.; Lee, J.W.; Oh, K.-S.; Kim, N.R.; Yoon, J.P.; Kim, J.Y.; Moon, S.H.; Kwon, J.; Lee, H.-J.; et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop.* **2018**, *89*, 468–473. [CrossRef]
11. Tomita, N.; Cheung, Y.Y.; Hassanpour, S. Deep neural networks for automatic detection of osteopo-rotic vertebral fractures on CT scans. *Comput. Biol. Med.* **2018**, *98*, 8–15. [CrossRef]
12. Heo, M.-S.; Kim, J.-E.; Hwang, J.-J.; Han, S.-S.; Kim, J.-S.; Yi, W.-J.; Park, I.-W. Artificial intelligence in oral and maxillofacial radiology: What is currently possible? *Dentomaxillofacial Radiol.* **2021**, *50*, 20200375. [CrossRef]
13. Hung, K.; Montalvao, C.; Tanaka, R.; Kawai, T.; Bornstein, M.M. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiol.* **2020**, *49*, 20190107. [CrossRef] [PubMed]
14. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
15. Nagi, R.; Aravinda, K.; Rakesh, N.; Gupta, R.; Pal, A.; Mann, A.K. Clinical applications and performance of intelligent systems in dental and maxillofacial radiology: A review. *Imaging Sci. Dent.* **2020**, *50*, 81–92. [CrossRef]
16. Python. Available online: https://www.python.org/ (accessed on 24 June 2021).
17. PyTorch. Available online: https://pytorch.org/ (accessed on 3 July 2020).
18. Fastai. Available online: https://docs.fast.ai/ (accessed on 3 February 2021).
19. Scikit-Learn. Available online: https://scikit-learn.org/stable/ (accessed on 6 July 2020).
20. Pydicom. Available online: https://pydicom.github.io/ (accessed on 8 July 2020).
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
22. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
23. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 4700–4708.
24. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
25. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.12808.
26. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
27. Learning Rate Finder. Available online: https://fastai1.fast.ai/callbacks.lr_finder.html (accessed on 3 February 2021).
28. Howard, J.; Gugger, S. Fastai: A Layered API for Deep Learning. *Information* **2020**, *11*, 108. [CrossRef]
29. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
30. Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. *J. Comput. Mol. Des.* **2014**, *28*, 887–918. [CrossRef]

31. Murero, M. Building Artificial Intelligence for Digital Health: A socio-tech-med approach and a few surveillance night-mares. *Ethnogr. Qual. Res. Il Mulino* **2020**, *13*, 374–388.
32. Comelli, A.; Coronnello, C.; Dahiya, N.; Benfante, V.; Palmucci, S.; Basile, A.; Vancheri, C.; Russo, G.; Yezzi, A.; Stefano, A. Lung Segmentation on High-Resolution Computerized Tomography Images Using Deep Learning: A Preliminary Step for Radiomics Studies. *J. Imaging* **2020**, *6*, 125. [CrossRef]
33. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [CrossRef]
34. Pranata, Y.D.; Wang, K.-C.; Wang, J.-C.; Idram, I.; Lai, J.-Y.; Liu, J.-W.; Hsieh, I.-H. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. *Comput. Methods Programs Biomed.* **2019**, *171*, 27–37. [CrossRef]