Balancing Performance and Interpretability in Medical Image Analysis: Case study of Osteopenia

Mikulić, Mateo

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Diplomski sveučilišni studij računarstva

Diplomski rad

UTJECAJ ZBUNJUJUĆIH VARIJABLI NA INTERPRETABILNOST I PERFORMANSE MODELA DUBOKOG UČENJA NA STUDIJI SLUČAJA OSTEOPENIJE

Rijeka, srpanj 2024.

Mateo Mikulić

0069087917

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Diplomski sveučilišni studij računarstva

Diplomski rad

UTJECAJ ZBUNJUJUĆIH VARIJABLI NA INTERPRETABILNOST I PERFORMANSE MODELA DUBOKOG UČENJA NA STUDIJI SLUČAJA OSTEOPENIJE

Mentor: Prof. dr. sc. Ivan Štajduhar

Komentor: dr. sc. Franko Hržić

Rijeka, srpanj 2024.

Mateo Mikulić

0069087917

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA DIPLOMSKE ISPITE

Rijeka, 12.03.2024.

Zavod: Zavod za računarstvo

Predmet: Strojno učenje

ZADATAK ZA DIPLOMSKI RAD

Pristupnik: **Mateo Mikulić (0069087917)**

Studij: Sveučilišni diplomski studij računarstva (1400)

Modul: Programsko inženjerstvo (1441)

Zadatak: **Utjecaj zbunjujućih varijabli na interpretabilnost i performanse modela dubokog učenja na studiji slučaja osteopenije / Impact of confounding variables on the interpretability and performance of deep learning models on the case study of osteopenia**

Opis zadatka:

Razmotriti metode dubokog učenja korištene za detekciju bolesti kostiju. Za studij Osteopenije potrebno je pronaći i sakupiti skup podataka te ga opisati. U skupu podataka potrebno je detektirati zbunjujuće varijable te kroz osmišljeni test ispitati interpretabilnost i performanse odabranih modela dubokog učenja kada su modeli obučeni na skupu podataka sa zamaskiranim zbunjujućim varijablama iskupu podataka bez maskiranih zbunjujućih varijabli. Interpretabilnost modela potrebno je verificirati s medicinskim stručnjacima.

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

Zadatak uručen pristupniku: 20.03.2024.

Mentor:
prof. dr. sc. Ivan Štajduhar

Komentor:
dr. sc. Franko Hržić

Predsjednik povjerenstva za diplomski ispit:
prof. dr. sc. Miroslav Joler

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Zavod za računarstvo

dr. sc. Franko Hržić
prof. dr. sc. Ivan Štajduhar

Rijeka, 3.7.2024.

Predmet: Izvješće o sadržaju rada, obimu i složenosti istraživanja te doprinosu studenta

Znanstveni rad naslova „*Balancing Performance and Interpretability in Medical Image Analysis: Case study of Osteopenia*" prihvaćen je za objavljivanje u časopisu *„Journal of Imaging Informatics in Medicine",* rangiranom u drugoj kvartili (Q2, IF 2.9) prema indeksnoj bazi Web Of Science, u kategoriji *„Radiology, nuclear medicine & medical imaging".* Student Mateo Mikulić je prvi autor na navedenom radu, dok je njegov komentor dr. sc. Franko Hržić dopisni autor.

Objavljeni rad rezultat je jednogodišnje suradnje između Medicinskog Sveučilišta u Grazu, Zavoda za dječju radiologiju, i Tehničkog fakulteta Sveučilišta u Rijeci, tijekom koje je Mateo Mikulić ispitao vrijednu hipotezu o relevantnosti maskiranja zbunjujućih varijabli na konkretnoj studiji osteopenije. Student je samostalno proveo istraživanje te ponudio niz potencijalnih rješenja od kojih je uz pomoć mentora i komentora te savjeta radiologa izabirao ona koja su polučila najbolje rezultate. Temeljni izvor podataka za istraživanje jest javno dostupan skup podataka, GRAZPEDWRI-DX, koji obuhvaća više od 20,000 radiograma zapešća.

Naslov formalnog zadatka za diplomski rad studenta na hrvatskom jeziku glasi: *Utjecaj zbunjujućih varijabli na interpretabilnost i performanse modela dubokog učenja na studiji slučaja osteopenije.* Naslov znanstvenog rada preveden na hrvatski jezik glasi: *Balansiranje izvedbe i interpretabilnosti u analizi medicinske slike: studija slučaja osteopenije.* Znanstveni rad sadržajno u potpunosti odgovara zadatku diplomskom rada.

Kao glavni cilj diplomskog rada navodi se ispitivanje zbunjujućih varijabli na performanse modela strojnoga učenja na slučaju osteopenije gdje će se konačne performanse usporediti i vrednovati od strane medicinskih stručnjaka. Svrha istraživanja je ponuditi novi pogled na ravnotežu između performansi modela mjerenih klasičnim metrikama, te rangiranja modela od strane radiologa koji u obzir uzimaju interpretabilnost modela s obzirom na fokus modela tijekom zaključivanja. Kako bi ostvario navedeni cilj, student je morao izvršiti sljedeće:

- Filtrirati skup podataka te dodatno označiti lažne maske koje su detektirane kao jedno od rješenja za pretjeranu prilagodbu modela.

- U suradnji s radiolozima detektirati zbunjujuće varijable na medicinskom slučaju osteopenije.
- Odabrati i istrenirati dovoljan broj modela dubokog učenja kako bi izvedeni zaključci bili pravovaljani i statistički signifikantni.
- Implementirati metodu GradCAM kao alat za interpretabilnost modela dubokog učenja.
- Osmisliti i provesti test u kojem radiolozi validiraju istrenirane modele.
- Izvesti zaključke o provedenom testiranju.
- Dokumentirati metodologiju, rezultate i opažanja.

Na temelju navedenoga, mišljenja smo da je student Mateo Mikulić sustavno proveo potrebne aktivnosti i kvalitetno obradio temu zadanu za diplomski rad. Uzimajući dodatno u obzir njegovu proaktivnost i ažurnost tijekom pripreme znanstvenog rada na engleskom jeziku, kao i složenost materije, ne nalazimo razloga za odbacivanje njegove molbe. Slijedom navedenog, **suglasni smo s priznavanjem predmetnog znanstvenog članka kao njegovog diplomskog rada**.

Franko Hržić

Ivan Štajduhar

IZJAVA O SAMOSTALNOJ IZRADI RADA

Izjavljujem da sam samostalno izradio diplomski rad.

U Rijeci, 9. srpnja, 2024.

Mateo Mikulić

ZAHVALA

Zahvaljujem dr. sc. Franku Hržiću na trudu i vremenu koje je uložio u mentorstvo, kao i na velikoj podršci i prijateljskim savjetima. Također, zahvaljujem obitelji i Martini na ogromnoj podršci koju su mi pružili tijekom studija. Hvala svima koji su sudjelovali u postupku izrade ovog znanstvenog rada.

Balancing Performance and Interpretability in Medical Image Analysis: Case study of Osteopenia

Mateo Mikulić[a], Dominik Vičević[a], Eszter Nagy[b], Mateja Napravnik[a,c], Ivan Štajduhar[a,c], Sebastian Tschauner[b], and Franko Hržić[a,c,*]

a) University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Vukovarska 58, Rijeka 51000, Croatia
b) Medical University of Graz, Department of Radiology, Division of Pediatric Radiology, Graz 8036, Austria
c) University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Radmile Matejčić 2, Rijeka 51000, Croatia
*) Corresponding author: Franko Hržić (e-mail: franko.hrzic@uniri.hr, tel: +385 51 505725)
Orcids: Eszter Nagy (0000-0001-7080-5996), Mateja Napravnik (0000-0002-3271-3342), Ivan Štajduhar (0000-0003-4758-7972), Sebastian Tschauner (0000-0002-7873-9839), Franko Hržić (0000-0003-1513-0337)

Abstract

Multiple studies within the medical field have highlighted the remarkable effectiveness of using convolutional neural networks for predicting medical conditions, sometimes even surpassing that of medical professionals. Despite their great performance, convolutional neural networks operate as black boxes, potentially arriving at correct conclusions for incorrect reasons or areas of focus. Our work explores the possibility of mitigating this phenomenon by identifying and occluding confounding variables within images. Specifically, we focused on the prediction of osteopenia, a serious medical condition, using the publicly available GRAZPEDWRI-DX dataset. After detection of the confounding variables in the dataset, we generated masks that occlude regions of images associated with those variables. By doing so, models were forced to focus on different parts of the images for classification. Model evaluation using F1-score, precision, and recall showed that models trained on non-occluded images typically outperformed models trained on occluded images. However, a test where radiologists had to choose a model based on the focused regions extracted by the GRAD-CAM method showcased different outcomes. The radiologists' preference shifted towards models trained on the occluded images. These results suggest that while occluding confounding variables may degrade model performance, it enhances interpretability, providing more reliable insights into the reasoning behind predictions. The code to repeat our experiment is available on the following link: https://github.com/mikulicmateo/osteopenia.

Keywords

Artificial Intelligence, Bias Mitigation, Image Processing, Interpretable Decision Making, Occlusion Learning, Osteopenia

Statements and Declarations

Acknowledgements

This work has been supported by the Croatian Science Foundation [grant number IP-2020-02-3770] and by the University of Rijeka [grant uniri-iskusni-tehnic-23-12 2947, and uniri-mladi-tehnic-23-19 3070].

Author Contributions

The authors of this research made significant contributions -- especially in the conceptualization of research. Eszter Nagy and Sebastian Tschauner verified, interpreted and labelled the data. Mateo Mikulić and Dominik Vičević developed the necessary software and monitored model training. Ivan Štajduhar and Franko Hržić secured funding and supervised the research. Mateja Napravnik and Mateo Mikulić conducted statistical tests and data analysis. Franko Hržić, Sebastian Tschauner, and Dominik Vičević wrote the initial draft, while Eszter Nagy, Ivan Štajduhar, and Mateo Mikulić reviewed and edited it.

Funding

This work has been supported by the Croatian Science Foundation [grant number IP-2020-02-3770] and by the University of Rijeka [grant uniri-iskusni-tehnic-23-12 2947, and uniri-mladi-tehnic-23-19 3070].

Ethics approval

In the conducted research, the publicly available dataset was leveraged which has all necessary ethical approvals presented in the associated manuscript available at: 10.1038/s41597-022-01328-z

Conflict of Interest

All authors declare that they have no conflict of interest: either financial interest (such as honoraria, educational grants, participation in speakers' bureaus, membership, employment, consultancies, stock ownership, or other equity interest, and expert testimony or patent-licensing arrangements) or nonfinancial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject discussed in presented manuscript.

Introduction

Osteopenia is a medical condition marked by suboptimal bone mass density (BMD). As per the World Health Organisation (WHO), osteopenia is characterised by a BMD with values ranging from 1.0 to 2.5 standard deviation below that of a "young normal adult" (T-score) [1]. Osteopenia is a medical condition that precedes osteoporosis, a more well-known and severe condition characterised by a T-score of ≤-2.5 [2-4]. Even though the risk of fracture is higher in individuals with osteoporosis compared to those with osteopenia, the considerably larger number of individuals with osteopenia signifies that this group constitutes a significant proportion of the population susceptible to fractures [5]. According to a study conducted by Pasco J. A. et al. [6], patients with prevalent fracture and osteopenia were found to have the same, if not greater, risk of subsequent fracture as patients with osteoporosis. Typical fractures in the hip, spine, and various non-vertebral areas are linked to elevated morbidity and mortality. The risk is more pronounced in the fracture's immediate aftermath than in later stages [7, 8]. This highlights the severity of osteopenia and the need for fast and effective detection and treatment. Convolutional neural networks (CNNs) have proven to be highly effective in medical imaging tasks, including fracture detection [9] and tumour detection [10, 11] in radiology images. B. Zhang et al. [12] conducted a study using X-ray images of the lumbar spine to train CNN models for predicting osteoporosis and osteopenia. They concluded that their deep learning model shows promising results in clinical tasks where BMD screening has not been performed, and X-ray images are readily available. According to a study [13] that utilised CT scan images, deep-learning models were found to accurately classify between "normal" osteopenia and osteoporosis, as well as predict BMD values. In a study [14] where researchers used X-ray chest images to train a deep-learning model, namely ResNet50, results were inferior compared to similar studies when classifying osteoporosis. Researchers also insisted that their model was not ready for clinical application and needed further improvement. The authors stressed that there were no studies to compare classification results of "normal" and osteopenia classes. The researchers in a study[15] divided their data into only two classes, "osteoporosis" and "non-osteoporosis," because they found it hard to differentiate osteopenia and osteoporosis in classification tasks.

For machine learning (ML) models to be successfully integrated into clinical practice, they must have high reliability and minimal errors. However, to earn the trust of medical practitioners, it is crucial to ensure that the models are interpretable, making it possible to understand the decisions they make [16]. Interpretable ML models make understanding and explaining the decision-making behind diagnoses and treatment recommendations possible [17,18]. Hence, interpretability is crucial in building trust and transparency between medical practitioners who rely on ML models and patients [19]. When medical practitioners can understand the decisions made by these models, it helps to prevent errors and raise their trust in them. Due to the difficulty of interpreting their decision-making process, neural networks are often called "black-box" ML models [20]. This is why a tremendous scientific effort has been put into enhancing the interpretability of both CNNs, and neural networks in general [21]. One such method is Gradient-weighted Class Activation Mapping (GradCAM) [22], which helps to visualise the decision-making process of a CNN by highlighting the regions of an input image that contribute the most towards the model's predictions.

However, it has been shown that deep learning models can pick up on unintended or unhelpful patterns in the data used for model training. In the literature, when a ML model learns to make predictions based on features other than the target feature, it is often referred to as confounding bias or confounding variables [23-25]. Zech et al. [23] found that pneumonia detection in chest radiographs was affected by internal hospital-specific biases. Also, laterality markings in radiographs were found to have mild negative effects. Based on research [26], CNNs are susceptible to ubiquitous confounding image features, like radiograph labels, when trained to detect bone abnormalities. The authors recommend covering these kinds of features when training a CNN model. According to a study [27], researchers have discovered that hidden stratification can lead to more than 20% performance differences in clinically important subsets. This can occur in unidentified imaging subsets with low prevalence, low label quality, subtle distinguishing features, or spurious correlates. One of the notable observations was in the CXR14 dataset [28], where a significant number of X-rays in the pneumothorax class revealed the existence of chest drains. It is essential to note that chest drains are not a causal factor in diagnosing pneumothorax. Moreover, the presence of chest drains indicates that these pneumothorax cases have already been treated and pose almost no risk of harm related to pneumothorax.

In the context of X-ray image-based osteopenia classification, the issue of confounding bias may arise due to factors other than the radiograph labels themselves. These factors may include bone abnormalities such as periosteal reaction, fractures, and metal insertions (used to aid in proper bone growth after significant fractures). Given the mentioned potential confounding variables, we opted to use GRAZPEDWRI-DX [29], a thoroughly documented public dataset that suits our research cause.

The hypothesis we wanted to research is that the occlusion of confounding variables' features would improve the models' performance, similar to the covering mentioned above [26]. We also speculate that adding an approach where we occlude "dummy" confounding variable features (randomly masked areas on the bones in the image) can further improve model interpretability. To confirm the stated hypotheses, a test with two radiologists was specifically tailored in addition to the classical ML algorithm evaluation. During the test, the radiologists were presented with an original X-ray image and X-ray images covered with a GradCAM heatmap by a model trained with occluded, and without occluded features. The radiologists must select a better heatmap, if there is any difference in the quality of heatmaps produced by the models, or mark them as "equal" if there is no difference in quality. This whole process is summarised in Fig. 1.

Materials and Methods

This section introduces the publicly available dataset GRAZPEDWRI-DX [29] and specifies the utilized modes and confounding variables required for reproducing our research. The necessary code, models, and demonstrative examples are available at the following GitHub repository https://github.com/mikulicmateo/osteopenia.

Utilised dataset: GRAZPEDWRI-DX

The chosen data source is GRAZPEDWRI-DX, a public, well-documented dataset validated by three radiologists with between 6 and 29 years of experience in musculoskeletal radiology [29]. The GRAZPEDWRI-DX dataset consists of annotated paediatric trauma wrist radiographs of 6,091 patients. 10,643 total studies (20,327 images) are made available, typically covering postero-anterior and lateral projections. The dataset is annotated with 74,459 image tags, and features 67,771 labelled objects [29]. We further adapted this dataset by only considering patients who have had osteopenia at one point during the study, which reduces the number of patients to 1,051. We also excluded images of cast-covered wrists, as the cast can further hinder the detection of osteopenia. From now on, we will refer to this adapted dataset as the filtered dataset. Fig. 2 shows the distribution of osteopenia within the filtered dataset.

Fig. 3 shows the Spearman correlation coefficients and distribution of correlated labels and potentially confounding variables in the filtered dataset. The Subfigure 3a shows the percentage of data which includes that label, and the Subfigure 3b shows the Spearman correlation coefficients of the respective label in relation to the presence of osteopenia.

The initial exam label denotes whether the radiograph is part of the patient's initial study. With the Spearman correlation coefficient of -0.82, a very strong negative correlation is observed between the initial exam label and the presence of osteopenia; thus, if a radiograph is part of the initial exam, it is very likely that osteopenia is not present. The study number label denotes the number of studies done for the patient in question and increases for each subsequent study. The Spearman correlation coefficient of 0.44 shows a moderate positive correlation between the study number label and the presence of osteopenia. This means that as the study number increases, the presence of osteopenia becomes more likely.

Based on the information above, osteopenia is not likely to be present in the initial exam but is more likely to be observed during subsequent exams of the same patient. From this, we can conclude that osteopenia arises more frequently after trauma.

The latter three labels (fracture, metal, periosteal reaction) in Fig. 3 were identified as potentially confounding variables. These variables, if present, could potentially lead to wrong classifications since their presence alone may decide the model's output. The fracture variable denotes if fractures are present on the radiograph, the metal variable denotes any internal or external metal implants, and the periosteal reaction variable denotes if a periosteal reaction is present. Periosteal reaction results when cortical bone reacts to insults [30] (in this case, trauma). Initially, the models were trained without occlusions or image augmentations. Afterwards, these models were trained on images with occluded fractures. We identified the fractures as a potentially confounding variable based on the fracture labels' prevalence in the dataset and the aforementioned tests. The metal label, closely associated with the fracture label, is present exclusively in studies after the occurrence of a fracture. Consequently, it was also recognised as a potentially confounding variable. Lastly, the periosteal reaction label was recognised as a potentially confounding variable because of its close link to fractures (as periosteal reaction only occurs after trauma), and because of its moderate positive correlation to the presence of osteopenia.

In earlier testing of models using GradCAM heatmaps, we noticed that models focus on the text present in the image (every radiograph contains the uppercase letter 'R' or 'L'). This coincides with research [26]. Because of the above reasons, text was also included as a confounding variable.

All mentioned confounding variables are annotated in the GRAZPEDWRI-DX dataset with their spatial positioning coordinates on the radiograph. Using this, we decided to occlude these variables by masking the area around them. The exact occlusion process is described in the following subsection.

Occlusion of confounding variables

Since not all radiographs contain the fracture, mask, and periosteal reaction variables, we artificially added "dummy" masks. Dummy masks were annotated by hand but in a random fashion. Each image was manually inspected, and a hand-drawn mask was added to a random location in each radiograph. The only prerequisite to adding these dummy masks was that they at least partially cover bones and/or bone parts. While adding these masks, special attention was given to their shape to look as similar as possible to the real annotations for the respective label. This was done to ensure no potential bias regarding the annotation mask shape.

For all the images in the filtered dataset that did not have some or all confounding variables, dummy occlusions were manually added so that each image could have every type of variable occlusion. These occlusions were the same shape as the real occlusions of that confounding variable, not to give away their validity. An example of all types of occlusion can be seen in Fig. 4.

Since it is undesirable to have all of the feature occlusions present on every X-ray image, their presence was decided randomly. The probability of application of any particular occlusion was derived using grid search. These probabilities can be found in Table 1. As a result, any single X-ray image could have any combination of different occluded features present, and these combinations could differ between epochs.

Table 1 Probabilities for applying occlusion

Confounding variable	Mask shape	Probability of application
Text	Rectangle	50%
Fracture	Rectangle	33%
Metal	Rectangle	50%
Periosteal reaction	Polygon	25.5%

Utilised models

Although CNN models have led to remarkable progress in computer vision tasks, their results are often difficult to interpret. The interpretability of models' outcomes is essential to effectively incorporate ML into medicine. It is necessary to identify and correct avoidable errors and to alleviate any concerns patients may have. For example, confounding bias would be difficult to detect without interpretability. GradCAM, a technique developed to produce "visual explanations" for decisions made by CNNs, is very helpful in identifying such bias in model learning [22]. When identified, it is factored into data preparation and augmentation. GradCAM is inspired by Class Activation Mapping (CAM) [31], which uses global average pooling and fully connected layers to produce a binary heatmap. GradCAM is a generalisation of that technique, does not alter the architecture of a model, and does not require re-training of the model. GradCAM works by computing the gradient of the predicted class score for the feature maps in a CNN. After global average pooling, the weighted sum of feature maps is obtained, emphasising important regions. A ReLU activation is applied to focus on positive contributions, and the resulting heatmap is upsampled to the input image size. Selvaraju et al. [22] show how GradCAM identifies biases in datasets. Additionally, they conducted a study where they found that GradCAM helps humans establish trust with ML models and helps even untrained users recognise a more robust performing model from a weaker one, even when predictions are equal.

For the osteopenia classification task, we utilised a range of CNN architectures with varying levels of complexity. These CNNs were selected based on their ability to effectively learn features from images and perform classification tasks with high accuracy. The selected models are as follows:

- As the first model we decided to test a simpler, yet very relevant, CNN architecture, *VGG* [32]. VGG advanced the state-of-the-art by using small 3x3 convolution filters and deeper networks. Such architecture enabled the model to capture more information while maintaining fewer parameters than models with larger filters but fewer layers. Earlier layers capture low-level features, and deeper layers capture higher-level concepts like textures. The architecture can be deepened if needed, which is great for research. VGG19 architecture was trained.

- As networks get deeper, the problem of vanishing and exploding gradients occurs more frequently, as well as the degradation problem [33-35]. Researchers have found that introducing skip connections helps with tackling these problems. Residual Networks (*ResNet*) [36] were the first architecture to introduce skip connections. Skip connections bypass layers of a neural network to make it easier for the network to learn identity mappings. ResNet34 and ResNet50 architectures were trained.
- Unlike summation used in ResNet architectures, Densely Connected CNNs (*DenseNets*) [37] utilise skip connections with concatenation, where each layer receives feature maps of all previous layers to improve information flow between layers further. DenseNet169 and DenseNet201 architectures were trained.

All employed models were pretrained with adjusted classifier layers to suit our classification objective. Transfer learning was used for faster convergence. Models were trained using the progressive unfreezing method in four stages with equal training conditions. The optimiser used was AdamW, with a learning rate of 1.65e-5. The learning rate was chosen using random grid search, where the tested range of values was [1e-4, 3e-6]. The loss function used was binary cross entropy.

Data augmentations

Image values were normalised to the distribution of the ImageNet dataset [38] on which all utilised models were pretrained. Additionally, some standard image augmentation techniques were used to artificially increase dataset size, thus preventing models from overfitting:

- *Rotation angle*: Each image in the dataset is accompanied by an annotation of a two-point line delineating the primary axis of the forearm bones. Given that X-ray images are often misaligned, we leveraged this axis annotation to centre the images along the edges. Subsequently, a random rotation of up to 20 degrees in either direction was applied with a probability of 0.5 to the centred image. The purpose of this augmentation was to mitigate the inherent bias introduced by variations in the positioning of the forearm during X-ray imaging. Since multiple images of the same individual were available in the dataset, this approach effectively reduced the confounding effects of inter-subject variability.
- *Colour augmentations*: A random factor within the range of 0.8 to 1.2 was used to jitter the brightness, while a range of 0.8 to 1.3 was used to jitter the contrast. The hue and saturation were adjusted by a random factor uniformly chosen between [-0.5,0.5].
- *Horizontal flip* was applied with a probability factor of 0.5.

Employed models were first trained using only the augmentations mentioned above. After this, in addition to the described augmentation, the models were trained using the hypothesised method: occlusion of confounding variables.

Radiologists' test

To test if the occlusion of confounding variables would enhance the models' performance, a blind test was carried out with the participation of two experienced radiologists. During the test, the radiologists were presented with three images arranged side-by-side: (i) the original, unchanged radiograph, (ii) the same X-ray image covered with a GradCAM heatmap of a model trained without occluded features, and (iii) also the same radiograph, but this time covered with a GradCAM heatmap of a model trained with feature occlusion. The radiologists were required to select the image depicting osteopenia most accurately. This test was conducted on 50 X-ray images from the test subset, all with osteopenia present. An example of the test query for the radiologists is shown in Fig. 5.

Results

Each model underwent twenty training rounds, with ten rounds conducted without occlusion of confounding variables and the other ten rounds with occluded confounding variables. Table 2 showcases the best result of each version of the model on the test subset. Table 3 shows the average result for each model when trained with and without confounding variables occlusion for ten rounds. The configuration of Table 3 is the same as Table 2, with the only difference being that Table 3 has the averaged results along with the standard deviation for all rounds of training of each version. All averaged results were statistically tested for significance with a two-tailed paired t-test.

Table 2 Results of best-performing models on the test subset. The left side of the table shows results for the model trained without occluded confounding variables. The right side of the table is reserved for the results of models trained with occluded confounding variables. Each column describes values for one performance metric. The metrics described are accuracy, precision, recall, and F1-score, respectively. Bolded values represent better performance

	Not occluded				Occluded			
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
VGG19	91.18%	90.00%	97.10%	93.41%	**91.71%**	**91.34%**	96.27%	**93.74%**
ResNet34	**92.78%**	**92.46%**	96.68%	**94.52%**	91.18%	90.31%	96.68%	93.39%
ResNet50	**89.04%**	88.46%	**95.44%**	**91.82%**	88.24%	**88.63%**	93.78%	91.13%
DenseNet169	**93.05%**	**92.16%**	97.51%	**94.76%**	91.18%	88.81%	**98.76%**	93.52%
DenseNet201	90.64%	**89.62%**	96.68%	93.01%	**90.91%**	89.06%	**97.93%**	**93.28%**

Table 4 depicts the outcomes of the blind radiologists' test. After the blind-test phase, the judgement of radiologists regarding the superior classifier was considered the ground truth for subsequent statistical analysis using the McNemar test. The McNemar test was used as a robust analytical instrument to ascertain the statistical meaningfulness of the observed differences in the model's performance [39,49]. Table 5 represents the p-values of the McNemar statistical test.

Table 3 Averaged results (10 trainings) on the test subset. Values where the difference is statistically significant (t-test, p-value < 0.05) are marked in bold

	Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Not occluded	**VGG19**	90.05 ± 0.73%	90.04 ± 0.95%	95.10 ± 0.89%	92.50 ± 0.54%
	ResNet34	**90.99 ± 0.98%**	**89.54 ± 1.36%**	97.43 ± 1.42%	**93.31 ± 0.72%**
	ResNet50	**88.00 ± 0.65%**	**89.29 ± 1.16%**	92.49 ± 1.63%	**90.85 ± 0.53%**
	DenseNet169	**91.42 ± 0.75%**	90.39 ± 1.08%	97.01 ± 0.85%	**93.58 ± 0.54%**
	DenseNet201	89.57 ± 0.52%	**88.67 ± 1.16%**	96.14 ± 1.37%	92.24 ± 0.37%
Occluded	**VGG19**	**90.72 ± 0.69%**	90.03 ± 0.86%	**96.27 ± 0.90%**	**93.04 ± 0.51%**
	ResNet34	89.55 ± 1.28%	87.29 ± 1.79%	98.13 ± 1.33%	92.37 ± 0.86%
	ResNet50	86.93 ± 1.23%	87.60 ± 1.43%	92.90 ± 1.91%	90.15 ± 0.95%
	DenseNet169	89.84 ± 1.08%	89.29 ± 1.02%	95.77 ± 3.13%	92.37 ± 1.01%
	DenseNet201	89.39 ± 0.91%	87.46 ± 1.38%	97.55 ± 1.18%	92.22 ± 0.62%

Table 4 Results of radiologists' tests

	DenseNet169			VGG19		
	Not occluded better	**Occluded better**	**Equal**	**Not occluded better**	**Occluded better**	**Equal**
Radiologist 1	22% (11 images)	**58% (29 images)**	20% (10 images)	4% (2 images)	**66% (33 images)**	30% (15 images)
Radiologist 2	18% (9 images)	**60% (30 images)**	22% (11 images)	14% (7 images)	**74% (37 images)**	12% (6 images)

Table 5 McNemar test results. The rows represents radiologists and the columns depict evaluated models. For example, the first row and the first column mark the p-value of the McNemar test for “Radiologist 1” and the DenseNet169 model. The tested null-hypothesis is “occluded versus non-occluded model performance is insignificant.”

	DenseNet169 p-value	**VGG19 p-value**
Radiologist 1	0.00443	$1.606 \cdot 10^{-7}$
Radiologist 2	0.00077	$6.106 \cdot 10^{-6}$

Discussion

From Table 2, we can see that measured performance metrics favour the model version without the occlusions in cases of both versions of ResNet and DenseNet169. VGG19 and DenseNet201, however, perform better when trained with occluded confounding variables. In Table 3, similar trends follow, except for DenseNet201, whose average performance leans on the side where confounding variables were not occluded. Our study aimed to explore the hypothesis that the model would exhibit improvement through training with occluded confounding variables. To achieve this, we identified the models that demonstrated the highest accuracy for both "occluded" and "not occluded" training scenarios. Specifically, DenseNet169 achieved the highest F1 score in the "not occluded" training category. In contrast, VGG19 performed the best in the "occluded" category.

As it was shown that the model could learn the right choice for the wrong reason, this means that the F1 score could not be considered as the sole metric for model performance. To add context to the F1 score, we opted for the visualisation tool GradCAM. The GradCAM technique superimposes heatmaps on the original X-ray images to depict the models' areas of focus. Examples of the GradCAM heatmaps can be seen in Fig. 6.

From GradCAM heatmaps, it is clear that the models with occluded confounding variables focus more on the important regions of bones in the radiographs, which are closer to what a human might focus on. To test this further, we performed a blind test with two radiologists. In Table 4, we see that the radiologists consider the models with occluded training type to be the better classifiers in most cases. As per the assessment made by the radiologists, the DenseNet model trained with occluded features was deemed superior in 58% and 60% of cases respectively for each radiologist. The VGG19 model trained with occluded features outperformed its counterpart in 66% and 74% of the cases.

The results of the blind test (images with the superior classifier) were used as ground truth for subsequent statistical analysis using the McNemar test. The null-hypothesis being tested was that the difference between model predictions is not statistically significant. Table 5 shows the results of said test. Since the p-values of the McNemar test for both models are less than 0.05, we reject the null-hypothesis and conclude that the difference in the performance of the models is significant.

In summary, while common ML model evaluation metrics may indicate better performance in non-occluded models, models with explainable predictions are deemed more valuable from the perspective of medical practitioners (in this case, radiologists). The case study presented herein on the classification of osteopenia underscores the importance of integrating medical domain knowledge into the development of ML models. This highlights the need for close collaboration with medical professionals. Moreover, training models with occlusions may be helpful in other domains where confounding variables pose a challenge. An important observation to note is that the accuracy and F1-score of all models in this study (both trained with and without occlusion) surpass the results of all models presented in a paper [41] that used the same GRAZPEDWRI-DX dataset to classify osteopenia.

Conclusions

This study tested the hypothesis that the occlusion of confounding variables' features enhances predictive models' performance. On the one hand, the conventional metrics favoured models trained on non-occluded data over models trained on occluded data. On the other hand, radiologists disagreed with that conclusion. This case study emphasises the importance of incorporating medical expertise into ML algorithms. It highlights the need for collaboration between medical professionals and ML specialists. As a potential avenue for further research, additional experiments involving models with occluded confounding variables within alternative problem domains should be conducted. This would help better understand the potential impact of training models with occluded confounding variables on their performance.

References

[1] G. Karaguzel and M. F. Holick, “Diagnosis and treatment of osteopenia,” Reviews in Endocrine and Metabolic Disorders, vol. 11, no. 4, p. 237–251, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1007/s11154-010-9154-0

[2] K. G. Faulkner, “Update on bone density measurement,” Rheumatic Disease Clinics of North America, vol. 27, no. 1, p. 81–99, Feb. 2001. [Online]. Available: http://dx.doi.org/10.1016/s0889-857x(05)70188-5

[3] R. POLLYCOVE and J. A. SIMON, “Osteoporosis: Screening and treatment in women,” Clinical Obstetrics & Gynecology, vol. 55, no. 3, p. 681–691, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1097/GRF.0b013e31825caa50

[4] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, “Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network,” European Radiology, vol. 30, no. 6, p. 3549–3557, Feb. 2020. [Online]. Available: http://dx.doi.org/10.1007/s00330-020-06677-0

[5] S. Khosla and L. J. Melton, “Osteopenia,” New England Journal of Medicine, vol. 356, no. 22, p. 2293–2300, May 2007. [Online]. Available: http://dx.doi.org/10.1056/NEJMcp070341

[6] J. A. Pasco, E. Seeman, M. J. Henry, E. N. Merriman, G. C. Nicholson, and M. A. Kotowicz, “The population burden of fractures originates in women with osteopenia, not osteoporosis,” Osteoporosis International, vol. 17, no. 9, p. 1404–1409, May 2006. [Online]. Available: http://dx.doi.org/10.1007/s00198-006-0135-9

[7] E. F. Eriksen, “Treatment of osteopenia,” Reviews in Endocrine and Metabolic Disorders, vol. 13, no. 3, p. 209–223, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1007/s11154-011-9187-z

[8] O. Johnell, J. A. Kanis, A. Od´en, I. Sernbo, I. Redlund-Johnell, C. Petterson, C. De Laet, and B. J¨onsson, “Mortality after osteoporotic fractures,” Osteoporosis International, vol. 15, no. 1, p. 38–42, Oct. 2003. [Online]. Available: http://dx.doi.org/10.1007/s00198-003-1490-4

[9] F. Hržić, S. Tschauner, E. Sorantin, and I. Štajduhar, “Fracture recognition in paediatric wrist radiographs: An object detection approach,” Mathematics, vol. 10, no. 16, p. 2939, Aug. 2022. [Online]. Available: http://dx.doi.org/10.3390/math10162939

[10] M. Khairandish, M. Sharma, V. Jain, J. Chatterjee, and N. Jhanjhi, “A hybrid cnn-svm threshold segmentation approach for tumor detection and classification of mri brain images,” IRBM, vol. 43, no. 4, pp. 290–299, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031821000713

[11] T. Hossain, F. S. Shishir, M. Ashraf, M. A. A. Nasim, and F. M. Shah, “Brain tumor detection using convolutional neural network,” 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1–6, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209456854

[12] B. Zhang, K. Yu, Z. Ning, K. Wang, Y. Dong, X. Liu, S. Liu, J. Wang, C. Zhu, Q. Yu, Y. Duan, S. Lv, X. Zhang, Y. Chen, X. Wang, J. Shen, J. Peng, Q. Chen, Y. Zhang, X. Zhang, and S. Zhang, “Deep learning of lumbar spine x-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study,” Bone, vol. 140, p. 115561, Nov. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.bone.2020.115561

[13] T. Peng, X. Zeng, Y. Li, M. Li, B. Pu, B. Zhi, Y. Wang, and H. Qu, “A study on whether deep learning models based on ct images for bone density classification and prediction can be used for opportunistic osteoporosis screening,” Osteoporosis International, ol. 35, no. 1, p. 117–128, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1007/s00198-023-06900-w

[14] Y. Sato, N. Yamamoto, N. Inagaki, Y. Iesaki, T. Asamoto, T. Suzuki, and S. Takahara, “Deep learning for bone mineral density and t-score prediction from chest x-rays: A multicenter study,” Biomedicines, vol. 10, no. 9, p. 2323, Sep. 2022. [Online]. Available: http://dx.doi.org/10.3390/biomedicines10092323

[15] R. Jang, J. H. Choi, N. Kim, J. S. Chang, P. W. Yoon, and C.-H. Kim, “Prediction of osteoporosis from simple hip radiography using deep learning algorithm,” Scientific Reports, vol. 11, no. 1, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1038/s41598-021-99549-6

[16] R. J. Woodman and A. A. Mangoni, “A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future,” Aging Clinical and Experimental Research, vol. 35, no. 11, p. 2363–2397, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1007/s40520-023-02552-2

[17] A. I. F. Poon and J. J. Y. Sung, “Opening the black box of ai-medicine,” Journal of Gastroenterology and Hepatology, vol. 36, no. 3, p. 581–584, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1111/jgh.15384

[18] P. Lisboa, S. Saralajew, A. Vellido, R. Fern´andez-Domenech, and T. Villmann, “The coming of age of interpretable and explainable machine learning models,” Neurocomputing, vol. 535, p. 25–39, May 2023. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2023.02.040

[19] A. Vellido, “The importance of interpretability and visualization in machine learning for applications in medicine and health care,” Neural Computing and Applications, vol. 32, no. 24, p. 18069–18083, Feb. 2019. [Online]. Available: http://dx.doi.org/10.1007/s00521-019-04051-w
[20] F. Fan, J. Xiong, M. Li, and G. Wang, “On interpretability of artificial neural networks: A survey,” 2020. [Online]. Available: https://arxiv.org/abs/2001.02522
[21] E. Sorantin, M. G. Grasser, A. Hemmelmayr, S. Tschauner, F. Hrzic, V. Weiss, J. Lacekova, and A. Holzinger, “The augmented radiologist: artificial intelligence in the practice of radiology,” Pediatric Radiology, vol. 52, no. 11, p. 2074–2086, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1007/s00247-021-05177-7
[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
[23] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” PLOS Medicine, vol. 15, no. 11, p. e1002683, Nov. 2018. [Online]. Available: http://dx.doi.org/10.1371/journal.pmed.1002683
[24] R. T. Tomihama, J. R. Camara, and S. C. Kiang, “Machine learning analysis of confounding variables of a convolutional neural network specific for abdominal aortic aneurysms,” JVS-Vascular Science, vol. 4, p. 100096, 2023. [Online]. Available: http://dx.doi.org/10.1016/j.jvssci.2022.11.004
[25] S. Venugopalan, A. Narayanaswamy, S. Yang, A. Geraschenko, S. Lipnick, N. Makhortova, J. Hawrot, C. Marques, J. Pereira, M. Brenner, L. Rubin, B. Wainger, and M. Berndl, “It’s easy to fool yourself: Case studies on identifying bias and confounding in bio-medical datasets,” 2019. [Online]. Available: https://arxiv.org/abs/1912.07661
[26] P. H. Yi, P. S. Malone, C. T. Lin, and R. W. Filice, “Deep learning algorithms for interpretation of upper extremity radiographs: Laterality and technologist initial labels as confounding factors,” American Journal of Roentgenology, vol. 218, no. 4, p. 714–715, Apr. 2022. [Online]. Available: http://dx.doi.org/10.2214/AJR.21.26882
[27] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, “Hidden stratification causes clinically meaningful failures in machine learning for medical imaging,” in Proceedings of the ACM Conference on Health, Inference, and Learning, ser. ACM CHIL ’20. ACM, Apr. 2020. [Online]. Available: http://dx.doi.org/10.1145/3368555.3384468
[28] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.369
[29] E. Nagy, M. Janisch, F. Hržić, E. Sorantin, and S. Tschauner, “A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning,” Scientific Data, vol. 9, no. 1, May 2022. [Online]. Available: http://dx.doi.org/10.1038/s41597-022-01328-z
[30] R. S. Rana, J. S. Wu, and R. L. Eisenberg, “Periosteal reaction,” American Journal of Roentgenology, vol. 193, no. 4, p. W259–W272, Oct. 2009. [Online]. Available: http://dx.doi.org/10.2214/AJR.09.3300
[31] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.319
[32] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014. [Online]. Available: https://arxiv.org/abs/1409.1556
[33] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html
[34] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE Transactions on Neural Networks, vol. 5, no. 2, p. 157–166, Mar. 1994. [Online]. Available: http://dx.doi.org/10.1109/72.279181
[35] K. He and J. Sun, “Convolutional neural networks at constrained time cost,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2015. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7299173

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2016. [Online]. Available:http://dx.doi.org/10.1109/CVPR.2016.90
[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.243
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, may 2017.
[39] N. Japkowicz and M. Shah, Performance Evaluation in Machine Learning. Springer International Publishing, 2015, p. 41–56. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18305-3 4
[40] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” Neural Computation, vol. 10, no. 7, p. 1895–1923, Oct. 1998. [Online]. Available: http://dx.doi.org/10.1162/089976698300017197
[41] I. M. Wani and S. Arora, “Osteoporosis diagnosis in knee x-rays by transfer learning based on convolution neural network,” Multimedia Tools and Applications, vol. 82, no. 9, p. 14193–14217, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1007/s11042-022-13911-y

Figures

All figures were made using: matplotlib, GIMP, Inkscape and diagrams.net.

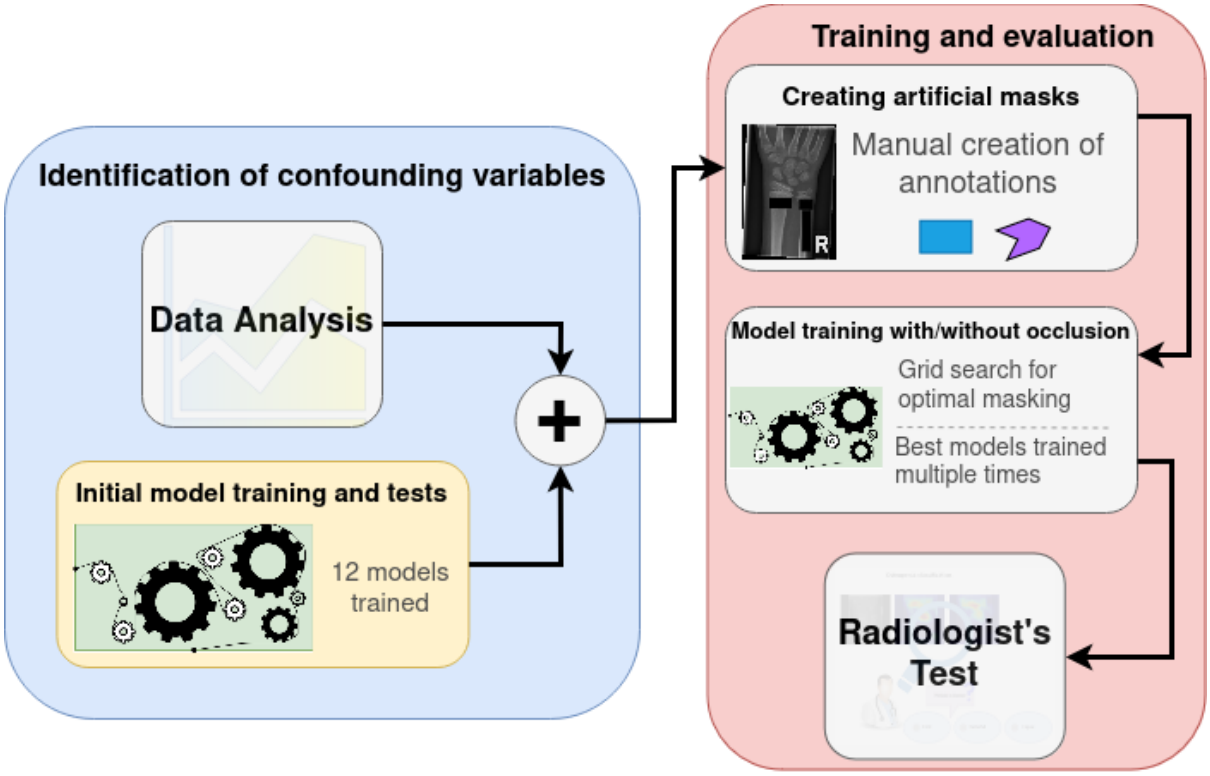

Fig. 1 An overview of the conducted research process. The research can generally be divided into two parts: detection of confounding variables and their evaluation

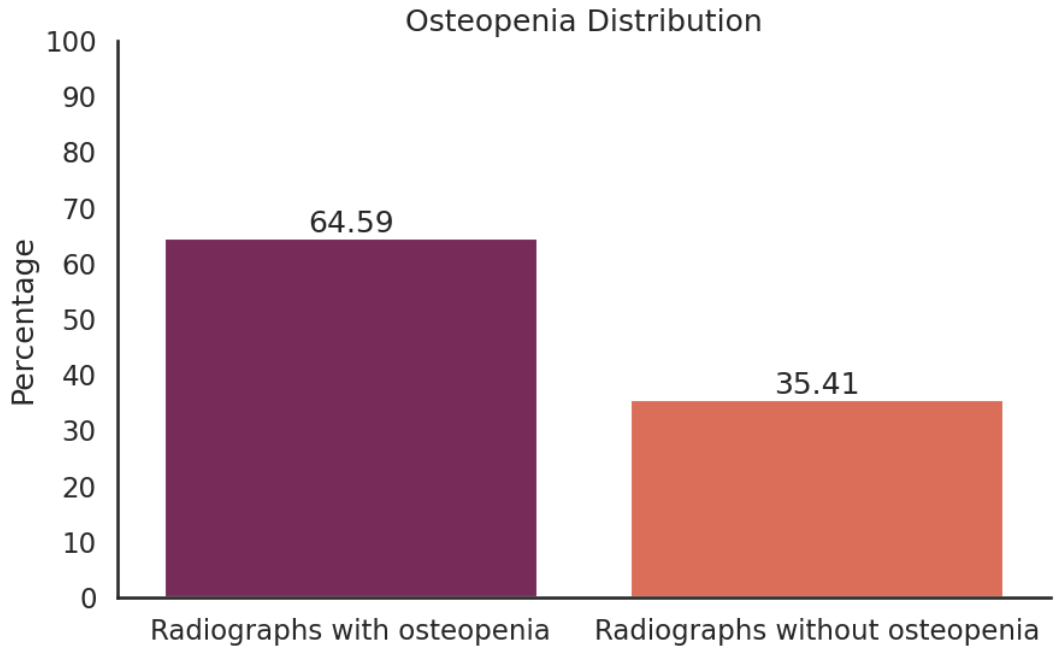

Fig. 2 Distribution of osteopenia presence in the filtered dataset

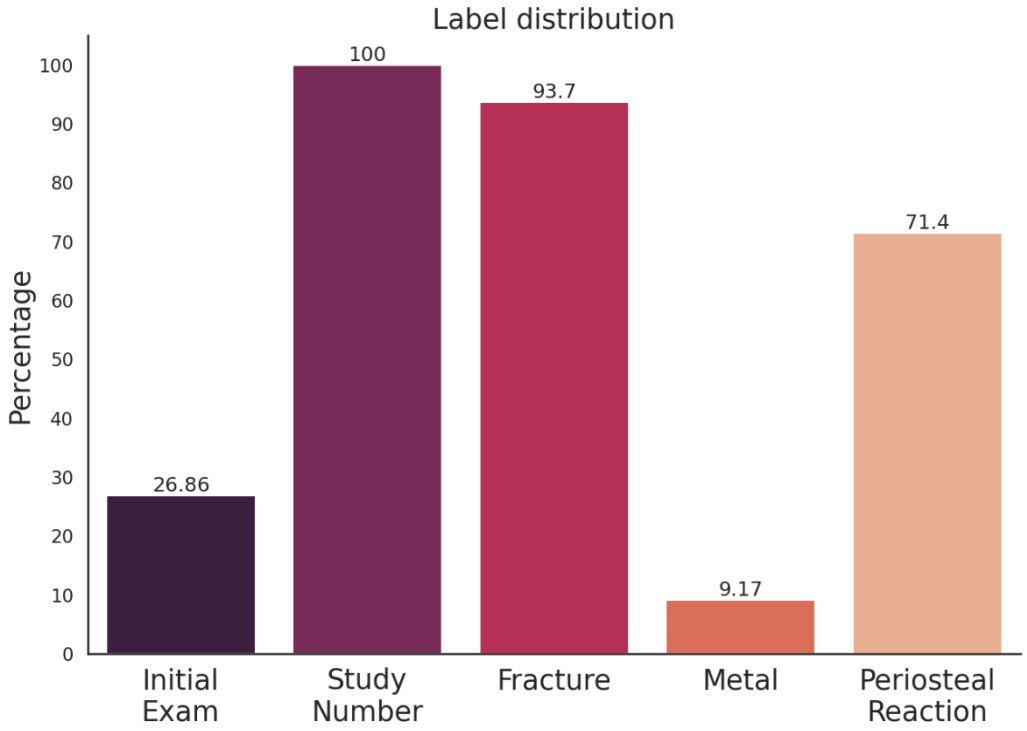

Fig. 3a Label distribution in the filtered dataset

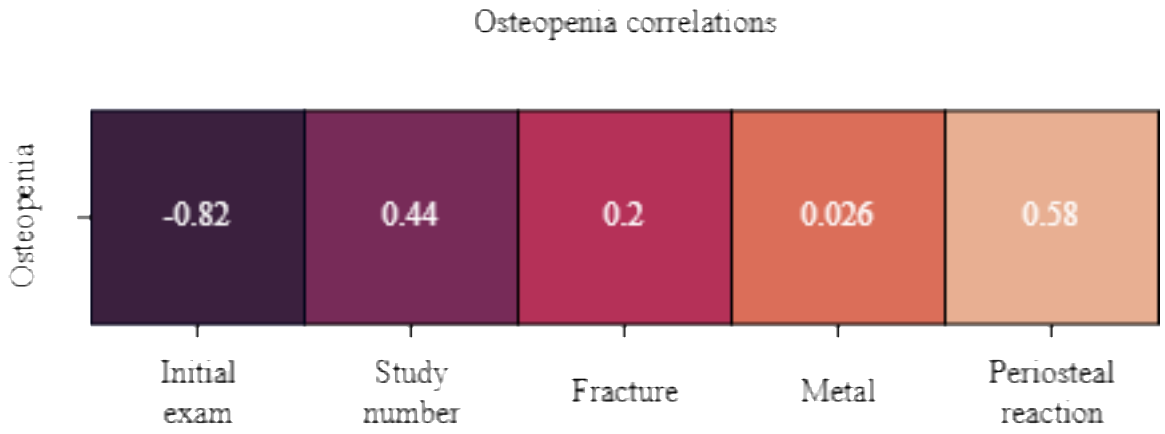

Fig. 3b The values of Spearman correlation coefficients of labels in the filtered dataset

Fig. 3 Relations between labels in the filtered dataset

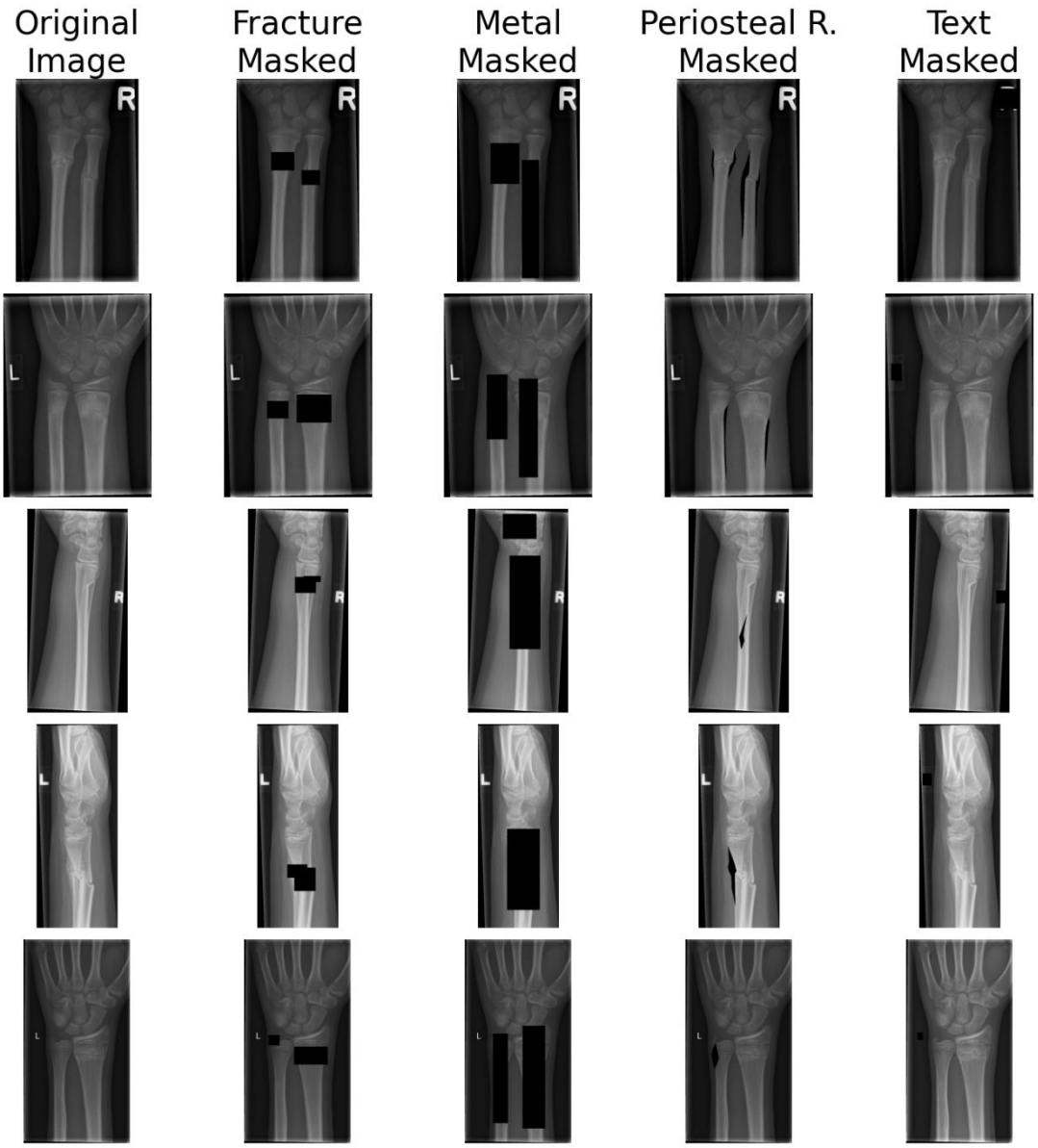

Fig. 4 A mosaic of proposed occlusion masks. Even for examples that did not have a variable present, a "dummy" mask was added nonetheless

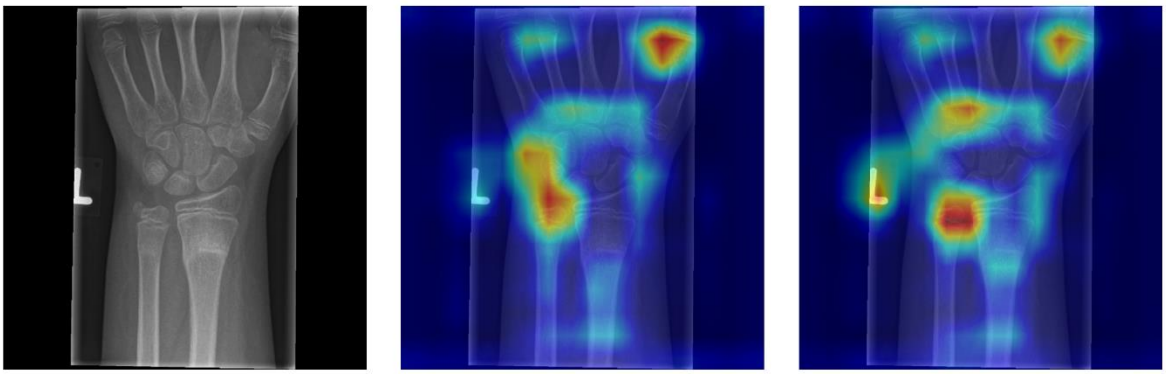

Fig. 5 An illustration of the test instance presented to the radiologists during a blind test. The X-ray image contains the letter 'L' which indicates the left hand. Upon examining the heatmaps, it is evident that one of the models exhibits greater focus on the letter 'L'

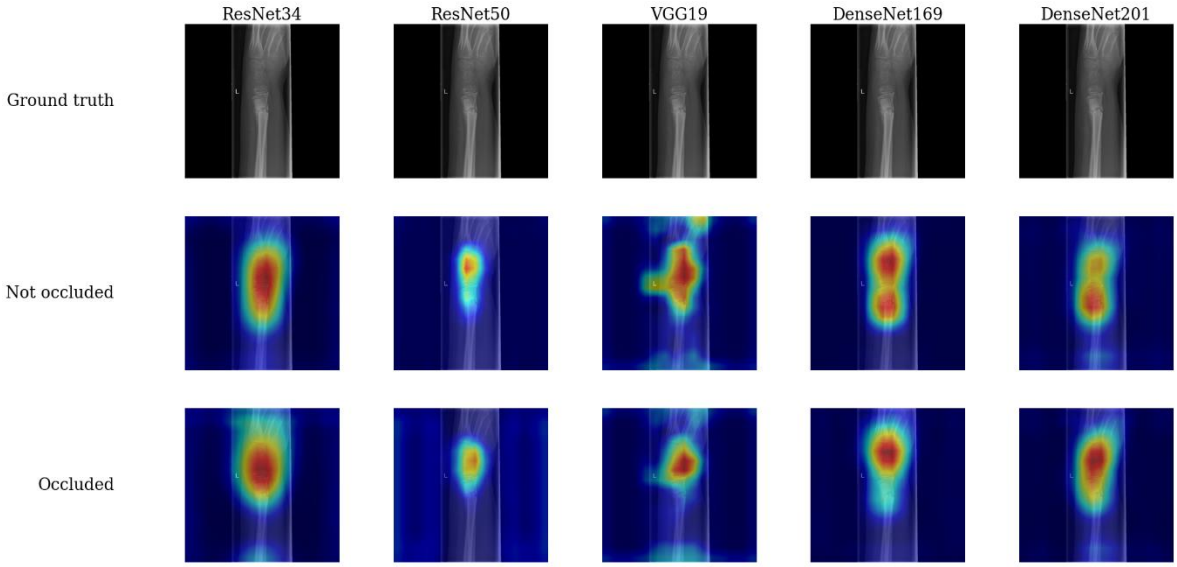

Fig. 6 GradCAM heatmaps of the best-performing instances of all trained models on the same X-ray image for each training type. Notably, the models trained with occluded confounding features exhibit a lower focus on the fractured area of the bone compared to their counterparts trained with unoccluded confounding features. Additionally, the VGG19 model displays an improved heatmap in the area of the letter 'L'

PROŠIRENI SAŽETAK

Brojne studije u medicini su istaknule izvanredne rezultate konvolucijskih neuronskih mreža u predviđanju medicinskih stanja pacijenata, a ponekad čak nadmašujući medicinske stručnjake. Unatoč njihovim izvanrednim performansama, konvolucijske neuronske mreže često se percipiraju kao algoritmi "crne kutije" (engl. *black box algorithm*), što znači da nije jasno koje značajke pridonose određenom predviđanju. Zbog principa "crne kutije" postoji pitanje dolaze li konvolucijske neuronske mreže do točnih zaključaka iz pogrešnih razloga i fokusiraju li se na pogrešna područja. Postoji mogućnost da slika sadrži varijable koje utječu na predviđanje, a nisu relevantne za to predviđanje, već se slučajno pojavljuju zajedno s relevantnim varijablama. Te varijable se nazivaju zbunjujuće varijable. Primjerice, kod smanjene gustoće kostiju česte su frakture kostiju, ali se fraktura ne mora dogoditi zbog smanjene gustoće kostiju – prema tome, ako želimo klasificirati smanjenu gustoću kostiju, fraktura predstavlja zbunjujuću varijablu.

Ovaj rad istražuje mogućnost ublažavanja utjecaja zbunjujućih varijabli unutar slika na predviđanje konvolucijskih neuronskih mreža na način da se zbunjujuće varijable identificiraju i uklone. Konkretno, fokus je na predviđanju osteopenije, ozbiljnog medicinskog stanja koje prethodi osteoporozi. Za ovaj zadatak korišten je javno dostupan GRAZPEDWRI-DX skup podataka, iz kojeg je odabran podskup rendgenskih snimki pacijenata s osteopenijom koji sadrži 3731 snimku. Nakon što su identificirane zbunjujuće varijable u korištenom skupu podataka, generirane su maske koje zaklanjaju regije slika povezane s tim varijablama. Također, označene su "lažne maske" koje se nasumično primjenjuju na ostale slike kako ne bi bile uvedene nove zbunjujuće varijable u obliku maski. Okluzijom dijelova slike maskama, modeli su prisiljeni fokusirati se na različite dijelove slike koji su relevantni za detekciju zadanog stanja.

Evaluacija modela pomoću uobičajenih metrika poput F1-mjere, preciznosti, odziva i točnosti, pokazala je da su modeli trenirani na slikama bez maski obično nadmašili modele trenirane na slikama s maskama. Međutim, nepristrani test s radiolozima pokazao je drugačije rezultate. Radiolozi su preferirali modele trenirane na slikama s maskama. U testu su radiolozi birali model na temelju regija slika na koje je model fokusiran prilikom predviđanja, pri čemu su regije dobivene GRAD-CAM metodom. Rezultati sugeriraju da uklanjanje zbunjujućih varijabli može smanjiti performanse modela prema uobičajenim metrikama, ali pruža pouzdaniji uvid u razloge koji utječu na predviđanja modela.

Ovaj rad naglašava važnost interpretabilnosti i performansi konvolucijskih neuronskih mreža iz perspektive kliničke prakse, a ne samo znanstvenih istraživanja.

KLJUČNE RIJEČI

Umjetna inteligencija, ublažavanje pristranosti, obrada slike, interpretabilno donošenje odluka, učenje s okluzijom, osteopenija