Tibial Injury Detection using Convolutional Neural Networks

Matheus Bonfim da Rocha[1], Bruno Uhlmann Marcato[1], Wally auf der Strasse[2], Maiara Mitiko Taniguchi[3], José Luis Seixas Junior[4], Daniel Prado Campos[1], Rafael Gomes Mantovani[1]

Federal University of Technology – Paraná (UTFPR), Campus of Apucarana, Paraná, Brazil
Pontifical Catholic University of Paraná (PUCPR), Curitiba, Paraná, Brazil
State University of Maringá (UEM) - Maringá, Paraná, Brazil
State University of Paraná (UNESPAR), Apucarana, Paraná, Brazil

Abstract. Bone fractures are common traumas in hospital orthopedic departments. Thermal images in an orthopedic emergency setting indicate the exact location of the traumatic injury, facilitating the acquisition of radiological images and the correct patient positioning, avoiding the acquisition of complementary images. Despite significant progress in the area, there is still a need to develop thermal image automated techniques that provide robust, accurate, and detailed classification. Most studies segment manually regions of interest and establish threshold temperature values using specific thermal image processing software. Thus, in this study, we evaluated the use and effectiveness of convolutional neural networks for tibia injury detection with thermographic images. Experiments were performed with a real dataset developed by UTFPR/UFPR universities under the ethical guidelines of Resolution 466/12, with the approval of the Research Ethics Committees of the Federal and Hospital das Clínicas of the Federal University of Paraná (UFPR). The results were promising, showing that VGG19 could accurately recognize healthy and unhealthy patients with an average F-Score of 0.894. Although not statistically accurate like VGG results, traditional ML baselines could unveil some important image features that could explain the decision process, most related to the red channel values, saturation, and image texture.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: deep learning, medical diagnosis, supervised classification, thermal images

1. INTRODUCTION

Bone fractures are common orthopedic traumas in hospitals orthopedic departments [Almigdad et al. 2022]. Depending on their severity, the diagnosis may evolve favorably or present recurrent bone infections and a prognosis of delayed bone healing and pseudarthrosis, presenting in this case as a public health problem due to the length of clinical treatment. Acute trauma and bone repair are directly related to metabolic and vascular alterations, which present thermal changes due to increased or decreased blood flow around the injury [Bixel et al. 2024]. Acute trauma thermal imaging in an orthopedic emergency setting indicate the exact location of the traumatic injury, facilitating the correct acquisition of radiological images, as well as the correct positioning of the patient, avoiding the acquisition of complementary radiographs [der Strasse et al. 2022]. However, current thermal imaging studies evaluates tibial bone trauma by manually segmenting the regions of interest and establishing threshold temperature values using specific thermal image processing software.

Despite significant progress, there is still a need to develop thermal image classification techniques that provide robust, accurate and detailed thermogram classification. They are also practical enough to be used by healthcare professionals in their clinical practices [Senalp and Ceylan 2022]. Recent advances in machine learning enable advances in computer-aided diagnosis systems, in the evaluation of skin cancer [Magalhães et al. 2021], diabetic foot [Khandakar et al. 2021], osteosarcoma bone cancer detection [Gawade et al. 2023], differential diagnosis of diabetic foot osteomyelitis and Charcot neuropathic osteoarthropathy [Cakir et al. 2024] and thyroid nodule [Etehadtavakol et al. 2025].

Thermal image analysis using automatic detection and feature extraction techniques can accelerate patient triage by enabling rapid identification of bone trauma, supporting initial diagnosis and follow-

up in orthopedic emergency settings. Thus, this paper investigates the hypothesis that Deep Learning (DL) [Aggarwal 2018], in specific, Convolutional Neural Networks (CNNs) models can be explored to perform automatic bone trauma detection of tibia injuries through thermal images. Experiments were performed with different DL architectures, from simple to state-of-the-art models.

2. RELATED WORKS

Recent studies have explored thermal imaging and DL for injury detection in sports and orthopedics. [Ergene et al. 2024] developed a pipeline using U-Net for segmentation and EfficientNet for classification, achieving up to 0.83 accuracy in detecting hamstring injuries in football players. Similarly, [Trejo-Chavez et al. 2022] proposed a methodology using CNNs to differentiate between healthy and injured knees, attaining 0.98 accuracy with various image processing techniques.

Beyond sports, [der Strasse et al. 2021] investigated the application of thermal imaging in monitoring tibia bone healing after severe trauma. This study demonstrated that thermography could detect temperature changes associated with healing processes and complications like bone infection. Regarding bone trauma, a study in pediatric patients stands out, which proposed automatic segmentation of thermal images and classification of diagnostic medical image data in wrist fractures [Shobayo et al. 2024]. The results demonstrated sensitivity and accuracy of 0.88 and 0.76 in identifying wrist fractures, respectively. A similar study evaluated rheumatoid arthritis and classified healthy and arthritis patients with 90% accuracy, with sensitivity and specificity of 0.96 and 0.85, respectively. The results of this study highlight the potential of machine learning models to identify the disease in its early stages. This approach can significantly improve clinical decision-making and patient outcomes by facilitating intervention [Ahalya and Snekhalatha 2025].

The use of a Light Weight Convolutional Neural Network (LWCNN) for classification of thermal images was investigated in [Taspinar 2023]. The author explored a set of low-dimensional transformations of the original raw data, generating different image inputs from: Histogram Oriented Gradients (HOG), Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT) and Gabor Filter (GB) methods. Experiments were performed in three different datasets, and VGG-16 was included as a baseline. The best results were obtained by LW-CNN fed with raw images, with F-Score values between 0.96 and 0.98 in all the datasets. These studies highlight the potential of thermal imaging as a non-invasive, radiation-free tool for injury detection and monitoring, complementing traditional diagnostic methods in sports medicine and orthopedics.

3. METHODOLOGY

An overview of the experiments flow, including sub-steps, is shown in Figure 1. The following subsections give additional details regarding them: the image dataset acquisition and preprocessing, learning algorithms and their model evaluation process, focusing on the reproducibility of the experiments.

3.1 Data Acquisition

The diagnostic tibia images were acquired using a professional thermal imaging camera[1], featuring a thermal sensitivity/NETD < 30 mK at 30°C, a 42° lens, focal plane sensor resolution of 320×240, and a minimum focal distance of 0.15 m. The camera emissivity was set to 0.98, as recommended by the manufacturer. Image acquisition was carried out at the Trauma and Bone Reconstruction Service of the Hospital Universitário Federal do Paraná (HUFPR) in Curitiba, Brazil, between June 2020 and June 2021. Images were captured in a controlled environment, with closed windows and minimal external interference.

[1]FLIR model T530, Professional Scientific, FLIR® Systems Inc., Wilsonville, Oregon, USA

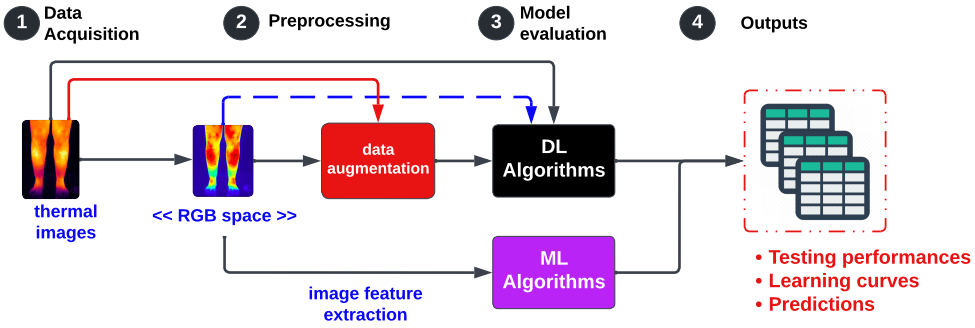
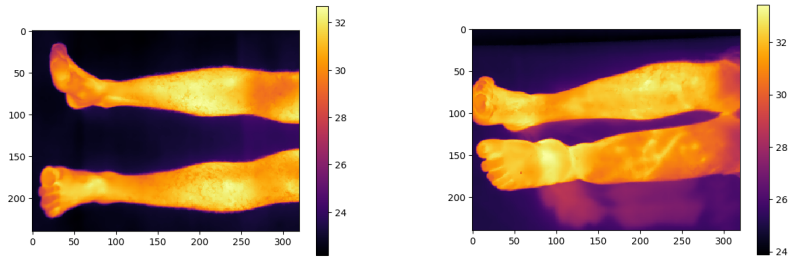


Fig. 1: Experimental methodology for automated diagnosis of tibia injuries.

The study followed the ethical guidelines of Resolution 466/12 and was approved by the Research Ethics Committee of the Federal Technological University of Paraná (UTFPR), under protocol number 3.014.748 (November 12, 2018), and by the Research Ethics Committee of the Hospital das Clínicas of the Federal University of Paraná (UFPR), under protocol number 3.067.005 (December 8, 2018). The sample consisted of patients with confirmed medical diagnosis of tibial bone injury, with no associated orthopedic trauma, as well as healthy volunteers who composed the control dataset. The resultant dataset comprises a total of 731 images.

3.2 Data Preprocessing

The raw thermal images were obtained directly from the FLIR T530 camera, which captures single-channel temperature matrices representing radiometric thermal values in each pixel. These raw data, stored as grayscale images, reflect surface temperature intensity in a 2D spatial distribution and do not include color information. The raw images were later converted into RGB pseudo-color representations to enable compatibility with DL architectures that require multi-channel input. They were generated using OpenCV Python library. Thus, experiments were performed with both raw and pseudo-color RGB images. All the images had their pixel values normalized from $[0, 255]$ to the interval $[0, 1]$.



(a) Healthy diagnosis. (b) Unhealthy diagnosis.

Fig. 2: Example of dataset image instances and their corresponding labels.

We anonymized sensitive data from patients and binarized the target feature to map healthy and unhealthy cases, due to the lack of representative samples for some types of diagnosis. It generated a dataset with 560 healthy samples and 171 with tibia injuries, leading to an imbalance rate of 0.78. Figure 2 shows an example for each class for a random patient. All the images in the dataset show only thermal recordings of the patient's lower body, focusing on the tibia. Data Augmentation (DA) was applied to increase the minority class of the dataset (positive diagnosis). Three transformations were applied: horizontal and vertical flips and a zoom-in transformation. None of these transformations

generates image types that are invalid to those originally contained in the dataset. DA expanded images for training, while the testing sets remained imbalanced.

3.3 Image Feature Extraction

We also evaluated traditional ML algorithms in experiments as baselines for DL architectures. To feed these classical algorithms, we extracted some features from RGB images and represented them as descriptors vectors. This process was carried out using the `Image Meta-feature Extractor` library developed by [Aguiar et al. 2019]. This tool generated a total of 97 features organized into six different categories: 3 simple image statistics; 36 color-based measures (RGB, HSV); 21 statistics from histograms of colors and intensity; 16 border descriptors; 2 quality assessment metrics; and 19 texture measures from FFT and LBT methods[2]. The resultant dataset had 731 samples and 97 features. A tabular preprocessing step was also conducted, removing i) constant and ii) highly correlated features with an absolute Pearson correlation value ≥ 0.95. The final dataset reduced the number of features to 65.

3.4 Deep Learning Algorithms

Three different DL architectures were considered in experiments: i) classical and straightforward Convolution Neural Network (CNN) architecture [Ribeiro et al. 2024]; ii) a state-of-the-art VGG19 architecture [Simonyan and Zisserman 2015]; and iii) a Light-Weight CNN (LWCNN) explored by similar studies with medical thermal images [Taspinar 2023]. The CNN baseline would be the simplest architecture designed for image recognition (lower baseline). At the same time, the VGG19 was the top-ranked model mentioned in several literature-related studies (higher baseline). The LWCNN would be a cheaper alternative, with fewer parameters and model size. The DL architectures used in experiments are the same as reported by the original studies cited above[3]. It is important to mention that the VGG19 models were pre-trained in the *ImageNet* dataset, and only the dense layers' weights placed before the output layer were trainable in our experiments (the base model was not).

3.5 ML Algorithms

Alternatively, a total of six traditional ML algorithms [Marsland 2015] were evaluated as baselines in the experiments: k-Nearest Neighbors (kNN), Decision Trees (DTs), Bagging of DTs, Random Forest (RF), Support Vector Machines (SVMs) with RBF kernel and a Linear classifier (Ridge). Each algorithm has a different inductive bias, resulting in different mappings between image characteristics and corresponding classes. All of them were coded with the `scikit-learn` library and their corresponding default hyperparameter values.

3.6 Experimental Setup

We evaluated induced models using a repeated stratified holdout resampling: 70% of the data for training and 30% for testing. Due to the stochasticity of the algorithms, experiments were repeated 10 times with different seeds. DL models minimize the binary cross-entropy loss in training via the Adam optimizer and its default learning rate. In a single execution, a total of 30% of the training data is used as the validation data. We empirically defined batch size = 16 and the maximum number of epochs = 100. Additionally, two callbacks were defined to avoid overfitting during training. There is an early stopping if no improvement is observed in the validation loss for 10 successive epochs, while a model checkpoint saves and restores the best weights found during learning. The F-Score evaluation measure assessed ML and DL induced models. To ascertain the statistical significance of our findings,

[2]More details can be found at `https://github.com/gabrieljaguiar/image-meta-feature-extractor`
[3]DL details can be found at: `https://github.com/ic2d/tibiaInjuryDetection/blob/master/src/deepModels.py`

we evaluated the results using non-parametric Wilcoxon with a significance level $\alpha = 0.05$. Most of the code was developed in Python, with the automatic analysis coded in R. Table I details the complete experimental setup necessary to replicate the current study. The code repository of this study is also publicly available[4].

Table I: Complete experimental setup.

Element	**Option**	**R/Python package**
Resampling	Stratified holdout Training set = 70% Testing set = 30%	`Python:scikit-learn`
Image descriptors	97 features [Aguiar et al. 2019]	`Python: image-meta-feature-extractor`
ML algorithms	kNN, DT, RF, Bagging, SVMs, Ridge	`Python: scikit-learn`
DL algorithms	CNN, VGG19, LW-CNN	`Python: Keras`
DL setup	epochs = 100 validation split = 0.3 batch size = 16 optimizer = Adam optimized measure = Binary cross entropy early stop criteria = 10 epochs (val loss)	`Python: Keras`
Data augmentation	Horizontal and vertical flip	`Python: albumentations`
Evaluation measure	F-Score	`Python: scikit-learn, Keras`
Repetitions	10 times with different seeds	-
Statistical evaluation	Wilcoxon - $\alpha = 0.05$ (95%)	`R: stats`
Automatic Graphical Analysis		`R: ggplot2, dplyr`

4. RESULTS

Figure 3 depicts the overall results of all the induced models. In the figure, violins highlight induced models' F-score distributions. There is also a boxplot contained in each violin showing median values and their quartiles. Induced models are decreasingly ordered in the x-axis from left to right according to their median F-Score values. The red dotted line at 0.8 in the a-xis separates models into the most accurate and regular ones. The best results were obtained by the VGG19 architecture using RGB images and no data augmentation, keeping the original imbalance in training. It obtained a mean F-Score of 0.894 with a standard deviation of 0.038. VGG19 is a well-known state-of-the-art architecture that took advantage of the number of layers and their weights pre-trained in ImageNet. The top-3 models are completed with Ridge and Random Forest classifiers, traditional ML algorithms induced with features extracted from the RGB images. They differ minimally in terms of mean F-Score (0.827, 0.814, respectively) with standard deviations $\in [0.033, 0.044]$.

All the induced models presented average F-Score values higher than 0.73. We may also note from the figure that traditional ML algorithms were competitive with DL architectures, except for kNN and DTs. The use of DA improved the simplest CNN architectures but significantly worsened VGG19. These models were better with raw data than using RBG images. On the other hand, image features extracted from RGB images were diverse to provide accurate ML models.

4.1 Top-3 Induced Models

While considering only the top-3 induced models which were VGG19, Ridge and RF, we assessed the statistical significance of their results. The non-parametric Wilcoxon paired-test with $\alpha = 0.05$ (95% significance) was applied to compare their F-Score distributions. For VGG19 $\times$ Ridge and VGG $\times$ RF, p-values < 0.001 were obtained, meaning a statistical difference favoring VGG19. Notwithstanding,

[4] `https://github.com/ic2d/tibiaInjuryDetection`

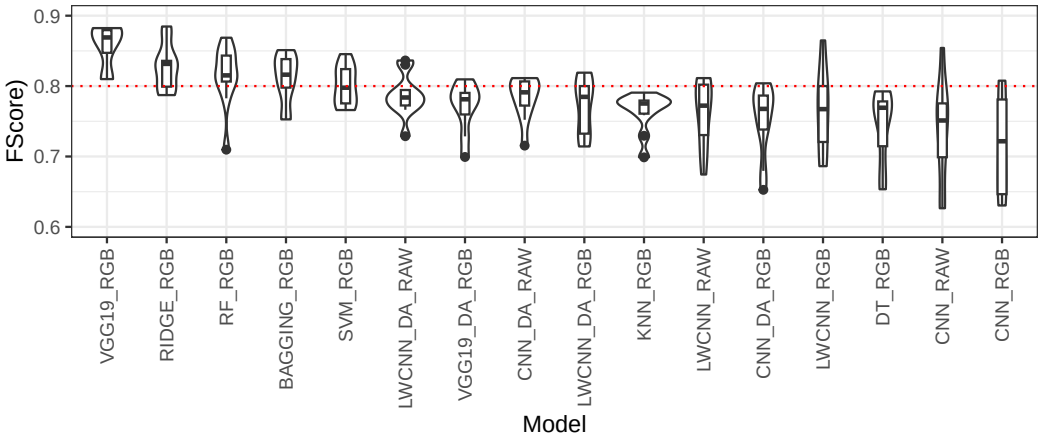


Fig. 3: Results obtained by DL and ML algorithms with thermal images.

test with Ridge × RF (p-value of 0.625), the null hypothesis is considered, meaning no statistical difference between the distributions.

Figure 4 depicts the confusion matrices obtained by the top-3 models in the testing sets. Values are rounded averaged considering the ten different seeds. In the figure, the zero (0) label denotes healthy images while one (1) indicates images with a disease diagnosis. One may note that all induced models accurately recognize the majority class (healthy images). The performance difference occurs when predicting the diagnosis: VGG better identifies the characteristics of these images through its abstract representations.

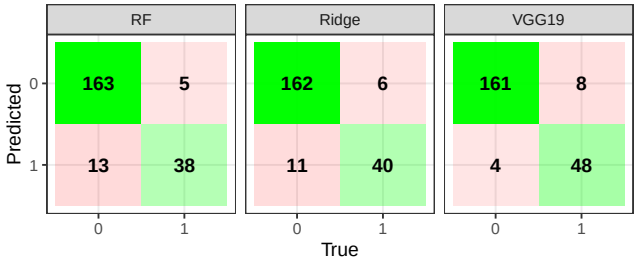
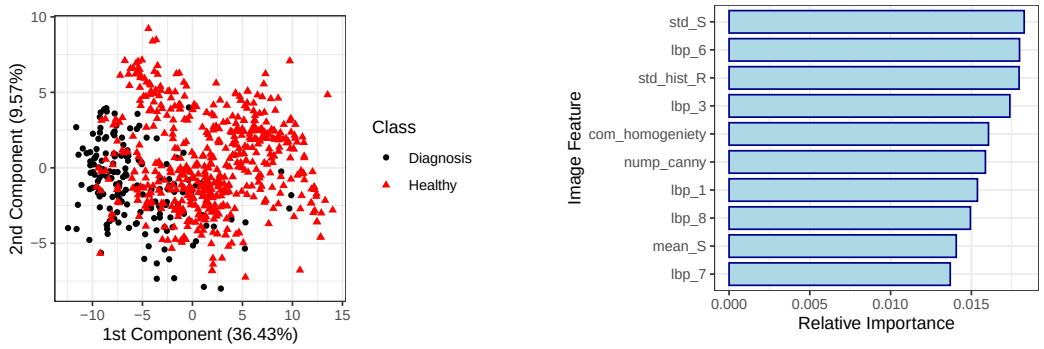


Fig. 4: Confusion matrices obtained by top-3 models in the testing sets.

Even statistically worse than VGG, ML algorithms presented quite interesting predictions. Figure 5 presents some meta-characteristics directly or indirectly extracted from induced models. Since Ridge is a linear classifier, we looked at the linearity of the generated image dataset. Sub-figure 5a presents a 2D-Principal Component Analysis (PCA) projection considering its first two components. Together, they describe 46% of the data variance, while adding the third component would increases this value to 55%. Most of the healthy and diagnostic images seem to be linearly separable, but there is an overlap region that may lead to misclassifications: these images might have similar image features and do not differ sufficiently.

Sub-figure 5b lists the top-10 most important features from RF models. The y-axis lists the image features while the x-axis projects their relative importance in terms of the Gini index. Among them: 6 are texture features (`lbp_6`, `lbp_3`, `com_homogeneity`, `lbp_1`, `lbp_8`, `lbp_7`); 2 color-based features (`std_S`, `mean_S`); one border feature (`nump_canny`); and one histogram feature (`std_hist_R`). Lpb_* are Local Binary Pattern (LBP) features that compare the intensity of a central pixel in a small



(a) 2D PCA dataset projections.

(b) Relative importance of the image features obtained by induced Random Forests.

Fig. 5: Meta-characteristics extracted from traditional ML induced models.

neighborhood with its surrounding pixels. The `std_hist_R` measures the standard deviation of the R channel histogram, while the S features determine the mean and standard deviation of the Saturation channel from HSV space. They measure the intensity or purity of the color.

Increased temperature regions are associated with acute inflammation or infection, while lower temperatures are linked to reduced perfusion, especially in chronic conditions such as osteomyelitis [der Strasse et al. 2021; der Strasse et al. 2022]. Therefore, the prominence of features derived from saturation values and the red channel values is aligned with the thermographic response to alteration of local blood flow. Texture-based features, also ranked highly, likely capture local irregularities in thermal distribution. Such patterns may reflect microvascular anomalies or localized metabolic activity, consistent with physiological responses described in the literature [Reed et al. 2020].

5. CONCLUSION

In this article, ML and DL algorithms were investigated for tibial injury detection on thermal images. Experiments were carried out with a real thermal dataset containing 731 images of healthy and tibial injured patients. The best model was VGG19, reaching 0.894 average F-Score, which accurately classified both patterns. Also, the model was statistically better than all other setups tried in experiments. However, it requires higher computation cost to train even using pre-trained ImageNet weights. VGG19 required 4 hours to be trained in a single seed using CPUs, while traditional ML algorithms run in few minutes[5]. Alternatively, traditional ML algorithms were assessed using image features as data descriptors. Using their default hyperparameter values, Ridge, RF, Bagging and SVM obtained F-Score values $\in [0.80, 0.828]$.

We could observe some data characteristics that may explain the classification process by looking into the induced models. Since Ridge and RF compose the top-3 induced models, they can explain some characteristics of the original problem. The decision boundary has a linearity degree, partially explaining Ridge success, but with a region containing overlapping instances (images) with similar features. When completing the analysis with the most important RF features, we observed that texture (LBP) and color features (Saturation and red values) differentiate healthy and unhealthy images. So, both strategies (ML and DL) provided accurate results for the tibia diagnosis problem.

For future works, we plan to understanding why VGG outperformed the other models, so we can explore eXplainable Artificial Intelligence (XAI) methods to unveil VGG predictions. We can also: explore different DL architectures (with and without pre-training weights); explore transfer learning from other thermal image problems; perform Neural Architecture Search (NAS); evaluate different ML

[5]When using GPUs provided by Google Colab, VGG training time was reduced to 10 minutes

algorithms tuning their hyperparameters, and perform data balancing before their training. Different sets of features can also be explored, such as low-size transformed images [Taspinar 2023] since DL models were better with the one-channel original data (except for VGG19).

REFERENCES

Aggarwal, C. C. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, Cham, 2018.

Aguiar, G. J., Mantovani, R. G., Mastelini, S. M., de Carvalho, A. C., Campos, G. F., and Junior, S. B. A meta-learning approach for selecting image segmentation algorithm. *Pattern Recognition Letters* vol. 128, pp. 480–487, 2019.

Ahalya, R. K. and Snekhalatha, U. Cnn transformer for the automated detection of rheumatoid arthritis in hand thermal images. In *Artificial Intelligence over Infrared Images for Medical Applications*. Lecture Notes in Computer Science, vol. 15279. Springer, Cham, pp. 29–40, 2025.

Almigdad, A., Mustafa, A., Alazaydeh, S., Alshawish, M., Bani Mustafa, M., and Alfukaha, H. Bone fracture patterns and distributions according to trauma energy. *Advances in Orthopedics* 2022 (1): 8695916, 2022.

Bixel, M., Sivaraj, K., Timmen, M., Mohanakrishnan, V., Aravamudhan, A., Adams, S., and Adams, R. Angiogenesis is uncoupled from osteogenesis during calvarial bone regeneration. *Nature Communications* 15 (1): 4575, 2024.

Cakir, M., Tulum, G., Cuce, F., Yilmaz, K., Aralasmak, A., Isik, M., and Canbolat, H. Differential diagnosis of diabetic foot osteomyelitis and charcot neuropathic osteoarthropathy with deep learning methods. *Journal of Imaging Informatics in Medicine* 37 (5): 2454–2465, 2024.

der Strasse, W. A., Campos, D. P., Mendonça, C. J. A., et al. Detecting bone lesions in the emergency room with medical infrared thermography. *BioMedical Engineering OnLine* 21 (1): 35, 2022.

der Strasse, W. A., Campos, D. P., Mendonça, C. J. A., Soni, J. F., Mendes, J., and Nohama, P. Evaluation of tibia bone healing by infrared thermography: A case study. *Journal of Multidisciplinary Healthcare* vol. 14, pp. 3161–3175, 2021.

der Strasse, W. A., Campos, D. P., Mendonça, C. J. A., Soni, J. F., Tuon, F., Mendes, J., and Nohama, P. Evaluating physiological progression of chronic tibial osteomyelitis using infrared thermography. *Research on Biomedical Engineering*, 2022.

Ergene, M. C., Bayrak, A., and and, M. C. A new deep learning based end-to-end pipeline for hamstring injury detection in thermal images of professional football player. *Quantitative InfraRed Thermography Journal* 0 (0): 1–18, 2024.

Etehadtavakol, M., Sirati-Amsheh, M., Moallem, G., and Ng, E. Enhancing thyroid nodule classification: A comprehensive analysis of feature selection in thermography. *Infrared Physics & Technology*, 2025.

Gawade, S., Bhansali, A., Patil, K., and Shaikh, D. Application of the convolutional neural networks and supervised deep-learning methods for osteosarcoma bone cancer detection. *Healthcare Analytics* vol. 3, pp. 100153, 2023.

Khandakar, A., Chowdhury, M., Reaz, M., Ali, S., Hasan, M., Kiranyaz, S., and Malik, R. A machine learning model for early detection of diabetic foot using thermogram images. *Computers in Biology and Medicine* vol. 137, pp. 104838, 2021.

Magalhães, C., Tavares, J., Mendes, J., and Vardasca, R. Comparison of machine learning strategies for infrared thermography of skin cancer. *Biomedical Signal Processing and Control* vol. 69, pp. 102872, 2021.

Marsland, S. *Machine learning: an algorithmic perspective*. CRC press, 2015.

Reed, C., Saatchi, R., Burke, D., and Ramlakhan, S. Infrared thermal imaging as a screening tool for paediatric wrist fractures. *Medical & Biological Engineering & Computing* vol. 58, pp. 1549–1563, 2020.

Ribeiro, J., Mileski, M., Seixas Junior, J. L., Carvalho, L. F., and Mantovani, R. G. Image classification for precision agriculture: A coffee study case. In *Anais do Computer on the Beach 2024*, 2024.

Senalp, F. and Ceylan, M. Effects of the deep learning-based super-resolution method on thermal image classification applications. *Multimedia Tools and Applications* 81 (7): 9313–9330, 2022.

Shobayo, O., Saatchi, R., and Ramlakhan, S. Convolutional neural network to classify infrared thermal images of fractured wrists in pediatrics. *Healthcare* 12 (10): 994, 2024.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. CoRR, San Diego, CA, USA, 2015.

Taspinar, Y. S. Light weight convolutional neural network and low-dimensional images transformation approach for classification of thermal images. *Case Studies in Thermal Engineering* vol. 41, pp. 102670, 2023.

Trejo-Chavez, O., Amezquita-Sanchez, J. P., Huerta-Rosales, J. R., et al. Automatic knee injury identification through thermal image processing and convolutional neural networks. *Electronics* 11 (23), 2022.