Performance Analysis of Feature-Based Transfer Learning Using VGG19 for Detecting Hairline and Spiral Fractures

Jiahua Xia[1], Anwar P.P. Abdul Majeed[2], Qifan Li[3], Yifeng Xu[1], Yi Chen[1], Xiaoyan Liu[1], Yang Luo[1,*] and Rabiu Muazu Musa[4]

[1]School of Intelligent Manufacturing Ecosystem, Xi'an Jiaotong-Liverpool University, Taicang, Suzhou, 215400, China
[2]School of Engineering and Technology, Sunway University, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia
[3]Beijing Anzhen Hospital, Capital Medical University, Beijing Institute of Heart, Lung, and Blood Vessel Diseases, 100029, Beijing
[4]Universiti Malaysia Terengganu, Persiaran Ktgr 21030 Kuala Terengganu, Malaysia

*yang.luo@xjtlu.edu.cn

Abstract. Correct classification of bone fractures is crucial for accurate medical treatment, as traditional diagnostic procedures can be time-consuming and prone to errors. This study explores the use of the VGG19 architecture combined with three classifiers, namely Support Vector Machine (SVM), Logistic Regression (LR), and k-Nearest Neighbors (kNN) for classifying hairline and spiral fractures in X-ray images. A dataset of 400 X-ray images was utilized, from which features were extracted using the VGG19 model. These features were then used to train the SVM, LR, and kNN classifiers. Among the models tested, the VGG19-LR pipeline demonstrated the best overall performance, achieving high accuracy and robustness in both validation and testing phases. The VGG19-SVM model also showed strong performance but was slightly less effective than the VGG19-LR. In contrast, the VGG19-kNN model yielded the weakest results, indicating lower suitability for this classification task. These results suggest that the VGG19-LR pipeline was suitable for the given dataset. The high accuracy demonstrated indicates that transfer learning can offer an efficient method for classifying bone fractures, particularly for hairline and spiral fracture, in contrast to training a deep neural network from the beginning. This enables the creation of automated bone fracture diagnosis using computer vision and deep learning.

Keywords: Bone fracture classification, VGG19Convolutional Neural Networks (CNNs), Support Vector Machine (SVM), Logistic Regression (LR), k-Nearest Neighbors (kNN), Medical imaging.

1 Introduction

Medical imaging serves as a fundamental pillar in contemporary healthcare, particularly for diagnosing and treating fractures. Traditional diagnostic methods, which often rely on the subjective interpretation of X-ray images by experts, can be time-consuming and may exhibit variability in accuracy due to the inherent subjectivity of human assessment. However, with the advent of digital imaging and artificial intelligence, there has been a significant shift towards automated fracture detection and classification using deep learning techniques, which have substantially enhanced the precision of identifying fracture types.

Advancements in deep learning, especially CNN, have profoundly influenced the field of medical image analysis. As noted by Krizhevsky et al., [1] CNNs pre-trained on extensive datasets such as ImageNet have demonstrated impressive accuracy in classifying images, setting a new benchmark for automated image analysis tasks.

Building upon this foundation, Yadav et al. [2] developed a deep neural network model for the classification of fractured and healthy bones. The challenge of overfitting on a small dataset was addressed by employing data augmentation techniques, effectively expanding the dataset size. The model's performance was evaluated through three sets of experiments utilizing softmax activation and the Adam optimizer. The results were promising, with the model achieving a classification accuracy of 92.44% for both healthy and fractured bones, based on 5-fold cross-validation. Moreover, the model demonstrated high accuracy rates of over 95% and 93% on 10% and 20% of the test data, respectively.

While these findings are encouraging, Ronneberger et al. [3] point out the necessity for a sophisticated approach when applying these models to specific tasks such as biomedical image segmentation. The characteristics of the dataset and the model architecture must be carefully considered to leverage the full potential of deep learning in this domain. Although previous studies [4-6] have explored the use of CNNs for bone fracture detection and have shown the technology's promise, there remains an opportunity for refining the models to improve accuracy and ensure consistent performance across various fracture types.

In this context, the current study aims to develop a sophisticated deep learning model capable of accurately categorizing bone fractures into specific types such as hairline, spiral, and non-fractured. Barhoom et al.'s [7] approach utilizes transfer learning with the VGG19 architecture, a pre-trained CNN model renowned for its effectiveness in image recognition, to extract features from medical images. The goal is to develop a model that not only identifies but also precisely represents the nuances of different fracture types. By harnessing the powerful feature-extraction capabilities of VGG19, the study seeks to enhance the precision and reliability of fracture detection in medical imaging.

To further validate the model's performance, this article conducts an analysis of three classifiers, SVM, LR, and KNN on the task of classifying bone fractures using features extracted by the VGG19 model. The study demonstrates the effectiveness of integrating deep learning features with traditional machine learning classifiers. A comprehensive evaluation will be performed using standard metrics such as accuracy, precision, recall,

and F1-score, along with a detailed examination of confusion matrices. This thorough assessment aims to determine the most effective classifier for the task of bone fracture classification, thereby contributing to the ongoing advancement of medical image analysis through deep learning applications.

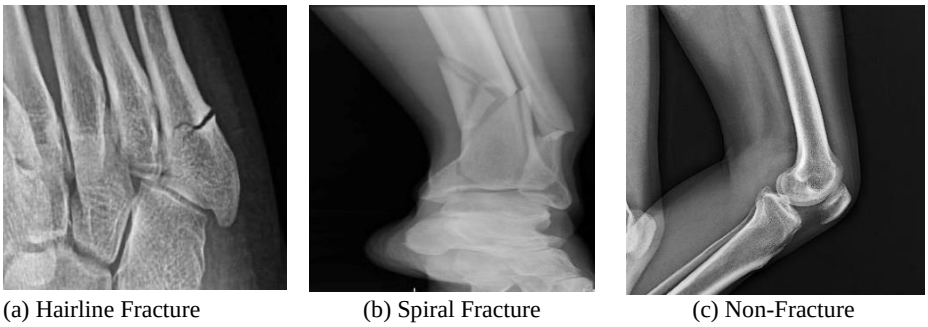
2 Methods

This study utilized X-ray images of bone fractures, obtained from openly accessible medical imaging libraries [8], initially consisted of 1,000 images. For this study, the sample size was reduced to 400 images, and each image was downsized to dimensions of 224 by 224 pixels. This resizing ensures compatibility with pre-trained models and facilitates consistent feature extraction. The dataset is classified into three categories: hairline fracture, spiral fracture, and non-fracture. Fig.1 displays sample photos representing these three groups. The dataset was partitioned using a 60:20:20 hold-out cross-validation procedure for training, validation, and testing, respectively.

The choice of a pre-trained CNN is critical for the task of bone fracture classification. In this study, the VGG19 architecture was selected due to its effectiveness in various image recognition tasks, owing to its deep and hierarchical structure. The VGG19 model, pre-trained on the ImageNet dataset, offers a robust feature extractor, which is leveraged through transfer learning for the specific task at hand. The decision to use VGG19 is based on its ability to capture complex patterns and textures in images, which is essential for distinguishing between different types of bone fractures [9].

Feature extraction is performed using the VGG19 model. The process involves feeding the pre-processed medical images through the VGG19 network and extracting the high-level features from the last convolutional layer before the fully connected layers. These features are then flattened and used as input for the classifiers. The rationale behind this approach is that the deep features learned by VGG19 can effectively capture the nuances of bone fractures, which are essential for accurate classification [10].

Three different classifiers are used in this study to evaluate their performance in classifying bone fractures based on the features extracted from the VGG19 model. These classifiers include:



(a) Hairline Fracture (b) Spiral Fracture (c) Non-Fracture

Fig.1. Sample x-ray image of Hairline Fracture, Spiral Fracture, and Non-Fracture.

1. **SVM:** Known for its effectiveness in high-dimensional spaces, SVM works by finding the optimal hyperplane that separates different classes. The choice of kernel (e.g., linear, RBF) and parameters (e.g., C, gamma) is crucial for the model's performance.
2. **LR:** A simple yet powerful linear model for binary classification tasks, which can be extended to multi-class problems using techniques like one-vs-rest. LR is chosen for its interpretability and efficiency.
3. **kNN**: A non-parametric method that classifies a sample based on the majority vote of its nearest neighbors. The number of neighbors (k) and the distance metric are key parameters that significantly influence the classifier's accuracy.

Popular machine learning libraries, such as scikit-learn in Python, were used for the implementation of classifiers and evaluation metrics. The classifiers are evaluated using standard metrics: accuracy, precision, recall, and F1-score. These measures collectively provide insight into the performance of the models regarding their ability to classify bone fractures correctly. Confusion matrices also allow analysis of the types of misclassifications for each set of models, as well as any patterns that might arise from the distribution of different fracture types [11].

3 Results and Discussion

The bar chart in Fig. 2 displays the classification accuracy of three distinct VGG19 architectures (VGG19-SVM, VGG19-LR, and VGG19-kNN) throughout training, validation, and test datasets.

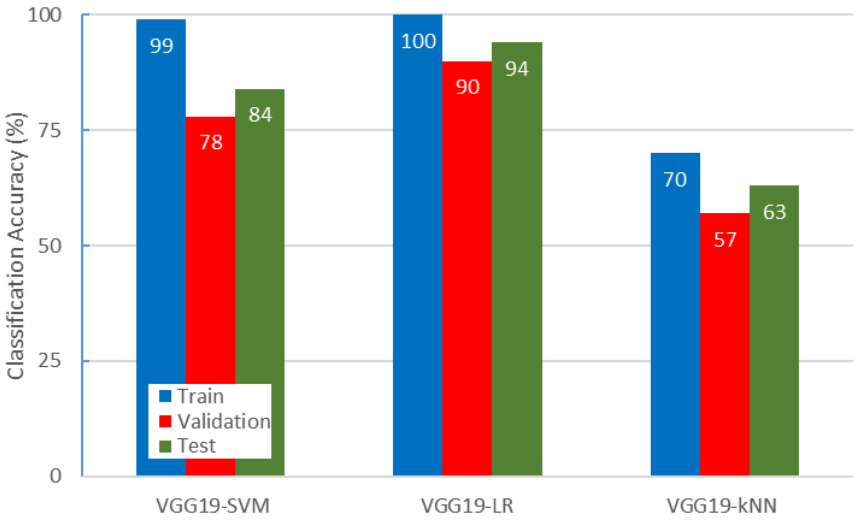


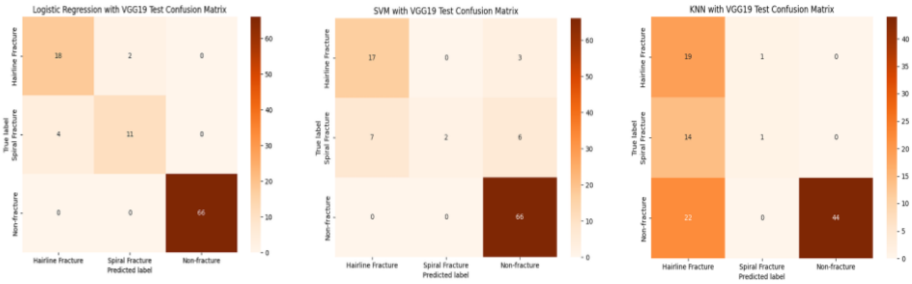
Fig. 2. Evaluated pipelines performance on detecting the bone fracture types

The VGG19-SVM pipeline exhibited an impressive training accuracy of 99%, demonstrating a strong model fit on the training data. However, the validation accuracy decreased to 78%, indicating a noticeable case of overfitting, as the pipeline did not maintain the same level of performance on the validation set. The presence of overfitting is further supported by the fall in test accuracy to 84%, indicating a consistent but inferior performance on independent data.

Comparatively, the VGG19-LR pipeline achieved a training accuracy of 100%, indicating that the model perfectly learned the training data. The validation accuracy was recorded at 90%, suggesting a good balance between fitting the training data and generalizing to unseen data. The test accuracy was slightly lower at 94%, showing reliable performance on new data, and indicating that the VGG19-LR model effectively minimized overfitting compared to VGG19-SVM.

The VGG19-kNN pipeline achieved a training accuracy of 70%, significantly lower than the other two models. This suggests that the kNN model struggled to fully capture the training data's patterns. The validation accuracy dropped to 57%, and the test accuracy was 63%, both indicating a lack of generalization and a significant disparity in performance across different datasets. The results highlight that the VGG19-kNN pipeline may not be as effective as VGG19-SVM and VGG19-LR in classifying bone fractures.

The confusion matrices for the different pipelines provide a comparative evaluation of their performance on the test set for hairline fractures, spiral fractures, and non-fractures, as shown in Fig.3. The combination of VGG19 and SVM demonstrated good performance with minimal misclassification in non-fractures but exhibited some confusion between hairline and spiral fractures. The higher number of false positives for spiral fractures suggests a need for better discrimination between these categories. The VGG19-LR model showed superior performance, particularly in accurately recognizing hairline fractures and non-fractures, with minimal misclassification. The slight confusion between hairline and spiral fractures indicates that while the model is highly effective, there is room for improvement in distinguishing these two types of fractures.



VGG19-LR VGG19-SVM VGG19-kNN

Fig. 3. Confusion Matrix of the VGG pipelines with LR, SVM and KNN on the test dataset.

The VGG19-kNN combination exhibited the highest level of misclassification, particularly confusing spiral fractures with hairline fractures and non-fractures. The high number of false positives in both categories indicates that the kNN model struggled significantly with this classification task, likely due to its reliance on the local structure of the data.

4 Conclusions

This study investigated the efficacy of the VGG19 architecture combined with various classifiers, such as SVM, LR and kNN, for the classification of hairline and spiral bone fractures from X-ray images. The comparative analysis of these models yielded insightful findings regarding their performance and applicability in medical imaging.

The VGG19-SVM model demonstrated high training accuracy, suggesting a robust model fit, yet it exhibited notable overfitting, as evidenced by the lower validation and test accuracies. This indicates challenges in generalizing to new data, particularly in distinguishing between hairline and spiral fractures. In contrast, the VGG19-LR model achieved perfect training accuracy and maintained strong validation and test performances, demonstrating its ability to balance training and generalization effectively. Its superior performance in differentiating between fracture types, especially hairline fractures and non-fractures, underscores its potential for reliable application in clinical diagnostics. The VGG19-kNN model, however, showed the least satisfactory results, with significant drops in accuracy across datasets. Its performance highlighted difficulties in capturing the intricate patterns necessary for accurate fracture classification, leading to considerable misclassification.

The findings of this study highlight the critical role of advanced feature-based transfer learning techniques, with VGG19 proving particularly effective in enhancing the precision and reliability of fracture diagnosis. Among the evaluated models, VGG19-LR emerged as the most promising for practical use in automated diagnostic tools within medical practice.

Future research should focus on validating these findings with larger, real-world clinical datasets to further establish the robustness and applicability of these methods. This will ensure their readiness for integration into clinical workflows, ultimately improving diagnostic accuracy and patient outcomes.

Funding: This research was funded by Research Development Fund, Grant Number: RDF-21-01-028; Summer Undergraduate Research Fellowship, Grant Number: SURF-2024-0355; and Project for Centre of Excellence for Syntegrative Education, Grant Number: COESE2324-01-07 of Xi'an Jiaotong-Liverpool University.

References

1. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.

2. Yadav, D. and S. Rathor. *Bone fracture detection and classification using deep learning approach*. in *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*. 2020. IEEE.
3. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. 2015. Springer.
4. Yadav, D.P., et al., *Hybrid SFNet model for bone fracture detection and classification using ML/DL.* Sensors, 2022. **22**(15): p. 5823.
5. Samothai, P., et al. *The evaluation of bone fracture detection of yolo series*. in *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. 2022. IEEE.
6. Luo, Y., et al., *Liquid Natural Gas Cold Energy Recovery for Integration of Sustainable District Cooling Systems: A Thermal Performance Analysis.* Inventions, 2023. **8**(5): p. 121.
7. Barhoom, A.M., M.R.J. Al-Hiealy, and S.S. Abu-Naser, *Bone abnormalities detection and classification using deep learning-vgg16 algorithm.* Journal of Theoretical and Applied Information Technology, 2022. **100**(20): p. 6173-6184.
8. Curso, *Bone Break Classification Dataset*, in *Roboflow Universe*. 2024.
9. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. pmlr.
10. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
11. He, H. and E.A. Garcia, *Learning from imbalanced data*. IEEE Transactions on knowledge and data engineering, 2009. **21**(9): p. 1263-1284.