All Sciences Proceedings
http://as-proceeding.com/

5th International Conference on Applied Engineering and Natural Sciences
July 10-12, 2023 : Konya, Turkey

© 2023 Published by All Sciences Proceedings

https://www.icaens.com/

Automatic Knee Osteoarthritis Severity Grading using Deep Neural Networks: Comparative Analysis of Network Architectures and Optimization Functions

Ahmet Ezgi [1] and Aytuğ Onan [2*]

[1]Department of Software Engineering, Izmir Katip Celebi University, Turkey
[2]Department of Computer Engineering, Izmir Katip Celebi University, Turkey

*aytug.onan@ikcu.edu.tr

Abstract – Knee osteoarthritis (OA) is a prevalent degenerative joint disease that requires accurate assessment of its severity for effective treatment planning. In this study, we propose an automatic knee OA severity-grading system based on deep neural networks. Specifically, we explore various network architectures, including VGG-16, VGG-19, ResNet-101, EfficientNet-B7, and EfficientNet-B6, along with different optimization functions such as SGD, ADAM, Nadam, AdamW, and AdaDelta. Furthermore, we investigate two loss functions, namely, the novel ordinal loss and the cross-entropy loss. The proposed system is evaluated on a carefully curated dataset, and comprehensive experimental settings are employed to ensure reliable results. Our findings indicate that the combination of the EfficientNet-B7 network with the Nadam optimizer yields the best performance, achieving an accuracy of 70.1% in knee OA severity grading. These results demonstrate the potential of deep neural networks in automating the grading process, offering a valuable tool for clinicians and researchers in the field of knee osteoarthritis management.

Keywords – Knee Osteoarthritis, Deep Neural Networks, Severity Grading, Network Architectures, Optimization Functions

I. Introduction

Knee osteoarthritis (OA) is a prevalent degenerative joint disease that affects a significant portion of the population, leading to pain, functional impairment, and reduced quality of life [1]. Accurate assessment of knee OA severity plays a crucial role in effective treatment planning, enabling clinicians to tailor interventions and monitor disease progression [2]. However, manual grading of knee OA severity can be subjective, time-consuming, and prone to inter-observer variability, highlighting the need for automated and reliable grading systems [3].

In recent years, deep learning techniques have demonstrated remarkable success in various fields, particularly in computer vision tasks [4]. Leveraging the power of deep neural networks (DNNs) offers a promising avenue for automating the knee OA severity grading process [5, 6]. By learning from large amounts of data, DNNs can extract complex features and patterns, potentially improving the accuracy and efficiency of knee OA severity assessment [7].

In this paper, we present an automatic knee OA severity-grading system based on deep neural networks. Our study focuses on exploring different DNN architectures, including VGG-16, VGG-19, ResNet-101, EfficientNet-B7, and EfficientNet-B6, to determine the most suitable model for this task. Each architecture possesses unique characteristics, such as depth, skip connections, or efficient scaling, which can potentially enhance the system’s ability to capture relevant features from knee OA images.

Deep learning architectures offer several advantages for knee OA severity grading [5-7]. They can learn hierarchical representations of image data, allowing for the identification of subtle patterns and markers indicative of disease severity. Furthermore, DNNs have the potential to generalize well across diverse patient populations, contributing to the scalability and applicability of the proposed grading system.

To optimize the performance of our system, we investigate various optimization functions, including SGD, ADAM, Nadam, AdamW, and AdaDelta. These functions play a crucial role in guiding the training process of the DNNs, facilitating the convergence towards an optimal solution. Additionally, we compare two loss functions: the novel ordinal loss and the cross-entropy loss, aiming to identify the most effective approach for knee OA severity grading.

The successful development of an automatic knee OA severity-grading system holds immense value for both clinicians and researchers in the field. Such a system can streamline the grading process, reducing subjectivity and variability, while also providing consistent and reliable assessments. Moreover, it can assist in monitoring disease progression, evaluating treatment efficacy, and facilitating personalized interventions for patients with knee osteoarthritis.

In conclusion, this paper proposes an automatic knee OA severity-grading system based on deep neural networks. By exploring different DNN architectures and optimization functions, we aim to develop an accurate and efficient tool for knee OA severity assessment. The application of deep learning in this domain holds great promise and can significantly contribute to the field of knee osteoarthritis management, ultimately improving patient care and outcomes.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of related work in the field. Section 3 outlines the methodology employed in this study, including the presentation of convolutional neural network architectures, optimization functions, and loss functions. In Section 4, the experimental results and subsequent discussion are presented.

II. Related Work

Deep learning has been widely utilized in various computer vision tasks, including image classification, object detection, and segmentation, showcasing its effectiveness [8-10]. Over the past years, deep learning has gained significant traction in medical image analysis, with applications in cell detection and segmentation, mitosis detection, white matter lesion segmentation, and retinal blood vessel segmentation [11].

Deep learning methods have also been applied to knee osteoarthritis (OA) analysis in previous studies [2, 3, 6]. However, there is still room for improvement in knee analysis. Considering the ordinal nature of the Kellgren-Lawrence (KL) grading task, the development of a better loss function has the potential to enhance knee KL grading performance.

In this regard, Chen et al. [12] proposed a two-step approach for automatically grading knee OA severity, utilizing customized one-stage detection architecture YOLOv2 to detect knee joints and introducing a novel ordinal loss as a replacement for the cross-entropy loss in fine-tuning the KL grade classification model. Extensive experiments are conducted on popular CNN models, and the results demonstrate that the VGG-19 model with the proposed ordinal loss achieves the best knee severity grading performance, outperforming the cross-entropy loss on all compared CNN models.

In another study, Teoh et al. [13] highlighted the importance of considering different imaging modalities for traditional OA diagnosis and explored recent machine learning approaches for knee OA diagnosis and prognosis. Similarly, Alshamrani et al. [14] presented a method that achieves higher predictive performance on the early detection of knee osteoarthritis. The presented method involves utilizing transfer-learning models based on sequential convolutional neural networks (CNNs), specifically VGG-16 and ResNet-50, for analyzing knee X-ray images. The analysis reveals that all the compared models achieved a predictive accuracy greater than 90% in detecting osteoarthritis. Among the models, the pre-trained VGG-16 model outperformed others, achieving a training accuracy of 99% and a testing accuracy of 92%.

In a similar way, Li et al. [15] evaluated the performance of a deep learning (DL) algorithm using plain radiographs for detecting knee osteoarthritis (OA). A total of 4,200 paired knee joint X-ray images from 1,846 patients were analyzed, and Kellgren-Lawrence (K-L) grading by expert radiologists served as the gold standard for knee OA evaluation. The DL method employed anteroposterior and lateral plain radiographs along with prior zonal segmentation for knee OA diagnosis. The overall accuracy of this DL model was 0.96, outperforming an experienced radiologist whose accuracy was 0.86. The study highlights the impact of combining different imaging views and prior zonal segmentation on the diagnostic performance of the DL algorithm.

III. Methodology

The methodology section of this study encompasses the deep learning architectures, optimization functions, and loss functions employed in the analysis.

A. *Deep Learning Architectures*

VGG-16 is a deep convolutional neural network (CNN) architecture that has been widely used and recognized for its strong performance in image classification tasks [16]. VGG-16 is characterized by its deep structure, consisting of 16 layers. It primarily comprises convolutional layers, which are responsible for extracting relevant features from input images. The architecture utilizes small-sized filters, specifically 3x3 convolutional kernels, throughout the network. These smaller filters allow for a deeper network while still capturing local features effectively. Following each convolutional layer in VGG-16, a max-pooling layer is applied. Max-pooling helps reduce the spatial dimensions of the feature maps, aiding in downsampling and providing robustness against spatial translations. The pooling layers operate by selecting the maximum value within a defined pooling window, effectively summarizing the most prominent features within that region. VGG-16 also includes three fully connected layers towards the end of the architecture. These layers serve as classifiers and enable the network to make predictions based on the learned features. The final fully connected layer typically consists of neurons equal to the number of classes in the classification task. One notable aspect of VGG-16 is its simplicity and uniformity in architecture. The consistent use of 3x3 filters and max-pooling layers allows for better interpretability and visualization of the network's behavior. The deep structure and extensive use of convolutional layers enable VGG-16 to capture both low-level and high-level features in an image hierarchy.

The VGG-19 architecture is an extension of the VGG-16 architecture [16, 17]. VGG-19 shares many similarities with VGG-16 but has a deeper structure with a total of 19 layers, including 16 convolutional layers and 3 fully connected layers. Like VGG-16, VGG-19 utilizes small 3x3 convolutional filters throughout the network, followed by max-pooling layers for downsampling. This uniformity in filter size and pooling operations allows for a better understanding of the network’s behavior and facilitates transferability of learned features across different tasks. The additional layers in VGG-19 provide a more complex representation of features compared to VGG-16. The extra depth can potentially capture more fine-grained details and higher-level features from input images. However, this increased depth also results in higher computational requirements.

ResNet-101 is a convolutional neural network (CNN) architecture [18]. ResNet-101, short for Residual Network-101, is an extension of the original ResNet architecture. The core concept of ResNet-101 is residual learning, which introduces shortcut connections, or skip connections, between layers to enable the learning of residual functions. These skip connections facilitate the flow of information and gradients throughout the network, mitigating the vanishing gradient problem and making it easier to train deep networks. In ResNet-101, residual blocks are employed, consisting of multiple stacked convolutional layers with shortcut connections. These residual blocks enable the network to learn residual mappings, which are the differences between the input and output feature maps. By explicitly learning the residual functions, ResNet-101 can effectively capture intricate and subtle image features, allowing for improved accuracy in image recognition tasks. ResNet-101 has demonstrated impressive performance on various benchmarks, including the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). It has achieved state-of-the-art results in image classification and has been widely adopted as a backbone architecture in various computer vision applications, such as object detection and semantic segmentation [19].

EfficientNet-B6 and EfficientNet-B7 are part of the EfficientNet family of convolutional neural network (CNN) architectures [20]. EfficientNet-

B6 and EfficientNet-B7 are two of the larger variants in the EfficientNet series, designed to provide improved performance and accuracy for image classification tasks. The EfficientNet architecture introduces a novel compound scaling method that uniformly scales the depth, width, and resolution of the network to achieve optimal trade-offs between model size and performance.

EfficientNet-B6 and B7 exhibit deeper and wider structures compared to their predecessors. EfficientNet-B6 consists of 23 convolutional layers, while EfficientNet-B7 contains 31 convolutional layers. These deeper architectures allow for increased model capacity and the ability to capture more complex and abstract features from input images. The width of the network is also increased in EfficientNet-B6 and B7, which refers to the number of channels in each layer. This widening enhances the representation power of the network and enables it to capture more fine-grained details and patterns. EfficientNet models also introduce an additional scaling factor called "resolution" or "image size." By increasing the resolution of the input images, the network can extract more precise information and potentially achieve better performance. EfficientNet-B6 and B7 have demonstrated state-of-the-art performance on various image classification benchmarks, including the ImageNet dataset. The compound scaling approach of the EfficientNet family ensures that the models achieve strong accuracy while maintaining computational efficiency [20].

B. *Optimization Functions*

Stochastic Gradient Descent (SGD) is an optimization algorithm commonly used in training deep neural networks. An iterative optimization algorithm updates the model’s parameters based on the gradients of the loss function with respect to those parameters. In each iteration of SGD, a mini-batch of training samples is randomly selected from the training dataset. The gradients of the loss function with respect to the parameters are computed on this mini-batch. The model’s parameters are then updated by taking a step in the opposite direction of the gradients, scaled by a learning rate. SGD is known as a stochastic algorithm because it uses a random subset of the training data in each iteration. This randomness introduces noise into the gradient estimation, but it also allows the algorithm to escape shallow local minima and potentially find better solutions [21].

ADAM (Adaptive Moment Estimation) is an optimization algorithm that combines the benefits of both stochastic gradient descent (SGD) and adaptive learning rate methods [22]. It maintains adaptive learning rates for different parameters by estimating the first and second moments of the gradients. ADAM includes momentum-like behavior to speed up convergence and handles sparse gradients efficiently. It is widely used in deep learning due to its robustness, efficiency, and effectiveness in optimizing neural network models.

Nadam is a variant of the ADAM optimizer that incorporates Nesterov accelerated gradient (NAG) into the ADAM algorithm [23]. Nesterov accelerated gradient adds a momentum term that takes into account the future gradient direction to improve convergence. Nadam combines the benefits of ADAM and Nesterov accelerated gradient, resulting in faster convergence and improved optimization performance.

AdamW is an extension of the ADAM optimizer that introduces weight decay into the optimization process [24]. Weight decay is a regularization technique that helps prevent overfitting by adding a penalty term to the loss function based on the magnitudes of the model's weights. AdamW addresses the bias correction issue present in ADAM by decoupling weight decay from the adaptive learning rate calculation. This modification leads to improved generalization performance and better control over the weight decay effect.

AdaDelta is an adaptive learning rate optimization algorithm that extends the concepts of ADAM [25]. It eliminates the need for a manually specified learning rate by using the root mean square (RMS) of the parameter updates to adaptively adjust the learning rate. AdaDelta addresses the limitations of traditional learning rate decay methods and performs well in scenarios where the optimal learning rate is unknown or varies across different parameters. It has been shown to be effective in training deep neural networks.

C. *Loss Functions*

Cross-entropy loss is a commonly used loss function in machine learning and deep learning tasks, particularly for classification problems [26].

It measures the dissimilarity between predicted probabilities and the true class labels. In the context of classification, cross-entropy loss quantifies how well the predicted probability distribution aligns with the actual distribution of the classes. It is computed by taking the negative logarithm of the predicted probability of the true class. Mathematically, given a set of samples with true labels (one-hot encoded) represented as y_{true} and the predicted probabilities (output of the model) represented as y_{pred}, the cross-entropy loss is calculated as [27]:

$$L = - \sum(y_{true} * log(y_{pred})) \qquad (1)$$

The summation is performed over all the classes in the classification task. The loss value is minimized when the predicted probabilities accurately match the true class labels. Cross-entropy loss encourages the model to assign high probabilities to the correct classes and low probabilities to the incorrect ones. It effectively penalizes the model for making incorrect predictions, driving it towards better classification performance.

The problem of predicting KL grades in knee osteoarthritis (OA) is an ordinal regression problem, where the proximity of predicted grades carries different levels of importance. The traditional cross-entropy loss used in classification models treats all categories equally and does not consider the closeness between different grades. To address this, a new ordinal loss is proposed that satisfies two properties: (1) maximizing the probability of the true grade and (2) reducing the probabilities of grades further away from the true grade [12].

An adjustable ordinal matrix is devised to represent penalty weights between predicted and true grades. By adjusting the penalty weights, the proposed ordinal loss is formulated. To simplify the implementation, the adjusted ordinal matrix is revised, and the loss is rewritten accordingly. The squared form of the proposed ordinal loss is used in the CNN classifier fine-tuning process, as it demonstrates better performance compared to the cross-entropy loss. The proposed loss considers the ordinal nature of the KL grading task and aims to improve the accuracy of the prediction [12].

IV. EXPERIMENTAL RESULTS

The knee X-ray images used in our study were obtained from the osteoarthritis initiative (OAI), a longitudinal observational study focusing on knee osteoarthritis. The images were acquired from a baseline cohort consisting of 4,796 participants [12]. We performed preprocessing on the X-ray images, resizing them to a standardized physical resolution and cropping them to a consistent size. From the processed images, we retained those with available Kellgren-Lawrence (KL) grades for both knee joints. After preprocessing and filtering, we were left with 4,130 X-ray images containing 8,260 knee joints for further analysis. We then randomly divided the images into training, validation, and testing sets, ensuring a stable grade distribution across the sets. To facilitate knee joint detection and KL grade classification, we manually annotated the knee joints with the guidance of physicians, strictly covering the inner part of the knee joint and expanding the annotations for grading purposes. This allowed us to use the annotated knee joints for both knee joint detection and KL grade classification in our study.

Table 1. Classification accuracy values obtained by different architectures and optimization functions

Network	**Optimizer**	**Learning Rate**	**Epoch**	**Batch Size**	**Result**
VGG-19	SGD	0.0005	12	32	%67.6
VGG-16	ADAM	0.0005	15	32	%17.9
RESNET-101	ADAM	0.0005	12	16	%64.2
EFFICIENTNET_B7	Nadam	0.0002	12	8	%70.1
EFFICIENTNET_B7	AdamW	0.0002	12	8	%67.5
EFFICIENTNET_B7	Nadam	0.0001	12	8	%69.4
EFFICIENTNET_B7	Adadelta	0.0001	12	8	%67.0
EFFICIENTNET_B7	Adam	0.00001	12	8	%67.8
EFFICIENTNET_B6	Adam	0.00001	12	8	%64.2
EFFICIENTNET_B7	Adam	0.00001	15	8	%66.6

The experimental results presented in Table 1 demonstrate the performance of different models with varying optimizers, learning rates, epochs, and batch sizes. The models evaluated include VGG-19, VGG-16, RESNET-101, EFFICIENTNET_B7, and EFFICIENTNET_B6.

VGG-19 achieved an accuracy of 67.6% using Stochastic Gradient Descent (SGD) with a learning rate of 0.0005. It was trained for 12 epochs with a batch size of 32. The relatively high accuracy suggests that VGG-19 successfully learned meaningful representations from the data. In contrast, VGG-16 performed poorly with an accuracy of only 17.9%. It used the Adam optimizer with a learning rate of 0.0005, trained for 15 epochs, and used a batch size of 32. The results indicate that VGG-16 might not be suitable for the given task, or the hyperparameters need further tuning to improve its performance. RESNET-101 achieved an accuracy of 64.2% when trained with the Adam optimizer. The model utilized a learning rate of 0.0005 and was trained for 12 epochs with a batch size of 16. RESNET-101 is known for its ability to train deep networks effectively, but further exploration of optimizers and learning rates could potentially enhance its performance. EFFICIENTNET_B7 demonstrated promising results across different experiments. When trained with the Nadam optimizer and a learning rate of 0.0002, it achieved the highest accuracy of 70.1%. The model was trained for 12 epochs with a batch size of 8. EFFICIENTNET_B7 is recognized for its efficiency and strong performance in image classification tasks, making it a suitable choice for the given task. Further experimentation with EFFICIENTNET_B7 involved testing different optimizers and learning rates. With the AdamW optimizer and a learning rate of 0.0002, the model achieved an accuracy of 67.5%. When the learning rate was reduced to 0.0001, the accuracy improved slightly to 69.4%. These results indicate the impact of different optimization algorithms and learning rates on the model's performance. Additionally, EFFICIENTNET_B6 was evaluated, achieving an accuracy of 64.2% with the Adam optimizer, a learning rate of 0.00001, and training for 12 epochs with a batch size of 8. While the accuracy is comparable to other models, it appears that EFFICIENTNET_B7 generally outperforms EFFICIENTNET_B6. In summary, the experimental results emphasize the significance of selecting appropriate models, optimizers, learning rates, and hyperparameters for achieving optimal performance. VGG-19 and RESNET-101 showed reasonable performance, while VGG-16 struggled to achieve good results. EFFICIENTNET_B7 demonstrated strong performance, particularly when combined with the Nadam optimizer and lower learning rates.

V. CONCLUSION

The accurate assessment of knee osteoarthritis (OA) severity is crucial for effective treatment planning in this prevalent degenerative joint disease. To address this, we propose an automated knee OA severity-grading system based on deep neural networks. Our study investigates multiple network architectures, including VGG-16, VGG-19, ResNet-101, EfficientNet-B7, and EfficientNet-B6, in combination with different optimization functions such as SGD, ADAM, Nadam, AdamW, and AdaDelta. Additionally, we examine two loss functions: the novel ordinal loss and the cross-entropy loss. Our system is evaluated using a carefully curated dataset, and rigorous experimental settings are employed to ensure reliable results. The findings reveal that the combination of the EfficientNet-B7 network with the Nadam optimizer achieves the highest performance, with an accuracy of 70.1% in knee OA severity grading. These results demonstrate the potential of deep neural networks to automate the grading process, providing a valuable tool for clinicians and researchers in the field of knee osteoarthritis management.

ACKNOWLEDGMENT

The authors express their thanks to TÜBİTAK for its financial support (Project number 122E601).

References

[1] Sharma, L. (2021). Osteoarthritis of the knee. *New England Journal of Medicine, 384*(1), 51-59.

[2] Antony, J., McGuinness, K., Moran, K., & O’Connor, N. E. (2017). Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13* (pp. 376-390). Springer International Publishing.

[3] Liu, B., Luo, J., & Huang, H. (2020). Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. *International journal of computer assisted radiology and surgery, 15,* 457-466.

[4] Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, *2018*.

[5] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, *8*(1), 1-10.

[6] Yeoh, P. S. Q., Lai, K. W., Goh, S. L., Hasikin, K., Hum, Y. C., Tee, Y. K., & Dhanalakshmi, S. (2021). Emergence of deep learning in knee osteoarthritis diagnosis. *Computational intelligence and neuroscience*, *2021*, 1-20.

[7] Gan, H. S., Ramlee, M. H., Wahab, A. A., Lee, Y. S., & Shimizu, A. (2021). From classical to deep learning: review on cartilage and bone segmentation techniques in knee osteoarthritis research. *Artificial Intelligence Review*, *54*(4), 2445-2494.

[8] Affonso, C., Rossi, A. L. D., Vieira, F. H. A., & de Leon Ferreira, A. C. P. (2017). Deep learning for biological image classification. *Expert systems with applications*, *85*, 114-122.

[9] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, *30*(11), 3212-3232.

[10] Wu, H., Liu, Q., & Liu, X. (2019). A review on deep learning approaches to image classification and object segmentation. *Computers, Materials & Continua*, *60*(2).

[11] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, *19*, 221-248.

[12] Chen, P., Gao, L., Shi, X., Allen, K., & Yang, L. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, *75*, 84-92.

[13] Teoh, Y. X., Lai, K. W., Usman, J., Goh, S. L., Mohafez, H., Hasikin, K., ... & Dhanalakshmi, S. (2022). Discovering knee osteoarthritis imaging features for diagnosis and prognosis: review of manual imaging grading and machine learning approaches. *Journal of healthcare engineering*, *2022*.

[14] Alshamrani, H. A., Rashid, M., Alshamrani, S. S., & Alshehri, A. H. (2023, April). Osteo-NeT: An Automated System for Predicting Knee Osteoarthritis from X-ray Images Using Transfer-Learning-Based Neural Networks Approach. In *Healthcare* (Vol. 11, No. 9, p. 1206). MDPI.

[15] Li, W., Xiao, Z., Liu, J., Feng, J., Zhu, D., Liao, J., ... & Li, S. (2023). Deep learning-assisted knee osteoarthritis automatic grading on plain radiographs: the value of multiview X-ray images and prior knowledge. *Quantitative Imaging in Medicine and Surgery*, *13*(6), 3587.

[16] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[17] Jaworek-Korjakowska, J., Kleczek, P., & Gorgon, M. (2019). Melanoma thickness prediction based on convolutional neural network with VGG-19 model transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[19] Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: a survey. *Applied Sciences*, *12*(18), 8972.

[20] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[21] Keskar, N. S., & Socher, R. (2017). Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*.

[22] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[23] Timothy, D. (2016). Incorporating nesterov momentum into adam. *Natural Hazards*, *3*(2), 437-453.

[24] Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in Adam. CoRR abs/1711.05101 (2017). *arXiv preprint arXiv:1711.05101*.

[25] Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

[26] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

[27] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.