1. Michał BUKOWSKI[1], 2. Albina JEGOROWA[2], 3. Jarosław KUREK[1]

Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences (1),
Department of Mechanical Processing of Wood, Institute of Wood Sciences and Furniture, Warsaw University of Life Sciences (2)
ORCID: 1. 0000-0003-1567-879X; 2. 0000-0002-8935-845X, 3. 0000-0002-2789-4732

doi:10.15199/48.2024.05.52

A Novel Approach using Vision Transformers (VIT) for Classification of Holes Drilled in Melamine Faced Chipboard

Streszczenie. *Artykuł ten przedstawia szczegółową ocenę wydajności różnych architektur sztucznej inteligencji do klasyfikacji otworów wiertniczych w płytach wiórowych laminowanych. Badanie obejmuje własną sieć neuronową konwolucyjną (CNN), pięciokrotną sieć CNN, VGG19, pojedyncze i pięciokrotne VGG16, zespół sieci CNN, VGG19 i 5xVGG16, oraz transformery wizyjne (ViT). Wydajność każdego modelu mierzono i porównywano na podstawie dokładności klasyfikacji. Modele transformatorów wizyjnych, szczególnie model B_32 trenowany przez 8000 epok, wykazały wyższą skuteczność, osiągając dokładność 71.14%. Pomimo tego osiągnięcia, badanie podkreśla potrzebę równoważenia wydajności modelu z innymi aspektami, takimi jak zasoby obliczeniowe, złożoność modelu i czas szkolenia. Wyniki zwracają uwagę na znaczenie starannego doboru i dopracowania modelu, kierując się nie tylko wskaźnikami wydajności, ale także konkretnymi wymaganiami i ograniczeniami zadania i kontekstu. Studium stanowi solidną podstawę do dalszych badań nad innymi modelami opartymi na transformatorach oraz zachęca do głębszych badań nad dopracowaniem modeli w celu w pełni wykorzystania potencjału tych architektur SI w zadaniach klasyfikacji obrazów.* (***Nowatorskie podejście z wykorzystaniem transformatorów wizyjnych (VIT) do klasyfikacji otworów wierconych w płytach wiórowych pokrytych melaminą***)

Abstract. *This paper presents a comprehensive performance evaluation of various AI architectures for a classification of holes drilled in melamine faced chipboard, including custom Convolutional Neural Network (CNN-designed), five-fold CNN-designed, VGG19, single and five-fold VGG16, an ensemble of CNN-designed, VGG19, and 5xVGG16, and Vision Transformers (ViT). Each model's performance was measured and compared based on their classification accuracy, with the Vision Transformer models, particularly the B_32 model trained for 8000 epochs, demonstrating superior performance with an accuracy of 71.14%. Despite this achievement, the study underscores the need to balance model performance with other considerations such as computational resources, model complexity, and training times. The results highlight the importance of careful model selection and fine-tuning, guided not only by performance metrics but also by the specific requirements and constraints of the task and context. The study provides a strong foundation for further exploration into other transformer-based models and encourages deeper investigations into model fine-tuning to harness the full potential of these AI architectures for image classification tasks.*

Słowa kluczowe: Vision Transformer, Convolutional Neural Network, monitorowanie stanu narzędzia, płyta wiórowa laminowana
Keywords: Vision Transformer, Convolutional Neural Network, tool state monitoring, melamine faced chipboard

Introduction

The process of manufacturing furniture involves intricate and precision-demanding steps. One such critical stage is drilling holes in melamine faced chipboard, where errors can lead to significant financial losses due to reduced product quality. Traditionally, the condition of the drill is manually monitored to identify the optimal moment for replacement, thus ensuring consistently high product quality. While manual monitoring offers some control, it is not sufficiently efficient. As such, there is a pressing need for a more automated, accurate, and efficient solution.

In the pursuit of this solution, tool condition monitoring (TCM) methodologies have been developed to evaluate and assess the state of various utensils, including drills. Such methods often require a multitude of diverse sensors to collect data, which is subsequently used to diagnose the drill's condition [7]. While these approaches can produce accurate results, they often necessitate extensive preprocessing, and mistakes at any stage can compromise the final result. Moreover, these solutions can be costly and complex to implement and maintain. Despite the advanced features generated from the vast array of registered signals, the accuracy of such solutions rarely surpasses 90% [3], [4].

The incorporation of machine learning algorithms into the wood industry is a growing trend. For example, algorithms have been developed to recognize wood species based on macroscopic texture images [2]. When using image-based samples, convolutional neural networks (CNN) are often applied [1], [5], [6], [8], [9], [17], [18]. However, their application often encounters obstacles such as the requirement of large datasets and the need for close cooperation with manufacturers to ensure task specificity, among other factors. Considering these limitations, this work presents a novel approach applying Vision Transformers for classifying holes drilled in melamine faced chipboard. This approach attempts to mitigate the complexities and enhance the adaptability of the system to specific manufacturing requirements. The primary enhancement lies in the elimination of complex equipment, reducing the requirement to a camera that captures images of the drilled holes. These images form the basis for assessing the drill's condition. Prior research, which tested various algorithms such as CNNs, transfer learning, and data augmentation methodologies, confirmed that this approach can accurately predict the state of the drill based solely on images, while improving overall prediction accuracy [5], [6], [8], [9]. Given the state of the art in the field of artificial intelligence, this paper makes use of Vision Transformers (ViT), a revolutionary architecture that is currently considered a benchmark in the field. Unlike traditional convolutional neural networks, which process image data in a local and hierarchical manner, Vision Transformers treat image data as a sequence of patches and leverage self-attention mechanisms to capture global dependencies. This technique has shown unprecedented success in various image classification tasks, outperforming established CNN architectures on multiple benchmarks. In this paper, we demonstrate the applicability and effectiveness of Vision Transformers in the context of classifying holes drilled in melamine faced chipboard, aiming to further advance tool condition monitoring practices.

Data Set

The dataset comprises images of holes drilled during the experiment. The images were collected in collaboration with the Institute of Wood Sciences and Furniture at the Warsaw University of Life Sciences. A standard CNC vertical machining centre, Busellato Jet 100, Thiene, Italy, was used for the drilling process. The material drilled was a standard laminated chipboard (U511SM – Swiss Krono 88 Group), typically used in the furniture industry, with dimensions of 2500x300x18. A 12mm Faba WP-01 drill with a tungsten carbide tip was utilized.

Five different drills were used during the drilling process. Each drill underwent cycles of operation, and the external corner wear parameter was monitored between cycles. This allowed for assigning appropriate classes (Green, Yellow, Red) to the obtained images based on the level of drill wear. Images produced by each drill were stored separately, preserving the order of creation to reflect the gradual deterioration of drill condition. This could serve as additional information during the learning process. Table 1 provides details about the data acquisition process and final corner wear measurements for each drill at the end of the last drilling cycle.

Table 1. Sample counts for each class before and after data augmentation. Values in each cell are presented in following order: Green, Yellow, Red.

No. of drill	Original	Augmented	Total
1	840/420/406	840/840/840	2,520
2	840/700/280	840/840/840	2,520
3	700/560/420	700/700/700	2,100
4	840/560/280	840/840/840	2,520
5	560/560/560	560/560/560	1,680

Mentioned three classes correspond to the condition of the drill used to make the holes:

- Green class: good condition, where the drill is new and not yet worn, can be further used.
- Yellow class: worn condition, where the drill is used and may require manual evaluation to determine if it is still good enough for production.
- Red class: requiring replacement, where the drill is used to a point of being unusable and should be replaced immediately.

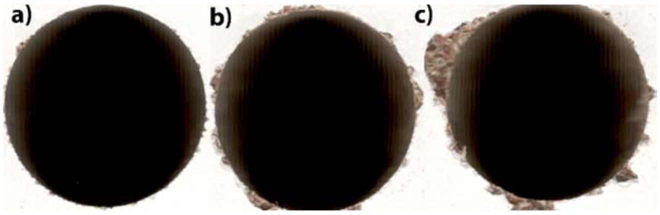

Figure.1. Evaluation of drill wear - Three classes of drill hole conditions in laminated chipboard: a) green class: hole made by a new, unworn drill, b) yellow class: hole made by a used drill, requiring manual evaluation for further use, c) red class: hole made by a drill that is too worn for use and needs immediate replacement.

Vision Transformer model

Convolutional Neural Networks (CNNs) have traditionally been the go-to models for computer vision tasks due to their ability to exploit local correlations in data, thanks to their convolutional nature. However, their efficiency can be undermined when the tasks require understanding the global context of an image or when the data lacks the aforementioned local correlation, due to which the CNN's hierarchical structure might introduce unnecessary inductive bias.

The Vision Transformer (ViT) model, introduced by Google Research [10]-[16], presents a departure from the typical CNNs by adopting a transformer architecture, originally de- signed for natural language processing tasks. Transformers rely on self-attention mechanisms, providing a better under- standing of global context as every part of the input contributes to the final representation of every other part.

Advantages of ViT:

ViTs possess several advantages over CNNs:

- Global Context Understanding: ViTs are more capable of understanding the global context in images as they are not restricted to local features.
- Parameter Efficiency: ViTs often require fewer parameters than CNNs for similar performance, which can lead to more efficient models.
- Transfer Learning: ViTs can leverage the benefits of transfer learning better, due to their capability to learn from both vision and language domains.

Disadvantages of ViT:

Despite their advantages, ViTs also have some limitations:

- Computational Requirements: ViTs are often more computationally intensive than CNNs, especially for larger inputs. This might be a limiting factor for real-time applications or for deployments on devices with limited computational power.
- Training Data: ViTs typically require more training data to outperform CNNs. This might be a constraint when only limited labeled data are available.

In consideration of these factors, the decision to utilize ViT for our study was informed by the global nature of our task, where understanding the entire context of each image is crucial. Furthermore, given that we have a sufficiently large dataset for training, the advantages of using ViT outweigh the potential drawbacks.

Numerical Experiments

We used transfer learning on the Vision Transformer (ViT) models, utilizing a range of pretrained models: R_Ti _16, S_32, B_32, R26_S_32, B_16, and S_16 [10]-[16].

The performance of each pretrained model was assessed in terms of mean accuracy, standard deviation of accuracy, and model size. The models were trained at different learning rates and over different numbers of epochs. The complete results of these experiments are presented in Table II.

Upon analysis of the experimental results, it can be observed that the model B 32, with a learning rate of 0.003 and trained over 8000 epochs, achieved the highest mean accuracy of 71.14% with a standard deviation of 0.35%. This model also had a size of 398 MiB.

When considering the balance between computational re- sources (model size and training time) and performance, the S 16 model also demonstrated impressive results. It achieved a mean accuracy of 70.54% with a standard deviation of 0.47%, trained with a learning rate of 0.01 over 4000 epochs, with a comparatively smaller model size of 115 MiB.

These experiments highlight the impact of different training parameters and model architectures on the performance of ViT models for the given task.

Table 2. Results of numerical experiments: Comparative Performance Metrics of Pretrained ViT Models.

Pretrained model name	Learning rate	Epochs	Mean Acc	Std Acc	Model Size
R_Ti_16	0.003	500	65.68%	1.14%	40 MiB
R_Ti_16	0.003	1000	64.90%	1.13%	40 MiB
R_Ti_16	0.003	2000	67.16%	0.35%	40 MiB
S_32	0.003	500	64.54%	1.21%	118 MiB
S_32	0.003	1000	68.06%	0.25%	118 MiB
B_32	0.003	500	65.60%	0.87%	398 MiB
B_32	0.003	1000	68.04%	0.98%	398 MiB
B_32	0.01	2000	68.55%	0.32%	398 MiB
B_32	0.003	4000	70.96%	0.22%	398 MiB
B_32	**0.003**	**8000**	**71.14%**	**0.35%**	**398 MiB**
R26_S_32	0.003	500	65.08%	0.97%	170 MiB
R26_S_32	0.003	1000	67.99%	0.71%	170 MiB
B_16	0.003	500	67.14%	1.52%	391 MiB
B_16	0.003	1000	68.69%	0.44%	391 MiB

B_16	0.01	1000	69.18%	0.53%	391 MiB
B_16	0.003	2000	70.00%	0.48%	391 MiB
S_16	0.003	500	66.71%	0.32%	115 MiB
S_16	0.003	1000	68.49%	0.27%	115 MiB
S_16	0.01	2000	69.83%	0.17%	115 MiB
S_16	0.03	2000	68.72%	0.95%	115 MiB
S_16	0.003	4000	69.54%	0.41%	116 MiB
S_16	0.01	4000	70.54%	0.47%	115 MiB

Numerical Experiments Using Different AI Architectures

The dataset comprises images of holes drilled during the experiment. The images were collected in collaboration with the Institute of Wood Sciences and Furniture at the Warsaw University of A wide array of machine learning models, including various artificial intelligence architectures, were applied in the numer- ical experiments conducted during this study. The primary fo- cus was on the assessment of their ability to perform accurate classifications and the results are compiled in Table 3.

The models under investigation included a self-designed Convolutional Neural Network (CNN-designed), five-fold implementation of the CNN-designed model (5xCNN-designed), VGG19, single and five-fold VGG16, an ensemble of CNN-designed, VGG19, and 5xVGG16 models, and the Vision Transformer (ViT) models.

The CNN-designed model and its five-fold counterpart were developed specifically for this task, adopting unique design principles derived from the nature of the data and the specifics of the classification problem.

The VGG models were implemented following their original design principles but were fine-tuned to the task at hand. Both the single and the five-fold VGG16 models were used, and an ensemble model was also implemented for a more diversified approach.

The Vision Transformer (ViT) models, a more recent development in the field of AI, were also used in the experiments. These models are based on the transformer architecture, which has shown exceptional performance in various machine learning tasks. Various configurations and training epochs of the ViT models were used in the numerical experiments.

The parameters and configurations for each model were meticulously selected and fine-tuned during preliminary testing and validation stages to optimize performance. The models were then trained on the same dataset to ensure a fair comparison of their performance.

Following the training phase, the models were evaluated on a test set, and their performance was assessed based on classification accuracy. The results of these experiments provide an in-depth understanding of how each model performs, allowing for a comprehensive comparison of different AI architectures.

Table 3. Classification results for chosen algorithms.

#	Model	Accuracy
1	CNN-designed	69.78%
2	5xCNN-designed	67.35%
3	VGG19	66.77%
4	5xVGG16	67.13%
5	10xVGG16	66.98%
6	Ensemble (1,3,4)	69.26%
7	**Vision Transformers**	**71.14%**

Discussion

The experimental results achieved during the performance evaluation phase offer several insightful takeaways, especially when the performance of Vision Transformer (ViT) models is juxtaposed with the performances of other AI modelling approaches. The performance comparison is presented in Table 3.

Considering the highest accuracy, ViT models outperform the rest of the models with an accuracy of 71.14%. Specifically, the B_32 model trained for 8000 epochs demonstrated the best performance among the pre-trained models used in this study. The ensemble approach combining CNN-designed, VGG19, and 5xVGG16 also resulted in high accuracy, but it was still lower than the top-performing ViT model.

The comparison also illustrates that the CNN-designed model and its five-fold implementation rendered decent performances with accuracies of 69.78% and 67.35% respectively. Nevertheless, their performances were not up to par with the B_ 32 ViT model.

For the VGG models, both single and five-fold VGG16 delivered similar results, and the performance was further improved when they were used in an ensemble with CNN-designed model, yet the results were less satisfactory than the ones obtained using ViT models.

This comparative evaluation thus accentuates the superior capability of ViT models for the task in focus. Despite using other high-performing AI models like CNN and VGG, the study confirmed the efficiency and effectiveness of pre-trained ViT models, especially with a large number of training epochs. In conclusion, while ViT models have proven to be a robust solution in this context, their use should be carefully assessed based on the available resources and the specific requirements of the task at hand.

Conclusion

This study presented an in-depth comparison of several AI architectures, including a custom Convolutional Neural Network (CNN-designed), five-fold CNN-designed, VGG19, single and five-fold VGG16, an ensemble of CNN-designed, VGG19, and 5xVGG16, and the Vision Transformers (ViT) for a classification task. This assortment of models allowed us to examine the advantages and disadvantages of each, using their performance on the same task as a point of comparison. Across the board, it was the ViT models, specifically the B_32 model trained for 8000 epochs, that outperformed all others, achieving an accuracy of 71.14%. The ViT models' success underscores the potential of transformer-based models in image classification tasks, a domain traditionally dominated by convolutional-based approaches. Despite this success, it is worth noting that the high performance of ViT models also came with increased model complexity and potentially longer training times, emphasizing the trade-offs often encountered in model selection.

The CNN-designed models and their five-fold counterparts also demonstrated respectable performance, highlighting the effectiveness of custom models tailored to a particular task. Meanwhile, the performance of the VGG models and ensemble model, although commendable, was still outclassed by the ViT models. The ensemble approach did demonstrate that combining models can yield higher performance, but not necessarily surpass the best individual models in this experiment.

To conclude, these experiments underline the importance of model selection and fine-tuning in achieving optimal performance. While the ViT models demonstrated superior performance in this study, it is critical to consider other factors, such as computational resources, model complexity, and training times. Indeed, the choice of model should be guided not only by performance metrics but also by the specific requirements and constraints of the task and context. Future work may include the exploration of other transformer-based models or further fine-tuning of the models studied here.

REFERENCES

[1] Hu, J., Song, W., Zhang, W., Zhao Y., Yilmaz A., (2019). Deep learning for use in lumber classification tasks Wood Sci Technol 53(2): 505-517.DOI: https://doi.org/10.1007/s00226-019-01086-z.

[2] Ibrahim, I., Khairuddin, A. S. M., Talip, M. S. A., Arof, H., Yusof, R.,

(2017). Tree species recognition system based on macroscopic image analysis. Wood science and technology, 51(2), 431-444.

[3] Jemielniak K., Urba´nski T., Kossakowska J., Bombi´nski S., (2012). Tool condition monitoring based on numerous signal features. Int J AdvManuf Technol 59: 73-81. DOI: https://doi.org/10.1007/s00170-011-3504-2.

[4] Kuo R., (2000). Multi-sensor integration for on-line tool wear estimation through artificial neural networks and fuzzy neural network. Eng Appl Artif Intell 13: 249-261. DOI: https://doi.org/10.1016/S0952-1976(00)00008-7.

[5] Kurek J., Antoniuk I., Górski J., Jegorowa A., Świderski B., Kruk M., Wieczorek G., Pach J., Orłowski A., Aleksiejuk-Gawron J., (2019a). Data Augmentation Techniques for Transfer Learning Improvement in Drill Wear Classification Using Convolutional Neural Network. Machine Graphics and Vision 28: 3-12.

[6] Kurek J., Antoniuk I., G´orski J., Jegorowa A., Świderski B., Kruk M., Wieczorek G., Pach J., Orłowski A., Aleksiejuk-Gawron J., (2019b). Classifiers ensemble of transfer learning for improved drill wear classification using convolutional neural network. Machine Graphics and Vision 28:13-23.

[7] Kurek J., Kruk M., Osowski S., Hoser P., Wieczorek G., Jegorowa A., Górski J., Wilkowski J., Śmietańska K., Kossakowska J., (2016). Developing automatic recognition system of drill wear in standard laminated chipboard drilling process Bulleting of the Polish Academy of Science. Technical Sciences 64: 633-640. DOI: https://doi.org/10.1515/bpasts-2016-0071.

[8] Kurek J., Swiderski B., Jegorowa A., Kruk M., Osowski S., (2017a). Deep learning in assessment of drill condition on the basis of images of drilled holes In: International Conference on Graphic and Image Processing. ICGIP. DOI: https://doi.org/10.1117/12.2266254.

[9] Kurek J., Wieczorek G., Świderski B., Kruk M., Jegorowa A., Osowski S., (2017b). Transfer learning in recognition of drill wear using convolutional neural network. 1. In: International Conference on Computational Problems of Electrical Engineering. IEEE. DOI: https://doi.org/10.1109/CPEE.2017.8093087.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words:Transformers for Image Recognition at Scale. ICLR. (2021)

[11] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M. & Dosovitskiy, A. MLP-Mixer: An all-MLP Architecture for Vision. ArXiv Preprint ArXiv:2105.01601. (2021)

[12] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. & Beyer, L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. ArXiv Preprint ArXiv:2106.10270. (2021)

[13] Chen, X., Hsieh, C. & Gong, B. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. ArXiv Preprint ArXiv:2106.01548. (2021)

[14] Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N.,Tatikonda, S., Duncan, J. & Liu, T. Surrogate Gap Minimization Improves Sharpness-Aware Training. ICLR. (2022)

[15] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A. & Beyer, L. LiT: Zero-Shot Transfer with Locked-image Text Tuning. CVPR. (2022)

[16] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. & Beyer, L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. (2022)

[17] Jegorowa, A.; Górski, J.; Kurek, J.; Kruk, M. Use of nearest neighbors (K-NN) algorithm in tool condition identification in the case of drilling in melamine faced particleboard. Maderas Cienc. Tecnol. 2020, 22, 189–196. https://doi.org/10.4067/S0718-221X2020005000205.

[18] Jegorowa, A., Kurek, J., Antoniuk, I., Dołowa, W., Bukowski, M. & Czarniak, P. Deep learning methods for drill wear classification based on images of holes drilled in melamine faced chipboard. Wood Science And Technology. 55, 271-293 (2021,1,1), https://doi.org/10.1007/s00226-020-01245-7