Scientific Papers. Series E. Land Reclamation, Earth Observation & Surveying, Environmental Engineering. Vol. XIII, 2024
Print ISSN 2285-6064, CD-ROM ISSN 2285-6072, Online ISSN 2393-5138, ISSN-L 2285-6064

WASTE CLASSIFICATION USING VISION TRANSFORMERS

Dan Constantin PUCHIANU

Valahia University of Targoviste, 13 Sinaia Alley Street, Targoviste, Romania

Corresponding author email: author_email@gmail.com

Abstract

Effective identification of recyclable waste is a major challenge in resource management and environmental protection. The present study explores the integration of transformer-based architectures for the accurate classification of recyclable waste, including plastic, glass, metal, and paper. A dataset consisting of digital images of different types of waste was used to train and evaluate the proposed architectures. To improve the generalization of the model a division of the data set was pursued for training, validation, and testing areas, as well as the implementation of data augmentation and transfer-learning techniques. Compared to traditional methods and different convolutional neural network architectures, transformer-based architectures have demonstrated superior performance both in terms of accuracy and computational efficiency. Analyzing the experimental results, the proposed models demonstrated accuracy values of over 95%. The study finally notes that the use of transformer-based architectures for the classification of waste from digital images presents a major potential in the development of efficient waste management practices and for reducing the impact of waste on the environment.

Key words*: deep learning, image classification, sustainability, vision transformers, waste management.*

INTRODUCTION

Pollution is one of the most pressing problems of modern society. Inefficient administration of waste management and pollution reduction solutions have a negative impact on the environment (Alrayes et al., 2023). The growth of the global population influences both the demand for and production of goods and services (Vîrsta et al., 2020; Agafitei & Pavel, 2023). Consequently, there is a greater risk of waste accumulation. Most of the time, the precise identification of waste resulting from the consumption and activities of modern society is based on manual work, being inefficient and time-consuming (Chen et al., 2023; Dong et al., 2022).

The field of Deep Learning and Computer Vision has experienced considerable advancement during the previous five-to-six-year period (Boldeanu et al., 2023). The advance of deep learning technologies and convolutional neural networks has brought to the fore innovative solutions in the field of detection and identification of objects illustrated in digital images and videos (An & Zhang, 2022). The processing and analysis of this data is a strong point of these architectures and at the same time an important research topic that is also applied in the field of waste management (Dookhee, 2022; Hu et al., 2022).

In this context and part of the present study, the automatic detection and classification of waste becomes a crucial step for its efficient management. The advanced technologies of convolutional neural networks (CNN) and vision transformers (VIT) offer modern solutions to this challenge (Huang et al., 2021; Kumar et al., 2023).

Because of their ability to analyse and understand digital images, convolutional neural networks have revolutionized the field of computer vision (Mao et al., 2021; Qin et al., 2024). These architectures consist of a multitude of convolutional layers and can extract relevant features from the input data. In the context of waste management, CNNs are used to identify and classify images containing waste of various categories (Kurz et al., 2022). These models can be trained on large data sets and their performance can be improved as the models are exposed to more examples (Ma et al., 2024).

Using Vision Transformer models, research is observed that capitalizes on the performance, features and popularity of these models. In the study carried out, the authors (Alrayes et al., 2023) discuss the rapid progress of deep

learning, facilitating the implementation of modern solutions for waste classification. To solve problems and limitations related to low accuracy and slow processing, the authors proposed a model based on the VIT architecture that was integrated with a hybrid multilayer network: VT-MLH-CNN. The developed model demonstrated remarkable performance with accuracy values up to 95.5%, outperforming state-of-the-art models and being effective for waste classification. Other key approaches presented the implementation of the model on a cloud service, highlighting the potential of the method to be integrated into real-time applications.

In the same context, another model, called WMC-Vit, was the basis of the authors' research (Kurz et al., 2022). The study proposes an efficient waste classification model using a combination of VIT structures and convolutional neural network models. The training and testing of the model was done based on a public dataset called TrashNet, the same being true for the case of the previous study. In this case, the model analyzes the images and achieves a maximum accuracy of 94.27%, integrating a Multi-Head block with parallel processing for transformer blocks. The authors note in the study the problem of constant growth of waste that can be effectively managed using carefully implemented deep learning models.

In another study, the authors (Huang et al., 2021), present an innovative method for waste identification, using the TrashNet dataset, transfer-learning techniques, and vision transformer-based architectures. The authors note the problem of manual sorting of waste, being expensive, time-consuming, and inefficient. To solve these limitations the authors proposed a VIT model, based on self-attention mechanisms to improve the accuracy of waste classification, finally obtaining an accuracy score of 96.98%. It was observed that the proposed model outperformed classical CNN and machine learning models. Moreover, to extend the applicability of the model, it was implemented on a cloud server that allows the automatic classification of waste using mobile devices.

The authors' study (Dong, Chen & Lu, 2022) explores the use of deep learning techniques and VIT models for the accurate identification of construction waste. The study proposes a model called BAT - Boundary-Aware Transformer for the precise identification of hard-to-classify waste. The key innovations of this model include a transformer structure with cascaded decoder and a careful edge processing model based on machine learning. The research also contributes to the implementation of a deep learning model for the semantic segmentation of construction waste with implementations aimed at waste sorting using systems with robotic arms and circular economy.

The present study proposes the implementation and optimization of architectures based on vision transformers for the automatic classification of various types of waste. The following sections detail the materials and methods used to implement the study, the experimental results obtained for waste classification, and finally the discussions and conclusions that note the relevance of the research.

MATERIALS AND METHODS

This part of the paper presents the materials used to implement the study, covering the dataset, the architectures used for the classification task, and the hardware and software details.

Creating a dataset is an important step for training and evaluating classification models as part of deep learning, especially for waste identification. For the present work, a dataset of RGB digital images was attached to the proposed classification models. Considering these details, it was aimed to divide the dataset into representative subsets for training, validation, and testing after a ration of 70%, 20% and 10% respectively. The dataset division process summed 1960 images for the training part, 560 images for the validation part, and 280 images for the test part. In total 2800 images were used.

The images in the dataset illustrate various types of waste, under varying conditions and from real contexts. Representative images taken from the dataset are illustrated in Figure 1.

No artificially generated images were used for the dataset of the present work, only pre-processing and augmentation processes were involved. Color transformations and geometric transformations have been implemented, as well as image resizing for an entry point standard specific to classification models (224x224px).

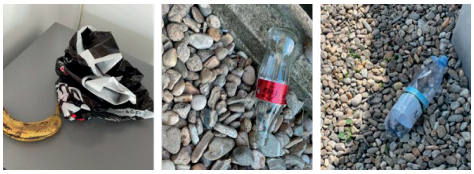

Figure 1. Example images from the dataset

Several vision transformer-based models were implemented and optimized for waste classification, following the dataset presented.
The first model used for the present study, VIT Base (Dosovitskiy et al., 2020), represents a modern architecture for image classification tasks. VIT uses a transformer-type architecture adapted from natural language processing that divides digital images into patches of fixed size for their analysis as sequences. At the same time, the architecture offers self-attention mechanisms to capture the global characteristics of the entire image. The next model, DeiT (Touvron et al., 2021), is an optimized version of the Vision Transformer. The core feature of this optimized architecture involves a teacher-student training strategy developed to improve classification performance and efficiency with less data. At the same time, the architecture presents structures of distillation tokens for improving accuracy and generalization, having applicability in real-time scenarios. Another model used, the MaxVIT Nano (Tu, Z. et al., 2022), completes the features of the vision transformer architectures, present as a compact version of the MaxVIT family. The main characteristic of this model is the combination of CNN and VIT strategies, using multi-axis self-attention mechanisms. This hybrid approach to the architecture translates into lightweight features that can be successfully applied in resource-constrained environments. Finally, the present study also used the Swin Transformer v2 (SwinV2) (Liu et al., 2022), an advanced architecture developed from the original Swin Transformer model. The present model represents a state-of-the-art architecture developed to improve image classification performance. The architecture stands out for the introduction of hierarchical shifted window structures that allow the model to capture spatial relationships at different scales and an improvement in the analysis of digital images. Other structures such as window-based attention mechanisms reduce the complexity of the architecture, increase the stability of training and the ability to generalize in various visual domains.
The last model used for this work, XCiT Small, represents another state-of-the-art architecture designed for image classification tasks based on Vision Transformer structures (Ali et al., 2021). The full name of the Cross-Covariance Image Transformer model brings key innovations compared to the VIT model and introduces new cross-covariance attention modules instead of the classic self-attention ones. In the same vein, the model introduces instead of spatial tokens modules that use channel-level attention, a key point to compute feature relationships, to reduce model complexity and optimize speed and accuracy. Key points of the architecture include feed-forward networks and multi-head attention to satisfy these features. For the present study, the small version of the model was used, being suitable for environments with low computing resources.
The hardware and software part of the system used in this study consisted of a customized setup with a Windows 10 Pro operating system, an AMD Ryzen 9 5900HS CPU, and a GeForce RTX 3060 6GB GPU. The selected programming language was Python v3.9, with PyTorch chosen as the framework.

RESULTS AND DISCUSSIONS

Analyzing the architectures used as well as the structure of the digital image dataset, the obtained results demonstrate the effectiveness of the proposed methodologies. For most of the models, accuracy values above 94-98% are noted, the metrics resulting from the evaluation of the models being presented in Table 1, using test images.

Each model demonstrated state-of-the-art classification performances for the waste dataset. The performance evolution for training and validation, loss and accuracy plots are presented in Figures 2-5, for each model.

Table 1. Waste classification performances for the proposed vision transformers models

Model	Accuracy
VIT Base	96.73%
MaxVit Nano	98.70%
SwinV2	97.39%
DeiT Base	93.46%
XCiT Small	94.10%

The presented results bring to the fore the stability of modern, optimized, and lightweight models such as SwinV2, XCiT or MaxVIT. The proposed architectures demonstrate notable performance for the task of the present study, focused on waste classification, being a good choice for the implementation of automatic identification applications.

For the present study, vision transformer-based models demonstrated very good accuracy values. Following the experimental results, all models show a good and fast convergence both in the case of the precision and accuracy functions. MaxVIT Nano and XCiT Small demonstrate the best and balanced performance, indicating a good generalization ability. The next top-performing model, SwinV2, shows slight fluctuations in validation accuracy, close to the results of the VIT Base model. DeiT performs below the previously presented models, an architecture that can highlight slight overfit/underfit issues, with a noticeable gap between validation and training performance. These features may indicate signs of considerable improvement, optimized for this task.

It can be noted that transformer models such as MaxVIT and SwinV2 are effective for the task of waste classification due to advanced architectures that combine innovative elements such as CNNs and Transformers to capture detailed features from digital images. Although performing well, the VIT Base and DeiT models can have difficulty generalizing, indicating a sensitivity to the analyzed data set.

The XCiT Small model shows remarkable performance during training and validation. It is observed that the accuracy during training and validation increases rapidly in the first 10-15 epochs, reaching values of about 0.94 and remaining constant with small fluctuations. At the same time, the stability observed on validation indicates that the model generalizes well and does not suffer from overfitting.

Another observation highlights the fact that a complex architecture is not always more performing and efficient, the analysis capacity and optimization being key points of these architectures, including for basic models and in relation to the proposed dataset. Following the evolution of the VIT Base model, a slight instability in the generalization capacity can be observed, but it notes a good learning capacity according to the evaluation results. The MaxVIT model achieves the best results and with a trend that suggests very efficient learning, observing accuracy and loss during training and validation.

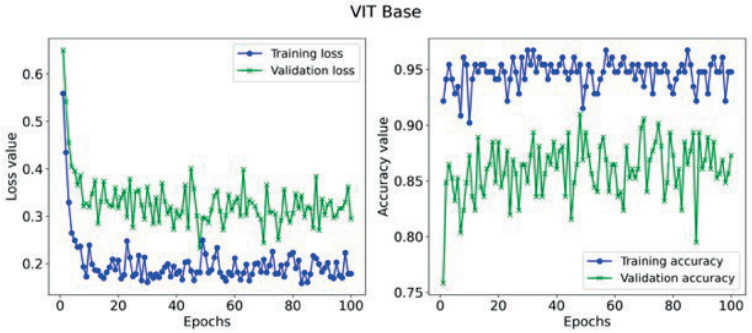

Figure 2. Training and Validation Metrics Over Epochs for VIT Base Model

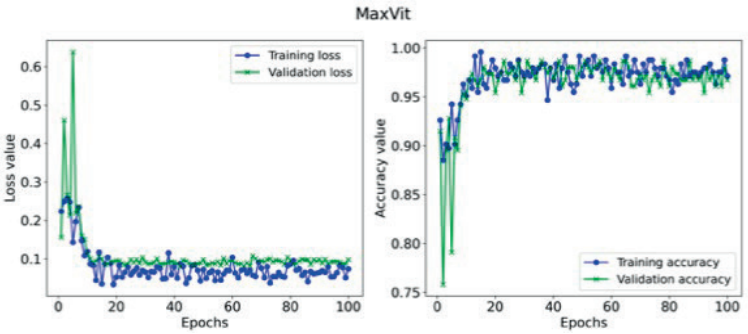

Figure 3. Training and Validation Metrics Over Epochs for MaxVit Model

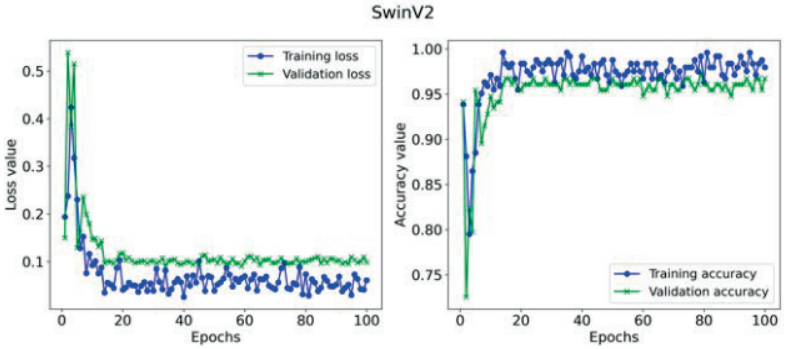

Figure 4. Training and Validation Metrics Over Epochs for SwinV2 Model

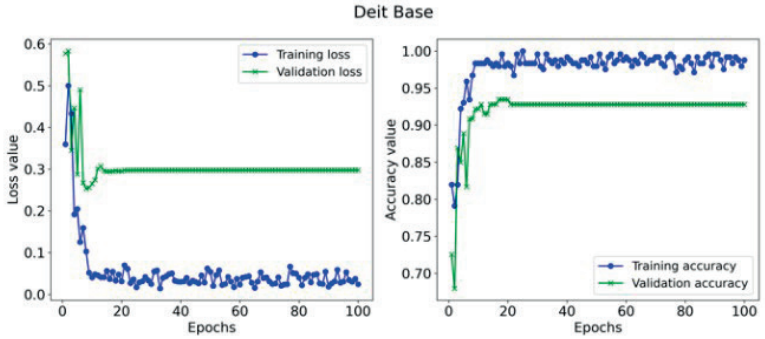

Figure 5. Training and Validation Metrics Over Epochs for DeiTBase Model

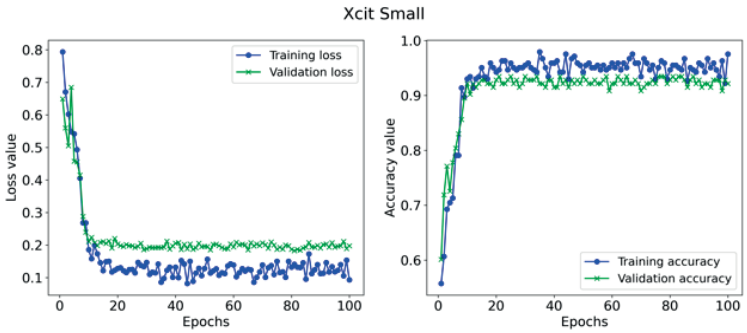

Figure 6. Training and Validation Metrics Over Epochs for XCiT Small Model

The training and validation metrics for DeiT Base suggest a possible complexity or under-training of the model, being more unstable than the other models and which may suggest room for improvement. The loss during training and evaluation decreases rapidly in the first epochs and stabilizes around 0.1-0.3, and the accuracy is lower compared to other analyzed models.

The SwinV2 model shows remarkable performance, both on the training and validation side, with low losses and high accuracy. It can be seen that the loss decreases rapidly in the first 10 epochs and stabilizes below the value of 0.1. The training and validation accuracies increase rapidly and suggest effective learning with very good values above 97%.

In the end, the analyzed models note a reasonable efficiency and performance, contributing to the development of modern solutions for automatic classification of waste, based on deep-learning techniques. According to the experiments and the comparative study, the MaxVIT Nano and SwinV2 models stand out for their high accuracy and stability. These metrics and features observed among top models may pave the way for implementations in real-time waste sorting applications, contributing to recycling and sustainability efforts.

CONCLUSIONS

Automated waste detection is essential for efficient waste management, sustainability, and a cleaner future. In the present study, modern and innovative solutions for this problem were presented, leveraging the characteristics of vision transformer architectures for automated image classification.

The results obtained in this study note the efficiency and performance of vision transformer, deep learning architectures for automatic waste classification. The method proposed in the present study was based on phases of training and evaluation of the implemented models in relation to transfer learning and fine-tuning techniques.

In this sense notable performances were observed, with waste classification accuracy values over 94%, even 97-98% for the top models, which can contribute to the implementation of high-performance and robust detection systems. The analyzed models included state-of-the-art architectures optimized for the classification task, the best and most stable models observed being MaxVIT, SwinV2, and XCiT. Notable results, but with small fluctuations in the metrics, were observed for the VIT Base and DeiT models, but which still provided notable accuracy values.

The practical development of these systems can significantly contribute to efficient recycling solutions and environmental protection. As part of future developments, recent vision transformer models can be implemented for waste classification and optimized accordingly. Furthermore, the best models can be developed as part of automated waste identification systems using various real-time and lightweight platforms.

REFERENCES

Agafitei, A., Pavel, V.L. (2021). Researches regarding the impact of waste landfill in Sibiu county, Romania on the environment and measures to reduce it. *Scientific Papers. Series E. Land Reclamation, Earth Observation & Surveying, Environmental Engineering, X*, 11-16, Print ISSN 2285-6064.

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., & Jégou, H. (2021). XCiT: Cross-covariance image transformers. *Advances in neural information processing systems, 34*, 20014-20027.

Alrayes, Fatma S., Mashael M. Asiri, Mashael S. Maashi, Mohamed K. Nour, Mohammed Rizwanullah, Azza Elneil Osman, Suhanda Drar & Abu Sarwar Zamani. (2023). Waste Classification Using Vision Transformer Based on Multilayer Hybrid Convolution Neural Network. *Urban Climate*, May, 101483. https://doi.org/10.1016/j.uclim.2023.101483.

An, Kang & Yanping Zhang. (2022). LPViT: A Transformer Based Model for PCB Image Classification and Defect Detection. IEEE Access, 42542–53. https://doi.org/10.1109/access.2022.3168861.

Boldeanu, G., Gheorghe, M., Moise, C., Dana Negula, I., Tudor, G. (2023). Earth observation techniques applied for land waste detection and monitoring. *Scientific Papers. Series E. Land Reclamation, Earth Observation & Surveying, Environmental Engineering, Vol. XII*, 383-388, Print ISSN 2285-6064.

Chen, Jinxiang, Yiqun Cheng & Jianxin Zhang. (2023). Anti-Local Occlusion Intelligent Classification Method Based on MobileNet for Hazardous Waste. *International Journal of Modelling, Identification and Control, 4*, 333–40. https://doi.org/10.1504/ijmic.2023.131203.

Dong, Zhiming, Junjie Chen, and Weisheng Lu. (2022). Computer Vision to Recognize Construction Waste Compositions: A Novel Boundary-Aware Transformer (BAT) Model. *Journal of Environmental Management*, 114405. https://doi.org/10.1016/j.jenvman.2021.114405.

Dookhee, Surajsingh. (2022). Domestic Solid Waste Classification Using Convolutional Neural Networks. 2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS), December. https://doi.org/10.1109/ipas55744.2022.10052971.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

Hu, Fan, Pengjiang Qian, Yizhang Jiang, and Jian Yao. (2022). An Improved Waste Detection and Classification Model Based on YOLOV5. In Intelligent Computing Methodologies, 741–54. Springer International Publishing. http://dx.doi.org/10.1007/978-3-031-13832-4_61.

Huang, Kai, Huan Lei, Zeyu Jiao & Zhenyu Zhong. (2021). Recycling Waste Classification Using Vision Transformer on Portable Device. *Sustainability, 21* 11572, https://doi.org/10.3390/su132111572.

Kumar, Avan, Bhavik R. Bakshi, Manojkumar Ramteke, & Hariprasad Kodamana. (2023). Recycle-BERT: Extracting Knowledge about Plastic Waste Recycling by Natural Language Processing. *ACS Sustainable Chemistry & Engineering, 32*, 12123–34. https://doi.org/10.1021/acssuschemeng.3c03162.

Kurz, Aidan, Ethan Adams, Arthur C. Depoian, Colleen P. Bailey & Parthasarathy Guturu. (2022). WMC-ViT: Waste Multi-Class Classification Using a Modified Vision Transformer. IEEE MetroCon, November. https://doi.org/10.1109/metrocon56047.2022.9971136.

Liu, Ze, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, et al. (2022). Swin Transformer V2: Scaling Up Capacity and Resolution. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June. https://doi.org/10.1109/cvpr52688.2022.01170.

Ma, Wanqi, Hong Chen, Wenkang Zhang, Han Huang, Jian Wu, Xu Peng & Qingqing Sun. (2024). DSYOLO-Trash: An Attention Mechanism-Integrated and Object Tracking Algorithm for Solid Waste Detection. *Waste Management,* 46–56. https://doi.org/10.1016/j.wasman.2024.02.014.

Mao, Wei-Lung, Wei-Chun Chen, Chien-Tsung Wang & Yu-Hao Lin. (2021). Recycling Waste Classification Using Optimized Convolutional Neural Network. *Resources, Conservation and Recycling*, 105132. https://doi.org/10.1016/j.resconrec.2020.105132.

Qin, Hai, Liye Shu, Li Zhou, Songyun Deng, Haihua Xiao, Wei Sun, Qiaokang Liang, Dan Zhang & Yaonan Wang. (2024). Active Learning-DETR: Cost-Effective Object Detection for Kitchen Waste. IEEE Transactions on Instrumentation and Measurement, 1–15. https://doi.org/10.1109/tim.2024.3368494.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International conference on machine learning, 10347-10357.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). Maxvit: Multi-axis vision transformer. In European conference on computer vision, 459-479. Cham: Springer Nature Switzerland.

Vîrsta, A., Sandu M.A., Daraban A.E. (2020). Dealing with the transition from in line economy to circular economy - public awareness investigation in Bucharest. *AgroLife Scientific Journal, 9*(1), 355–362.