Transformer-Based File Fragment Type Classification for File Carving in Digital Forensics

Andrey Guzhov and Christoph Tobias Wirth
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany

andrey.guzhov@dfki.de
tobias.wirth@dfki.de

Abstract: The recovery and reconstruction of fragmented data is a critical challenge in digital forensics, particularly when dealing with incomplete, corrupted, or partially deleted files in large-scale cybercrime investigations. Accurate classification of file fragment types is essential for reconstructing critical evidence, especially in environments characterized by high levels of data fragmentation, such as cyberattacks, data breaches, and the operation of illicit ("darknet") data centers. Traditional file carving methods often struggle to efficiently handle these fragmented files, limiting their reliability in complex investigations involving large volumes of data. This paper introduces a novel approach to classifying file fragment types using a Transformer-based model, designed to significantly enhance the speed and accuracy of forensic investigations. Unlike traditional methods, which rely on handcrafted rules or shallow machine learning techniques, our model leverages the powerful Swin Transformer V2 architecture, a state-of-the-art deep learning model tailored for sequence-to-sequence tasks. The model was trained to recognize complex, hierarchical patterns within raw byte sequences, enabling it to classify file fragments with high precision and reliability. We demonstrate that our model outperforms traditional methods on 512-byte file blocks, achieving superior classification accuracy on the File Fragment Type dataset (FFT-75), and also shows strong competitive performance with larger 4 KiB file blocks. Our approach represents a significant advancement in digital forensics, automating the classification of fragmented data and improving the reliability and efficiency of evidence recovery. Future work will focus on optimizing the model for different file block sizes and evaluating its application to real-world fragmented data scenarios. By automating the identification of file fragment formats, our approach not only improves classification accuracy but also reduces the time required for investigators to recover critical evidence from fragmented data sources. This work provides a promising tool for digital forensics practitioners, advancing recovery capabilities in the face of evolving cyber threats.

Keywords: Digital forensics, File carving, File fragment classification, Data fragmentation, Transformer models, Cybercrime investigations

1. Introduction

File carving is a critical technique in digital forensics used to recover fragmented data from files that have been damaged, deleted, or corrupted. The process involves identifying and reconstructing files and file fragments from raw data, which is often necessary when dealing with incomplete or partially overwritten files. Traditional file carving methods primarily rely on file signatures or heuristic rules to identify the structure and type of data fragments (Ali et al, 2018). These methods have been effective for certain types of file recovery but face limitations when dealing with large volumes of fragmented data or highly complex file structures (Pal and Memon, 2009).

In the recent years, machine learning approaches have been explored to enhance file carving techniques (Ramli et al, 2021). These methods aim to improve the accuracy and automation of fragment classification, potentially eliminating the need for manual feature engineering. However, despite promising results, current models typically rely on traditional machine learning (e.g., random forest (Breiman, 2001)) techniques or shallow neural networks (Liu et al, 2023; Mittal et al, 2020; Skračić et al, 2023), which may struggle to capture complex patterns within file fragments. As a result, these models can exhibit suboptimal performance in scenarios involving a high degree of fragmentation, where the relationships between fragments are not easily identified through traditional methods (Pal and Memon, 2009).

This paper aims to address these challenges by applying a Transformer-based deep learning model, specifically the Swin Transformer V2 (Liu et al, 2022a), to the task of file fragment type classification. The use of this model allows for the recognition of complex, hierarchical patterns within raw byte sequences, improving the accuracy of fragment classification, as shown in Section 5. Additionally, the approach seeks to provide a more efficient and scalable solution to file fragment identification, reducing the reliance on predefined rules or manual intervention. By evaluating the model on the File Fragment Type dataset (FFT-75) published by Mittal et al (2020), we demonstrate that it outperforms other models for shorter fragments and offers competitive accuracy with larger file block sizes.

The remainder of the paper is structured as follows: Section 2 provides a review of related work in the field of file carving and machine learning approaches to file fragment classification. Section 3 details the Transformer-based model used in this paper, including its architecture and key modifications. Section 4 describes the dataset and the training process employed in the experiments. Section 5 presents the results and a comparison with baseline models. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

Digital forensics encompasses a range of techniques aimed at recovering, preserving, and analyzing data from digital devices to support legal proceedings (Casino et al, 2022). A critical component of this field is file carving, which involves reconstructing files from fragmented data without relying on file system metadata (Pal and Memon, 2009). File carving is particularly helpful when dealing with corrupted, deleted, or partially overwritten files, as it enables the recovery of data that might otherwise be inaccessible. Traditional file carving techniques often depend on predefined file signatures and heuristics to identify and reconstruct file fragments. While effective in certain scenarios, these methods can be limited when handling large volumes of fragmented data or complex file structures (Ramli et al, 2021).

The exponential growth of digital data (Taylor, 2024) presents a significant challenge that potentially impacts the effectiveness of classical file carving methods. As data volumes increase, the complexity and fragmentation of files also rise, making it more challenging to accurately and efficiently reconstruct files using traditional techniques. This surge in data necessitates the development of more advanced methods capable of handling large-scale data with greater precision. The digital forensics research community has been explored a wide range of approaches to address these challenges (Li et al, 2021), which resulted in recent advances in file carving, as discussed in Section 2.2, thus offering the potential to alleviate the constraints imposed by the sheer volume of data and thereby enabling more effective analysis and recovery of fragmented files.

2.1 Digital Forensics

Recent advancements in digital forensics have been significantly influenced by the integration of artificial intelligence (AI) and machine learning (ML) technologies (Ademu et al, 2011). These innovations aim to enhance the efficiency and accuracy of forensic investigations by automating complex tasks and enabling the analysis of large datasets. AI-powered tools can process vast amounts of data (Qiu et al, 2016), identifying patterns and anomalies that might be overlooked by human analysts. This capability is particularly beneficial considering increasing data volumes and the growing sophistication of cyber threats (Balantrapu, 2024).

In the area of file carving, which involves recovering files without relying on file system metadata, recent research has focused on developing more advanced techniques employing deep learning models (Liu et al, 2023; Mittal et al, 2020; Skračić et al, 2023). Studies have highlighted the need for effective methodologies to retrieve data from fragmented files (Ramli et al, 2021), emphasizing the importance of developing and validating carving techniques and tools.

Unlike conventional recovery techniques that rely on file system metadata, file carving works by scanning raw disk images and identifying data fragments based on their content, structure, and known signatures (Pal and Memon, 2009). This method is particularly useful when file system information is not available, as is often the case when files have been deleted or corrupted. Over the years, file carving techniques have been evolving, although recovering highly fragmented files still remains a hard challenge nowadays (Ramli et al, 2021).

Recent research has explored the limitations of traditional carving tools, highlighting issues such as false positives (Pahade et al, 2015), incomplete file recovery (Ali et al, 2018), and the challenge of handling fragmented files that span across non-contiguous blocks of data (Pal and Memon, 2009). In response, several approaches have been proposed that incorporate machine learning techniques to refine the identification and reconstruction process (Ramli et al, 2021). These efforts aim to enhance the automation of file carving, reduce reliance on predefined rules, and better handle the growing volume and complexity of digital evidence. Despite these advancements, the field still faces challenges in achieving reliable results across a diverse spectrum of file types and fragmentation patterns.

2.2 Deep Learning in File Carving

In the past years, deep learning has emerged as a powerful tool for enhancing file carving techniques, aiming to improve the accuracy and efficiency of fragment classification (Mittal et al, 2020). Particularly, convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has been explored to automate the process of fragment classification and recovery (Liu et al, 2023; Mittal et al, 2020; Skračić et al, 2023; Zhu et al, 2023). These models are capable of learning complex patterns and features within data, enabling them to classify file fragments without relying on manual feature extraction or explicit signature-based approaches. Several studies have applied deep learning techniques to file carving, with encouraging results in improving classification accuracy for common file types and fragment sizes, as discussed in Section 5.1.

One of the key advantages of deep learning in file carving is its ability to recognize more abstract relationships in raw data. For instance, CNNs have been applied to treat file fragments as images (Liu et al, 2023) or sequences (Skračić et al, 2023), where the network can identify hierarchical patterns between byte sequences. Other approaches have utilized long short-term memory (LSTM) networks (Zhu et al, 2023) to capture causal dependencies between adjacent bytes within file fragments. The latter can be particularly useful when dealing with fragmented files that span across non-contiguous disk blocks, although it requires new datasets providing extended metadata. The aforementioned approaches allow for more robust and adaptable carving methods that are less dependent on predefined rules or file signatures, thus advancing research on file carving.

Despite the progress made in applying deep learning to file carving, several challenges remain. One of the primary difficulties is the lack of large, annotated datasets that model real-world data for training deep learning models. Many of the existing datasets are relatively small or limited to specific types of file fragments (Mittal et al, 2020), which makes it challenging to build models that can generalize across a wide range of forensic cases. Additionally, while deep learning models show promise in improving accuracy, their computational cost remains a concern, particularly when applied to large-scale forensic investigations. As a result, some studies have focused on optimizing network architectures and training strategies to reduce computational overhead while maintaining performance (Felemban et al, 2024; Mittal et al, 2020; Saaim et al, 2022; Skračić et al, 2023). To summarize, while deep learning has proven to be a valuable addition to the field of file carving, further research is needed to address these limitations and improve the scalability of these approaches in real-world applications.

3. Model

Designing a model for file fragment classification requires balancing computational efficiency, scalability, and the ability to capture complex patterns within raw binary data. The task defined by the FFT-75 dataset—classifying file types based on fixed-size file fragments—requires dealing with randomly fragmented data grouped by file formats into up to 75 classes, with no metadata present. To address these challenges, a Transformer-based (Vaswani, 2017) architecture was chosen for its ability to model hierarchical patterns and long-range dependencies in sequential data (Wen et al, 2022).

The Swin Transformer V2 architecture was identified as a strong candidate due to its modularity, scalability, and efficient computation by using shifted window-based attention (Liu et al, 2022a). Originally designed for visual tasks, its hierarchical structure makes it adaptable to the sequential nature of file fragment data when appropriately modified. The decision to base the model on Swin Transformer V2 reflects these strengths, while specific adaptations, as discussed in the Section 3.2, were made to tailor the architecture to the non-visual, integer input format and task-specific requirements of the FFT-75 dataset.

3.1 Swin Transformer V2

The Swin Transformer V2 is an evolution of the Swin Transformer architecture (Liu et al, 2021), a hierarchical vision transformer initially designed for visual recognition tasks. This architecture uses shifted window-based attention to balance computational efficiency and modeling capability, making it suitable for high-resolution input data and improving upon its predecessor by introducing scaled cosine attention mechanism, logarithmic positional encoding, which contributes to more stable training and improved performance on complex tasks (Liu et al, 2022a). At the same time, Swin Transformer V2 is scalable to moderate models offering competitive performance in downstream tasks.

The aforementioned advancements make the architecture capable of handling diverse tasks beyond image classification, including semantic segmentation (He et al, 2022) and video action classification (Liu et al, 2022b), suggesting applicability of the architecture to the file fragment classification task. The hierarchical structure of the Swin Transformer V2 allows it to capture both local and global relationships effectively (Kim et al, 2024), offering flexibility in processing inputs of varying resolutions and sizes.

The architecture exists in several variants, as shown in Table 1, each tailored for different computational budgets and performance requirements. The "Tiny" (Swin-T), "Small" (Swin-S), "Base" (Swin-B), and “Large”

(Swin-L) models differ in the number of parameters, computational complexity, and layer depth. In this paper, we used Swin Transformer V2 (Tiny) pre-trained on the ImageNet1k (Fei-Fei et al, 2009) dataset, as it offers a competitive performance for a moderate computational cost.

Table 1: Comparison of variants of Swin Transformer V2 (Liu et al, 2022a)

Variant of the architecture	Accuracy on ImageNet1k (%)	Computational cost (GFLOPs)
Swin-T	81.80	5.90
Swin-S	83.70	11.50
Swin-B	84.20	20.30
Swin-L	86.90 (ImageNet21k (Ridnik et al, 2021))	47.50

3.2 CarveFormer

To adapt the Swin Transformer V2 architecture for non-visual inputs, specifically for the classification of file fragments, several modifications were implemented, resulting in the model referred to as CarveFormer (shown in Figure 1). The original Swin Transformer V2 is designed for image processing tasks and begins with a 2D-convolutional (LeCun et al, 1998) layer that maps three-channel two-dimensional inputs (RGB images) into 96-channel outputs, followed by a LayerNorm (Ba, 2016) normalization layer. In CarveFormer, instead, the first convolutional layer and the subsequent normalization layer were replaced with an embedding layer of dimension 96. This embedding layer transforms the sequential byte data of file fragments into a suitable representation for the transformer model. The output from the embedding layer is then reshaped to mimic a 2D structure, aligning with the input requirements of the Swin Transformer V2 architecture. This modification allows the model to process non-visual data natively by leveraging powerful Swin Transformer V2 model.

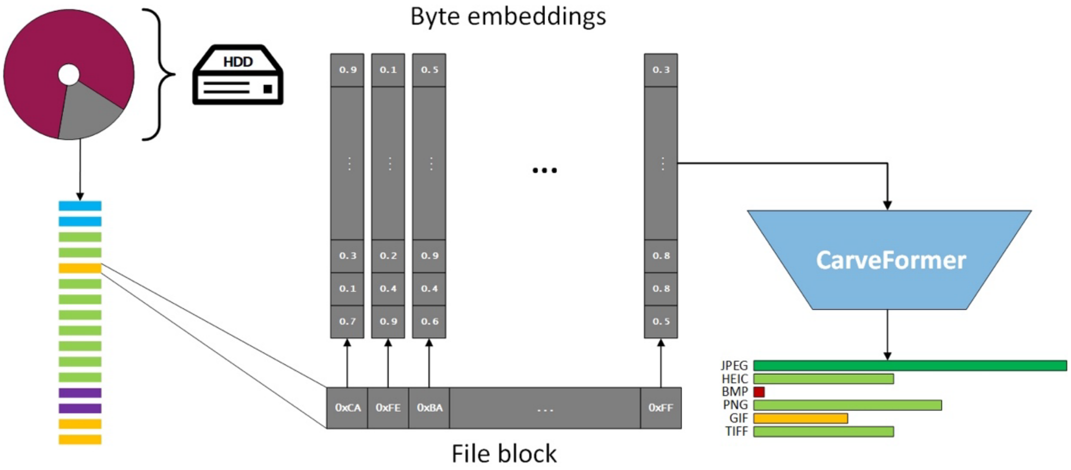

Figure 1: Overview of the CarveFormer model handling file fragments at the level of individual file blocks

CarveFormer was built upon a snapshot of Swin Transformer V2 (Tiny) pre-trained on the ImageNet1k dataset. This choice was driven by the absence of substantially large-scale datasets specifically tailored to file carving tasks (i.e. exceeding FFT-75), which might become an issue when training a Transformer from scratch (Steiner et al, 2021). Pre-training on ImageNet1k provides the model with a good weight initialization, as it learns to recognize patterns and hierarchical structures that, while originally intended for visual data, can still be adapted effectively to other types of input such as videos (Liu et al, 2022b) or audio (Guzhov et al, 2021).

4. Experimental Setup

The experiments aimed at evaluating the effectiveness of the CarveFormer model in classifying file fragments obtained from the FFT-75 dataset, which encompasses a diverse range of file formats and two file block sizes: 512 and 4096 bytes. The test subset of the FFT-75 dataset served as a basis for benchmarking of CarveFormer, those training and optimization of hyper-parameters were performed on respective subsets of FFT-75.

4.1 Dataset

In this paper, we utilized the FFT-75 dataset, a benchmark specifically designed for evaluating file fragment classification models. FFT-75 contains labeled file fragments sampled from 75 diverse file types, spanning formats commonly encountered in digital forensics, including text, image, audio, video, and archive files. These fragments are extracted at random offsets within files and are divided into two categories based on their sizes: 512 bytes and 4096 bytes and organized into six scenarios, as shown in Table 2. Within each scenario, the number of samples is balanced across file formats. This design aims to address a range of real-world applications where file fragments can vary in size and distribution of formats.

Table 2: Description of the FFT-75 dataset (scenarios)

Scenario	# of classes	# of samples	Description
#1	75	7500k	All file formats
#2	11	1935k	Common file formats
#3	25	2300k	Image and video formats
#4	5	1054k	Image formats
#5	2	1036k	JPEG or any other format
#6	2	1000k	JPEG or another image format

While FFT-75 provides a good starting point for assessing classification models, it does have certain limitations. The dataset is constructed with randomly sampled file fragments, which do not capture the sequential fragmentation patterns typically observed in real-world forensic cases, such as those caused by file corruption or deletion processes. This random sampling simplifies the task compared to the complexities of real-world data, where context and order often play critical roles (Garfinkel, 2007). These limitations highlight the need for future datasets that more accurately reflect the challenges of practical file carving scenarios.

4.2 Model Training

The CarveFormer model was trained on the FFT-75 dataset for all combinations of scenarios and file block sizes, thus resulting in 12 benchmarking experiments. Model weights were initialized from a Swin Transformer V2 (Tiny) pre-trained from scratch on ImageNet1k to obtain a good initialization before fine-tuning. This initialization strategy allowed the model to benefit from robust feature extraction capabilities developed during pre-training, even though the domain of the input data differed, as discussed in Section 3.2.

The training procedure followed hyper-parameter settings similar to those recommended for Swin Transformer V2 pre-training, ensuring consistency and compatibility with the architecture. The effective batch size was set to 1024, and the AdamW optimizer (Loshchilov, 2017) was used with a learning rate of $3.75 \cdot 10^{-4}$ and a weight decay of 0.05. Training was conducted for 50 epochs, ensuring the model adapts to the FFT-75 dataset and underlying distribution of data. In the test phase, the best model snapshot was chosen according to the validation accuracy measured after every epoch.

5. Results

The evaluation of CarveFormer on the FFT-75 dataset provided insights into its capabilities and limitations in file fragment classification. The results demonstrated that CarveFormer is effective in identifying file format-specific patterns, showcasing competitive performance when compared to existing methods.

While the model performed well in many aspects, the experiments also suggested an existence of challenges inherent to the dataset itself and the way a Transformer-based model can be applied to the task. Certain limitations stem from the design of FFT-75, which may not fully reflect real-world fragmentation patterns. These challenges, along with observations regarding intrinsic ambiguities in classifying some file formats, highlight the importance of addressing both model and dataset limitations for future work in this area, as discussed in a more detail in Section 6.

5.1 Model Performance

The performance of CarveFormer was evaluated on the FFT-75 dataset to assess its ability to classify file fragments accurately across two block sizes: 512 bytes and 4096 bytes, as presented in Table 3 and Table 4 respectively. Accuracy was used to compare our and the baseline models in the multiclass classification task.

At the high level, the obtained results allow for identifying two groups of scenarios, irrespective to the block size. Scenarios #1-4 represent more challenging multiclass classification tasks. In contrast, all models demonstrate a nearly perfect performance in scenarios #5 and #6 , suggesting that the latter scenarios can be considered as solved. Also, with the block size increasing, the difficulty level decreasing, as larger 4096-byte blocks provide the model with more format-specific information, making the task easier. Specifically, when going from 512-byte to 4096-byte blocks, the accuracy gradually increases for each model in the comparison.

Table 3 presents results for 512-byte blocks. Here, the proposed CarveFormer model outperforms other methods in the most challenging scenarios #1 and #2, achieving accuracy of 72.10% and 90.62% respectively. CarveFormer demonstrates second best performance in scenarios #3-5 (93.44%, 92.63%, and 99.04%), with a slight decrease in scenario #6 (98.93%).

Table 3: Performance comparison on FFT-75 (512-byte blocks), accuracy (%)

Model	Reference	Scenario					
		#1	#2	#3	#4	#5	#6
FiFTy	Mittal et al, 2020	65.60	78.90	87.90	90.20	99.00	99.30
DSC	Saaim et al, 2022	65.89	75.84	80.79	87.14	98.94	98.76
ResNet-18	Liu et al, 2023	71.00	90.40	**93.50**	**93.60**	**99.20**	99.20
CNN-LSTM	Zhu et al, 2023	66.50	–	–	–	–	–
ByteRCNN	Skračić et al, 2023	71.10	87.50	91.00	92.00	99.00	**99.50**
DSC-SE	Felemban et al, 2024	66.33	74.99	80.79	87.32	98.96	98.65
CarveFormer	Ours	**72.10**	**90.62**	93.44	92.63	99.04	98.93

Similarly, the proposed model performs competitively on 4096-byte blocks, achieving new state-of-the-art accuracy of 96.87% in scenario #3 and demonstrating second best accuracy in scenarios #1 (82.99%), #2 (93.96%), #4 (96.91%), and #5 (99.30%).

Table 4: Performance comparison on FFT-75 (4096-byte blocks), accuracy (%)

Model	Reference	Scenario					
		#1	#2	#3	#4	#5	#6
FiFTy	Mittal et al, 2020	77.50	89.80	94.60	94.10	99.20	99.60
DSC	Saaim et al, 2022	78.45	85.70	93.06	94.17	99.28	99.59
ResNet-18	Liu et al, 2023	82.10	**94.20**	96.80	**96.10**	99.30	99.40
CNN-LSTM	Zhu et al, 2023	78.60	–	–	–	–	–
ByteRCNN	Skračić et al, 2023	**83.90**	93.10	96.50	95.40	99.30	99.50
DSC-SE	Felemban et al, 2024	79.27	87.10	93.32	94.61	**99.37**	**99.69**
CarveFormer	Ours	82.99	93.96	**96.87**	95.91	99.30	99.42

5.2 Limitations

The quadratic computational complexity inherent in the Transformer architecture challenges scalability of the proposed CarveFormer model. File blocks are currently processed individually by converting them into sequences using an embedding layer, where 512-byte and 4096-byte blocks are treated as sequences of length 512 and 4096, respectively. This design is in alignment with recently published approaches (Felemban et al, 2024; Skračić et al, 2023) and ensures compatibility with the base Swin Transformer V2 model but significantly increases computational costs when the block size increases. The high memory and processing requirements restrict scalability, limiting the feasibility of extending the model to handle larger file fragments (above 8 KiB).

Another limitation arises from the design of the FFT-75 dataset. While the dataset provides a diverse collection of randomly sampled labeled file blocks, it does not capture the sequential patterns of fragmentation observed in real-world scenarios (Garfinkel, 2007). This absence of realistic fragmentation patterns prevents CarveFormer from fully leveraging the Transformer's ability to model long-range dependencies within sequences. As a result, the embedding layer is necessary to prepare the data for processing, introducing an additional transformation that increases computational costs.

Additionally, the comparative analysis of CarveFormer and competing models reveals an upper boundary on achievable accuracy, rooted in the fundamental characteristics of file fragments, as can be seen in Table 3 and Table 4. Specifically, in the most challenging scenario #1, the best performing models seem to asymptotically achieve accuracy of ~73% for 512-byte blocks and ~84% for 4096-byte blocks. This can be attributed to the observation that it is often impossible to distinguish between fragments of plain file formats, such as texts or images, and those embedded into a container format, such as PDF. Disregarding this observation during labeling of the FFT-75 data (Mittal et al, 2020) introduced an inherent ambiguity that potentially confuses models in the absence of contextual information. Such challenges highlight intrinsic constraints of the FFT-75 dataset, irrespective of the model employed.

6. Conclusion

This work presented CarveFormer, a Transformer-based model designed for file fragment classification, a critical task in digital forensics. By adapting the Swin Transformer V2 architecture to non-visual data, we developed a model capable of processing raw file fragments and classifying them with competitive accuracy across two file block sizes: 512 bytes and 4096 bytes. The model's performance on 512-byte file blocks in scenario #1 and #2 demonstrated its ability to outperform current carving methods, setting a new state-of-the-art performance baseline in this task.

CarveFormer effectively learns patterns corresponding to raw file fragments at block level, leveraging the Transformer's potential to model complex relationships within the data. However, inherent constraints, such as the quadratic complexity of the Transformer architecture and ambiguities in dataset labeling for certain file types, highlight the importance of exploring alternative strategies to enhance scalability and further align dataset characteristics with real-world forensic scenarios.

Future research needs to focus on enhancing both the model and the data. One key direction is to establish an improved alternative to the FFT-75 dataset. This involves considering container file formats for assigning appropriate class labels. Additionally, the goal is to better follow realistic fragmentation patterns commonly encountered in practical forensic cases (Garfinkel, 2007). These improvements aim at reducing the computational overhead when applying sequence-to-sequence–primarily, Transformer-based–models. While challenges remain, the results affirm the potential of Transformer-based models in digital forensics, and we are optimistic that continued exploration in this direction will yield even more robust and versatile solutions.

Acknowledgements

This work was funded by the BMBF project Carve-DL (grant 13N16405).

References

Ademu, I. O., Imafidon, C. O. & Preston, D. S., 2011. A new approach of digital forensic model for digital forensic investigation. IJACSA: International Journal of Advanced Computer Science and Applications, 2(12).

Ali, R. R., Mohamad, K. M., Jamel, S. A. P. I. E. E. & Khalid, S. K. A., 2018. A review of digital forensics methods for JPEG file carving. J. Theor. Appl. Inf. Technol., 96(17), pp.5841-5856.

Ba, J. L., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Balantrapu, S. S., 2024. Current trends and future directions exploring machine learning techniques for cyber threat detection. International Journal of Sustainable Development Through AI, ML and IoT, 3(2), pp.1-15.

Breiman, L., 2001. Random forests. Machine Learning, 45, pp. 5-32

Casino, F., Dasaklis, T. K., Spathoulas, G. P., Anagnostopoulos, M., Ghosal, A., Borocz, I. et al., 2022. Research trends, challenges, and emerging topics in digital forensics: A review of reviews. IEEE Access, 10, pp.25464-25493.

Fei-Fei, L., Deng, J. & Li, K., 2009. ImageNet: Constructing a large-scale image database. Journal of Vision, 9(8), p.1037.

Felemban, M., Ghaleb, M., Saaim, K., Al-Saleh, S. & Almulhem, A., 2024. File Fragment Type Classification using Light-Weight Convolutional Neural Networks. IEEE Access.

Garfinkel, S. L., 2007. Carving contiguous and fragmented files with fast object validation. Digital Investigation, 4, pp.2-12.

Guzhov, A., Raue, F., Hees, J. & Dengel, A., 2021. ESResNet: Environmental sound classification based on visual domain models. In 2020 25th International Conference on Pattern Recognition (ICPR), pp.4933-4940.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R. & Xue, Y., 2022. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60, pp.1-15.

Kim, J. H., Kim, N. & Won, C. S., 2024. Global–local feature learning for fine-grained food classification based on Swin Transformer. Engineering Applications of Artificial Intelligence, 133, p.108248.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G. et al., 2021. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. Mobile Networks and Applications, 26, pp.234-252.

Liu, W., Wang, Y., Wu, K., Yap, K. H. & Chau, L. P., 2023. A Byte Sequence is Worth an Image: CNN for File Fragment Classification Using Bit Shift and n-Gram Embeddings. In 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), pp.1-5.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y. et al., 2022a. Swin Transformer V2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12009-12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. et al., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.10012-10022.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. & Hu, H., 2022b. Video Swin Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3202-3211.

Loshchilov, I., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Mittal, G., Korus, P. & Memon, N., 2020. FiFTy: large-scale file fragment type identification using convolutional neural networks. IEEE Transactions on Information Forensics and Security, 16, pp.28-41.

Pahade, R. K., Singh, B. & Singh, U., 2015. A survey on multimedia file carving. International Journal of Computer Science & Engineering Survey, 6(6).

Pal, A. & Memon, N., 2009. The evolution of file carving. IEEE Signal Processing Magazine, 26(2), pp.59-71.

Qiu, J., Wu, Q., Ding, G., Xu, Y. & Feng, S., 2016. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016, pp.1-16.

Ramli, N. I. S., Hisham, S. I. & Badshah, G., 2021. Analysis of file carving approaches: A literature review. In Advances in Cyber Security: Third International Conference, ACeS 2021, Penang, Malaysia, August 24–25, 2021, Revised Selected Papers 3, pp.277-287.

Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L., 2021. Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972.

Saaim, K. M., Felemban, M., Alsaleh, S. & Almulhem, A., 2022. Light-weight file fragments classification using depthwise separable convolutions. In IFIP International Conference on ICT Systems Security and Privacy Protection, pp.196-211.

Skračić, K., Petrović, J. & Pale, P., 2023. ByteRCNN: Enhancing File Fragment Type Identification with Recurrent and Convolutional Neural Networks. IEEE Access.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. & Beyer, L., 2021. How to train your ViT? Data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270.

Taylor, P., 2024. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028. [Online] Available at: https://www.statista.com/statistics/871513/worldwide-data-created/

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. & Sun, L., 2022. Transformers in time series: A survey. arXiv preprint arXiv:2202.07125.

Zhu, N., Liu, Y., Wang, K. & Ma, C., 2023. File Fragment Type Identification Based on CNN and LSTM. In Proceedings of the 2023 7th International Conference on Digital Signal Processing, pp.16-22.