Enhancing AI-Generated Image Detection with a Novel Approach and Comparative Analysis

Stuart Weir*, Muhammad Shahbaz Khan*, Naghmeh Moradpoor (SMIEEE)*, and Jawad Ahmad†
40479348@live.napier.ac.uk, {muhammadshahbaz.khan, n.moradpoor}@napier.ac.uk, Jahmad@pmu.edu.sa
*School of Computing, Engineering and the Built Environment
Edinburgh Napier University, Edinburgh, UK
†Cybersecurity Center
Prince Mohammad Bin Fahd University, Alkhobar, Saudi Arabia

***Abstract*—This study explores advancements in AI-generated image detection, emphasizing the increasing realism of images, including deepfakes, and the need for effective detection methods. Traditional Convolutional Neural Networks (CNNs) have shown success but face limitations in generalization and accuracy, particularly with newer technologies like Diffusion Models. With the evolution of AI image generation models, from CNNs to Generative Adversarial Networks (GANs) and Diffusion Models, detecting synthetic images has become more challenging. Issues include dataset diversity, adversarial attacks, and inconsistencies in pre-processing methods. While state-of-the-art models like CNNs, Vision Transformers (ViTs), and hybrid approaches exist, their accuracy in detecting increasingly sophisticated fake images remains suboptimal. This research proposes a novel hybrid detection model combining CNNs and ViTs with an additional attention mechanism layer. This structure aims to improve the interaction between local and global features, enhancing detection accuracy. The model was trained using the CIFAKE dataset, which contains 120,000 real and AI-generated images. The added attention mechanism enhances feature extraction, addressing limitations in existing models when faced with next-generation synthetic images. The hybrid CNN/ViT+Attention model demonstrated improved detection accuracy, achieving 99.77%, surpassing previous methods. This research lays a foundation for stronger AI-generated image detection, helping to mitigate the risks of synthetic image fraud.**

***Index Terms*—Vision Transformer, Convolutional Neural Networks, Hybrid models, attention mechanism, CIFAKE dataset**

I. Introduction

Image manipulation has been around since the advent of photography. In recent years, advancements in technology have facilitated the creation of powerful editing tools like Photoshop and GIMP [1]. Consequently, research in multimedia forensics has been ongoing for almost 20 years, with growing interest from academia, IT companies, and funding agencies.

With the rise of deep learning and AI, the production of synthetically generated image content has increased exponentially. This technology gave rise to the 'deepfake,' where authentic images were manipulated to appear as if they were something or someone else. This development was primarily driven by the use of Generative Adversarial Networks (GANs) [1] & [2], which produced highly realistic images. As these systems became more sophisticated, the generated images became increasingly difficult to detect. Research has focused on combating the widespread misuse of such images, which poses significant threats to privacy, democracy, and potentially national security.

These networks not only produce 'deepfake' images and videos but also have applications across many industries, including entertainment and media (for special effects), marketing and advertising (for promotional content), and academia (for generating graphical visualizations in research materials). However, the latter is becoming an increasingly fertile area for the misuse of this technology, leading to image fraud in scientific publications [3].

The evolution of AI-generated image technology has progressed further with the shift from GANs to more advanced models, such as diffusion models. These models now allow even inexperienced users to generate photorealistic images from text prompts [4]. This development could significantly increase the potential for spreading misinformation, as the high fidelity of these images makes them indistinguishable from photographs to the naked eye.

Detection methods for AI-generated images have also evolved, moving from reliance on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to more recent Vision Transformer (ViT) models, which incorporate attention mechanisms [5]. ViT models can achieve high accuracy with larger datasets, while traditional CNN models tend to perform better with smaller, controlled datasets. Hybrid models combining CNN, RNN, and ViT architectures have also made advancements in detecting generated images [6], [7] & [8].

One limitation in detection methods lies in their ability to generalize across different image generation models, as each model produces specific artifacts that aid detection. Additionally, image compression methods, such as JPEG, can affect the detectability of generated images.

In this paper, we developed a hybrid CNN/ViT model with an additional attention mechanism, as suggested by [5], to enhance the accuracy of AI-generated image detection.

Therefore, there is one research question we aim to answer through this study:

- Is a hybrid CNN/ViT model capable of enhancing AI-

generated image detection accuracy when integrated with an additional attention mechanism?

The remainder of this paper is structured as follows: Section II reviews related work, Section III discusses the methodology and design, Section IV presents the implementation, Section V covers the results and evaluation, and Section VI concludes the paper with future work directions.

II. Literature Review

In this section, we review some of the related work in the field, categorizing it into three groups: CNN models, ViT and Attention Models, and Hybrid Models, as follows.

A. CNN models

In [9], the authors introduce an innovative technique for identifying images produced by diffusion models. The researchers have developed a spectral analysis-based method that detects subtle frequency artifacts resulting from the diffusion process. By using a cross-difference high-pass filter alongside Fourier transform analysis, the technique highlights these artifacts and trains a classifier to differentiate between authentic and fabricated images. This approach proves resilient against mild JPEG compression and generalizes well to previously unseen diffusion models. To evaluate their method, the authors compiled a dataset of synthetic images from multiple diffusion models and compared them with genuine images. The model yielded impressive results. However, it showed slightly lower accuracy with certain models, such as DALL·E 2 and challenges remain in reducing false positives and accommodating a broader range of models.

The authors in [10] developed an enhanced CNN model aimed at identifying counterfeit images. They initially conducted a comparative study of six conventional machine learning algorithms but found that these models produced unsatisfactory results. Consequently, they deployed six well-known CNN models, such as ResNet50, VGG16, and MobileNetV2, with ResNet50 delivering the highest performance. To further boost accuracy, the authors employed various preprocessing techniques, including data augmentation, adaptive learning rates, model checkpointing, and dropout layers. These improvements mitigated overfitting and significantly increased ResNet50's accuracy. The study emphasized the effectiveness of advanced CNN techniques in detecting sophisticated counterfeit images and proposed future applications, such as mobile deployment.

In [11], the authors address the challenge of distinguishing between AI-generated and authentic images by creating a unique dataset and classification method. They developed the CIFAKE dataset, which contains 120,000 images. The researchers utilized a CNN to classify these images as either real or AI-generated, achieving a good classification accuracy. To explain the model's decisions, the authors employed Gradient Class Activation Mapping (Grad-CAM), which revealed that the CNN focused on minor flaws in the backgrounds of synthetic images, rather than the primary subject, for its classification.

B. ViT and Attention Models

The researchers in [12] investigated the application of ViT models to enhance the detection of deepfake images, an escalating concern in cybersecurity and media integrity. They employed a fine-tuned ViT model pre-trained on the ImageNet-21k dataset and evaluated it using a well-balanced dataset of 100,000 images, evenly split between genuine and GAN-generated deepfakes. By leveraging the ViT's ability to capture both localized and global image features through self-attention mechanisms, the researchers demonstrated the model's exceptional performance in detecting deepfakes across various datasets. In multiple experiments, the ViT consistently outperformed existing deepfake detection techniques. It provides a strong example of a ViT-based detection model that delivers high-accuracy results compared to existing methods. However, the focus is on deepfake images, where genuine images are manipulated to appear as something or someone else, meaning the model may miss entirely generated fake images.

The researchers in [13] introduced an innovative method to enhance CNN-based image classification by integrating Discrete Wavelet Transform (DWT) with attention mechanisms. They developed a Wavelet-Attention (WA) block that divides feature maps into low- and high-frequency components, applying attention exclusively to the high-frequency parts to capture intricate details while reducing noise, thereby preserving essential structural features in the low-frequency range. This WA block was incorporated into a newly developed architecture called Wavelet-Attention CNN (WA-CNN). The authors evaluated WA-CNN using the CIFAR-10 and CIFAR-100 datasets, showing significant improvements in classification accuracy. In comparison to other attention models like GCNet, SE-Net, and CBAM, WA-CNN demonstrated competitive or even superior performance, particularly in larger networks. It illustrates the effectiveness of incorporating an attention mechanism alongside DWT in a CNN model to boost accuracy. However, this implementation was only tested on two specific datasets, and its generalization capabilities remain unclear.

C. Hybrid Models

In the study by researchers in [14], the authors address the challenge of identifying increasingly sophisticated deepfake videos by introducing a hybrid model named 'HCiT', which combines a CNN with a ViT. Their aim is to overcome the limitations of current deepfake detection methods, which often struggle to generalize to new types of fake videos. To achieve this, they use the Xception CNN model to extract local features from cropped facial images and then input these features into a ViT, which captures global dependencies through self-attention mechanisms. The model achieved impressive results. HCiT also demonstrated strong generalization across different deepfake manipulation techniques. An ablation study—removing certain parts of the network to better understand its behavior—confirmed that the hybrid

CNN-ViT model outperformed both individual CNN and ViT models. It highlights how hybrid CNN/ViT models can be used to capture both local and global dependencies to enhance the detection of fake images. While focused on deepfake detection, it also demonstrates the effectiveness of attention mechanisms in image classification tasks.

In [15], the researchers introduce two innovative models aimed at improving the detection of deepfake images. The first model combines a ViT with a Convolutional Autoencoder (CAE), where the CAE is trained on authentic images to reconstruct them, and the ViT is used to classify both real and deepfake images. The second model utilizes the encoded features produced by the CAE and applies traditional machine learning techniques such as Support Vector Machines (SVM) and Logistic Regression for classification. The authors used the OpenForensics dataset, which contains over 115,000 real and fabricated images, to train and test their models. Overall, the study showcases the potential of combining different models to enhance the effectiveness of deepfake detection. It demonstrates the value of experimenting with various deep learning model combinations to improve accuracy and generalization in deepfake detection.

The integration of CNNs and ViTs in hybrid models, as discussed by [16], provides a crucial framework for balancing local feature extraction with global context understanding and is the closest research to this paper. However, while their model offers a solid foundation, it may not fully capture the interaction between these features, especially in scenarios requiring the detection of fine-grained image details. Our work builds on this foundation by enhancing the interaction between local and global features, which is critical for sophisticated image analysis. The authors in [20] emphasize the importance of selectively enhancing features before classification, a strategy we adopt by introducing an attention layer that refines the feature selection process. The attention mechanisms introduced by [17] are a key component of our model, allowing us to dynamically prioritize the most critical features to improve accuracy and generalization. Additionally, the robustness to adversarial examples noted by [18] and [19] aligns with our objective of making the model more resilient to adversarial noise, ensuring that only the most relevant features influence the final decision. By addressing these identified gaps, our proposed enhancements aim to advance current hybrid CNN-ViT models, particularly in complex tasks requiring detailed image classification and robust detection of manipulated images.

III. Methodology & Design

Based on research into the effectiveness and accuracy of hybrid detection models that utilize attention mechanisms, such as the work in [17], we propose a new strategy in this paper, employing a hybrid CNN and ViT model with an additional attention layer inserted between the ViT and the dense layers. The CNN-ViT model will be developed and trained on a selected dataset, with hyperparameters optimized and the model fine-tuned to achieve a target validation accuracy of 93-94%. Therefore, the implementation strategy includes a CNN model, a ViT model, a combined CNN/ViT model, and an attention layer. The experiments will be conducted using Jupyter Notebook for its ease of use and ability to run code sections independently. Python 3.10 will be used, along with the TensorFlow and Keras modules for program development. The hardware used is a Windows 10 desktop PC with an AMD Ryzen 5 3600 6-Core Processor (3593 MHz), 16GB DDR4 RAM, and an AMD RX6600 8GB Graphics Card.

Due to hardware processing constraints, the dataset selection must balance diversity, to ensure the model is well-generalized, and size, where larger datasets would improve training but significantly increase processing time. Several datasets were considered for this project, as listed below.

A. GenImage

The GenImage dataset is a large and diverse collection that includes images from multiple image generators, such as BigGAN, Midjourney, Wukong, and Stable Diffusion versions 1.4 and 1.5 [21]. The dataset contains 3 million images (1.3 million real and 1.35 million fake) and exceeds 500GB in size. Although this would be an ideal candidate for a generalized detection model, I determined that this dataset is too large for successful model training to occur within the project's timescale.

B. CIFAKE

The CIFAKE dataset contains 60,000 real and 60,000 AI-generated images. The 'real' images are collected from the CIFAR-10 dataset [22], while the 'fake' images were generated using Stable Diffusion 1.4 and are equivalent to CIFAR-10 images [11]. All images are pre-processed to a standard size of 32 pixels by 32 pixels, and the total download size is 105 MB.

C. Kaggle-ai-generated-images-vs-real-images dataset

Kaggle also provides another promising dataset titled 'ai-generated-images-vs-real-images' [24]. This dataset is not supported by a research paper; however, it contains 30,000 real images and 30,000 fake images. The real images were collected from Pexels, Unsplash, and WikiArt, while the fake images were generated by Stable Diffusion, MidJourney, and DALL-E (10,000 each). The download size for this dataset exceeds 52GB, and the images are not of uniform size. Additional processing would be required to standardize the images.

For the sake of expedience, and somewhat due to the author's inexperience in the field, the CIFAKE dataset was selected. This dataset was chosen because it includes 120,000 images, all of small proportions (32 pixels by 32 pixels). It was therefore expected that this dataset would be easier to work with and faster to process and train the models on.

Figure 1 represents the model architecture overview, while Figure 2 represents the ViT layer architecture, and Figure 3 represents the attention mechanism architecture.

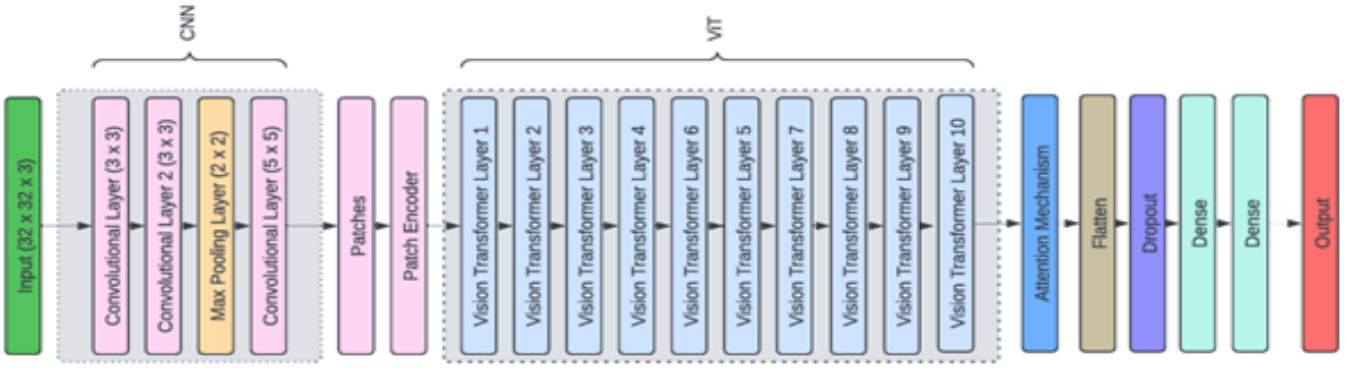

Fig. 1. Model Architechture Overview

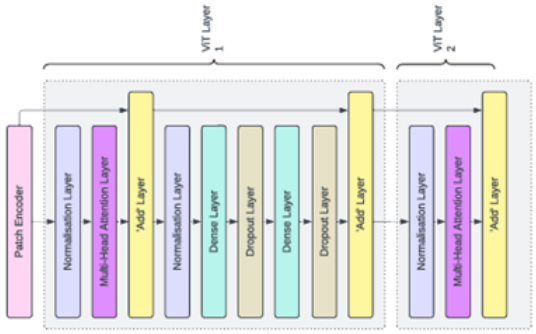

Fig. 2. ViT Layer Architecture

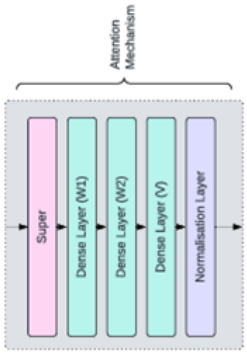

Fig. 3. Attention Mechanism Architecture

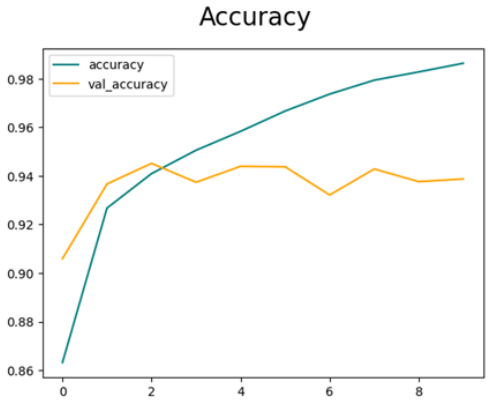

Fig. 4. Accuracy from CNN model training

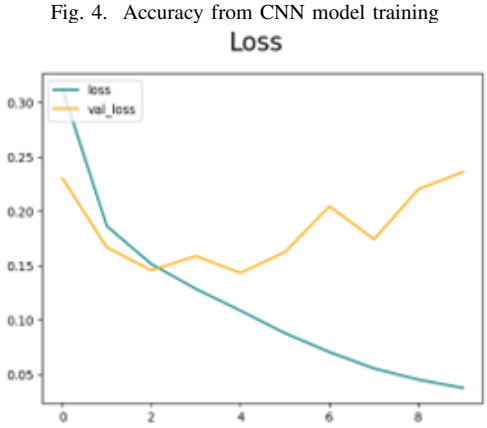

Fig. 5. Loss curves from CNN model training

IV. Implementation

With the dataset selected and the hardware programming methods set, several experiments were conducted where models were constructed and trained on the CIFAKE dataset. The code was largely based on examples from the Keras documentation in [25], with particular emphasis on the KerasTuner page [24], which was invaluable for training the models. For each experiment, the training folder within the CIFAKE dataset, containing 100,000 images, was used and split into 70% training, 20% validation, and 10% testing data. The testing folder was retained as 'unseen' data for any additional testing, if required.

A. Experiment 1 – CNN

The first experiment involved constructing a CNN model and training it on the CIFAKE dataset. Hyperparameter tuning was performed using KerasTuner, as mentioned above, to determine the optimal layers for the CNN model. KerasTuner revealed the optimal layers for the CNN model. With these optimal parameters, the model achieved a validation accu-

racy of 95%. The model was compiled using the optimal parameters, and a summary of the model was produced. It was then trained on the dataset for 10 epochs, achieving a validation accuracy of just under 94%, as seen in Figure 4 & Figure 5. However, the validation loss was nearly 25%, indicating some overfitting in the model. Despite this, the results represent a strong starting point with relatively high accuracy. The overfitting issue can be addressed when the model is combined with the Vision Transformer model.

B. Experiment 2 – Vision Transformer

This model was based on example code from the Keras website [25]. It includes a data augmentation layer that randomly flips, rotates, and zooms the images, effectively increasing the size of the training dataset. The model also incorporates a multilayer perceptron. It was configured with default settings. This model achieved a validation accuracy of 93.3% with a validation loss of 18%, which was lower than the training loss of 20%, as shown in the Figure 6. With this success, the decision was made to proceed with combining the CNN and ViT models in Experiment 3.

C. Experiment 3 – Combined CNN/ViT model

This experiment combines the code of the CNN model and the ViT model. The initial configuration yielded poor results. KerasTuner was incorporated into the code for both the CNN and ViT components to perform hyperparameter tuning.

a) Hyperparameter tuning: The grid method of hyperparameter tuning would have taken too long and would have tested many ineffective parameter sets due to the sheer number of parameters, each with multiple variations. Therefore, random tuning was employed to test multiple variations of each individual parameter within the model. Given the number of parameters in the combined model, tuning took roughly 10 hours per iteration, with most iterations producing unsatisfactory results. As a result, the original CNN layers were configured, and hyperparameter tuning was focused on the ViT layers.

b) Hardcoded Parameters: Once successful, the parameters were hardcoded into the final model. In addition to the hardcoded parameters, a learning rate scheduler was added to the code to help reduce validation loss, and an early stopping function was included to halt training if the validation loss became too high. These adjustments were necessary because the training epochs were set to 100 to fully train the model.

c) Training the model: The model was trained for 100 epochs and produced better results than initially expected. It achieved a training accuracy of 99.79%, a training loss of 0.6%, a validation accuracy of 99.69%, and a validation loss of only 1%. This data is presented in Figure 7 & Figure 8.

As it is shown in Figure 7 &Figure 8, the validation curves closely follow the training curves, except for a few dips and peaks. This indicates that there is significantly less overfitting compared to the initial CNN and ViT models. When tested on unseen data, the accuracy was reported as 99.67%, with a recall of 99.84% and a precision of 99.52%.

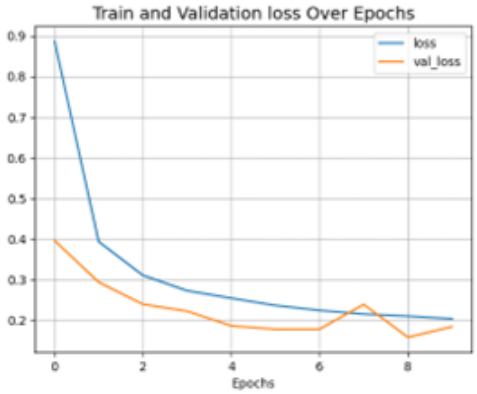

Fig. 6. Training curve for the ViT model

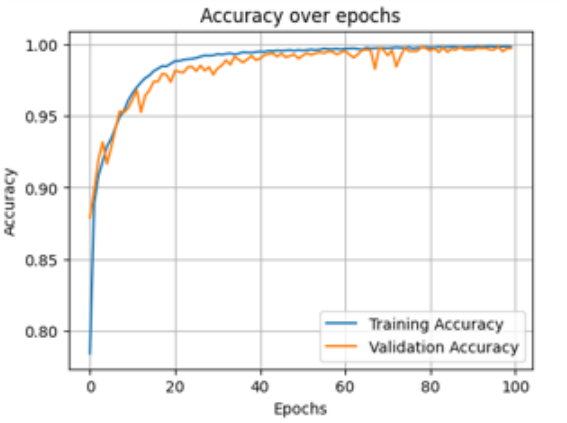

Fig. 7. Accuracy over epochs for the combined CNN/ViT model

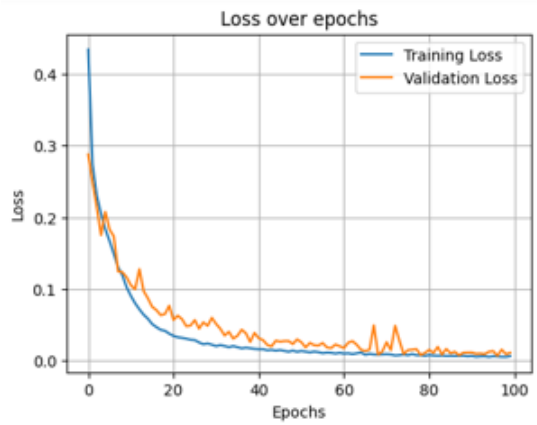

Fig. 8. Loss over epochs for the combined CNN/ViT model

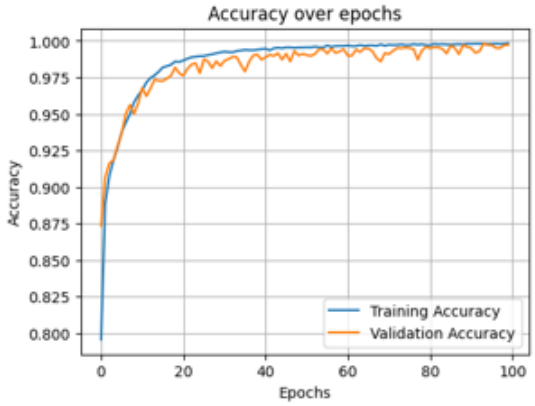

Fig. 9. Accuracy over epochs for the final CNN/ViT+Attention model

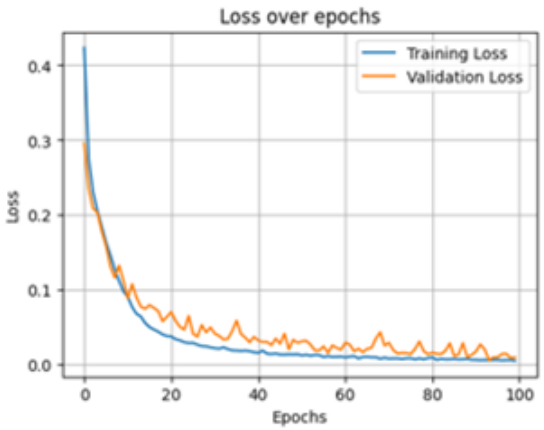

Fig. 10. Loss over epochs for the final CNN/ViT+Attention model

D. Experiment 4 - Combined CNN/ViT + Attention (Final Build)

This final experiment takes the combined CNN/ViT model and adds an additional attention layer. The additional attention layer can learn to prioritize the most critical features extracted by the CNN and ViT layers and, therefore, ensure the optimization of both local and global features prior to final classification.

TABLE I
DETECTION MODEL RESULTS

Model	Accuracy	Loss	Precision
CNN	94%	23%	96.3%
ViT	93.15%	18%	Not recorded
CNN/ViT	99.67%	0.6%	99.52%
CNN/ViT + Attention	99.77%	0.5%	99.8%
CIFAKE (CNN)	92.98%	18%	93.6%

a) Additional Attention Layer: The additional attention layer was added to the model between the CNN/ViT layers and the dense layers. The attention layer is defined in the final code and the model was trained multiple times using different values for the attention layer units. The values started at 32 and were incremented by 32 each time. After several iterations, it was determined that the optimal units value was 128.

b) Training the model: In training, the model achieved a training accuracy of 99.83%, a training loss of 0.5%, a validation accuracy of 99.7%, and a validation loss of 0.9%. This data can be visualized in Figure 9 & Figure 10.

c) Results: When tested with unseen data, the model returned an accuracy of 99.77%, a recall of 99.76%, and a precision of 99.8%. While this is only a marginal improvement over the CNN/ViT model without the added attention layer, there is undoubtedly an overall enhancement.

V. RESULTS & EVALUATION

The results achieved with the CNN/ViT+Attention model are surprisingly good, which may be attributed to the dataset itself. The dataset contains 60,000 real images and 60,000 generated images created using a latent diffusion model. Because of this, the experiments are considered to have been conducted in a controlled environment, and the model has not been exposed to uncontrolled image data. The results of each model are presented in the Table I, along with the results of the CIFAKE CNN detector.

It is worth noting that the VGG16 and ResNet50 pre-trained models were trained on this dataset. However, the results with VGG16 failed to achieve over 90% validation accuracy, with a validation loss greater than 10%. ResNet50 fared worse, achieving just over 75% validation accuracy and over 23% validation loss. This highlights the unsuitability of these models for the selected dataset, further emphasizing the difficulties in producing detection models that generalize well.

Compared to the CIFAKE CNN model, the CNN/ViT+Attention model demonstrates significantly greater accuracy with the same dataset. In fact, even the basic CNN configured in Experiment One shows an improvement over the CIFAKE model. While there is some overfitting in this model, it has been largely addressed in the subsequent experimental models. Improvements can also be observed from the base CNN model to the standalone ViT model, the hybrid CNN/ViT model, and finally to the hybrid model. Although the improvement between the CNN/ViT and CNN/ViT+Attention models is marginal, there is clearly an improvement.

VI. CONCLUSION

In this research, we propose a novel hybrid detection model that combines CNNs and ViTs with an additional attention mechanism layer to improve the interaction between local and global features, leading to enhanced detection accuracy. The proposal for the new hybrid CNN/ViT+Attention model was prototyped and assessed. We describe the hybrid model's architecture, highlighting its features and producing favorable results. When compared to the accuracy achieved by the

hybrid CNN/ViT model, it shows a marginal improvement in its output. However, the comparison against the CIFAKE CNN detection model is somewhat unbalanced, as the hybrid model's architecture is capable of much more detailed feature extraction and reasoning, making it an unfair comparison. In this paper, the following research questions were addressed through experiments:

- Is a hybrid CNN/ViT model capable of enhancing AI-generated image detection accuracy when integrated with an additional attention mechanism? Yes, the hybrid CNN/ViT+Attention model proposed in this paper demonstrated improved detection accuracy by achieving 99.77%, surpassing previous methods.

The proposed prototype can be improved in several ways. One area could be the incorporation of real-world data and scenarios. To increase the practical applicability of the research, future work could involve testing the models on real-world data, including images generated by emerging AI tools and in uncontrolled environments. This would help assess the model's performance outside of a controlled experimental setup and identify potential limitations in real-world applications.

Another area for future work could be an increased emphasis on interpretability. Given the complexity of hybrid models, focusing on improving the interpretability of the model's decisions could be valuable. Techniques like Grad-CAM or other visualization tools can help understand which features the model is focusing on, making the model's predictions more transparent and potentially revealing areas for further improvement.

By addressing these areas, future work can build on the current study's foundation, enhancing the proposed model's effectiveness and contributing to the broader field of AI-generated image detection.

REFERENCES

[1] Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview [Article]. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.

[2] Shahzad, H. F., Rustam, F., Flores, E. S., Luis Vidal Mazon, J., de la Torre Diez, I., & Ashraf, I. (2022). A Review of Image Processing Techniques for Deepfakes [Article]. Sensors (Basel, Switzerland), 22(12), 4556.

[3] Gu, J., Wang, X., Li, C., Zhao, J., Fu, W., Liang, G., & Qiu, J. (2022). AI-enabled image fraud in scientific publications [Article]. Patterns (New York, N.Y.), 3(7), 100511–100511.

[4] Bammey, Q. (2024). Synthbuster: Towards Detection of Diffusion Model Generated Images [Article]. IEEE Open Journal of Signal Processing, 5, 1–9.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need [Article]. ArXiv.Org.

[6] Janutenas, L., Janutenaite-Bogdaniene, J., & Sesok, D. (2023). Deep Learning Methods to Detect Image Falsification [Article]. Applied Sciences, 13(13), 7694.

[7] Passos, L. A., Jodas, D., Costa, K. A. P., Souza, L. A., Rodrigues, D., Del Ser, J., Camacho, D., & Papa, J. P. (2024). A review of deep learning-based approaches for deepfake content detection [Article]. Expert Systems.

[8] Sandotra, N., & Arora, B. (2024). A comprehensive evaluation of feature-based AI techniques for deepfake detection [Article]. Neural Computing & Applications, 36(8), 3859–3887.

[9] Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing.

[10] Hamid, Y., Elyassami, S., Gulzar, Y., Balasaraswathi, V. R., Habuza, T., & Wani, S. (2023). An improvised CNN model for fake image detection. International Journal of Information Technology, 15(1), 5-15.

[11] Bird, J. J., & Lotfi, A. (2024). Cifake: Image classification and explainable identification of ai-generated synthetic images. IEEE Access.

[12] Arshed, M. A., Alwadain, A., Faizan Ali, R., Mumtaz, S., Ibrahim, M., & Muneer, A. (2023). Unmasking Deception: Empowering Deepfake Detection with Vision Transformer Network. Mathematics, 11(17), 3710.

[13] Zhao, X., Huang, P., & Shu, X. (2022). Wavelet-Attention CNN for image classification. Multimedia Systems, 28(3), 915-924.

[14] Kaddar, B., Fezza, S. A., Hamidouche, W., Akhtar, Z., & Hadid, A. (2021, December). HCiT: Deepfake video detection using a hybrid model of CNN features and vision transformer. In 2021 International Conference on Visual Communications and Image Processing (VCIP) (pp. 1-5). IEEE.

[15] Shahin, M., & Deriche, M. (2024, April). A Novel Framework based on a Hybrid Vision Transformer and Deep Neural Network for Deepfake Detection. In 2024 21st International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 329-333). IEEE.

[16] Shao, R., Bi, X.-J., & Chen, Z. (2024). Hybrid ViT-CNN Network for Fine-Grained Image Classification [Article]. IEEE Signal Processing Letters, 31, 1109–1113.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need [Article]. ArXiv.Org.

[18] Khoo, B., Phan, R. C. -W., & Lim, C. (2022). Deepfake attribution: On the source identification of artificially generated images [Article]. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 12(3), e1438-n/a.

[19] Le, B., Shahroz Tariq, Alsharif Abuadbba, Moore, K., & Woo, S. (2023). Why Do Facial Deepfake Detectors Fail? [Article]. ArXiv.Org.

[20] Lin, M., Shang, L., & Gao, X. (2023). Enhancing Interpretability in AI-Generated Image Detection with Genetic Programming [Proceeding]. 2023 IEEE International Conference on Data Mining Workshops (ICDMW), 371–378.

[21] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., & Wang, Y. (2023). GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image.

[22] Krizhevsky, A., Hinton, G., & others. (2009). Learning multiple layers of features from tiny images.

[23] Zhang, T. (2022), ai-generated-images-vs-real-images. Retrieve from: https://www.kaggle.com/datasets/tristanzhang32/ai-generated-images-vs-real-images/data; [accessed September 2024.

[24] O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., & others. (2019). KerasTuner.

[25] Keras (2023), Simple. Flexible. Powerful. Retrieve from: www.keras.io; [accessed September 2024].