JIRBDAI Journal of Institutional Research, Big Data Analytics and Innovation

DEVELOPMENT OF AN INTELLIGENT DISTRACTED DRIVER DETECTION SYSTEM USING DEEP LEARNING TECHNIQUE

Nwadialor Calista Uchenna[1*], Ogbuikwu Rowland I.[2], Olua Henry Ifeanyi[3]

[1]Department of Computer Science, Federal polytechnic Oko, Anambra State
[2]Department of Electrical Engineering, Federal Polytechnic Ohodo. Enugu State
Corresponding Author: okpalacalista23@gmail.com

ABSTRACT: This study presents an intelligent approach to detect distracted driver by integrating a high-tech Deep Learning (DL) model. The DL model adopted in the system is based on the Efficient Channel Attention (ECA) mechanism integrated with the Swin Transformer. By strengthening the capacity of the model to focus on essential aspects of images, the goal is to enhance the classification of driver behaviour. The ECA module is integrated with the Swin Transformer, which is known for its window-based multi-head self-attention mechanism that can observe and document both local and global features. This allows the ECA module to dynamically modify channel weights, for the improvement of feature extraction process. A dataset from the 2016 State Farm Distracted Driving Detection Challenge, which is made up of 22,124 images depicting different distracted driving behaviours, is used to test the algorithm. To maximise the model's performance, the images go through preprocessing, which includes scaling and data augmentation. The implementation's results, which were achieved using MATLAB, demonstrate notable improvements in classification resilience and accuracy. The model applies ECA for improved feature focus and efficiently manages spatial connections among images using methods such as the Adam optimiser.

KEYWORDS: Driver Behaviour Detection; Efficient Channel Attention (ECA), Swin Transformer, Deep Learning; Distracted Driving

I. INTRODUCTION

Human-machine co-driving has grown into a major area of research in autonomous driving as a result of the ongoing progressions in autonomous driving technology (Zhao et al., 2021; Bu et al., 2020). The precise identification of driver behaviour states is especially important in this field as it highlights the cooperative relationship between automated systems and human drivers (Su et al., 2024). However, it is difficult to accomplish completely autonomous driving at L4 or L5 levels in the near future because present automotive intelligence levels mostly fall between L0 and L3 (Ge, 2022). Accordingly, the risk of driver attention rises with extended handoffs of control to automated systems (Yi et al., 2022; Lioi and Bassani, 2024). In complicated and dangerous driving situations, drivers still make the final decisions despite the help of autonomous driving, which calls for constant attention (Jo et al., 2024; Hollosi et al., 2024). As a result, identifying driver behaviour in situations where humans and machines co-drive is crucial to maintaining driving safety (Li et al., 2024; Huang et al., 2024). Manual feature extraction has been the backbone of traditional driver behaviour detection techniques, but it is not very flexible in complicated driving situations and has trouble detecting tiny behavioural changes. To cut down on resource use, He et al. (2024) applied the Histogram of Orientated Gradients (HOG) for feature extraction. For hand gesture identification, Houssem-Lahiani et al., (2018) used the HOG and Local Binary Pattern (LBP) feature extraction approach, which enhanced accuracy and decreased detection time. Nevertheless, these traditional methods frequently face some challenges with capturing delicate driver actions, especially in difficult settings, resulting in less-than-ideal accuracy and resilience.

Convolutional Neural Networks (CNNs) is usually used in a wide concept for the identification of driver behaviour in current years because to the rapid growth of deep learning, particularly for tasks involving behaviour recognition from image or video data (Yang et al., 2024; Li et al., 2023; Tang et al., 2024). CNNs are more effective than standard human feature extraction methods for automatically learning deep features from the data provided. Shahverdy et al., (2020) proposed the use of two-dimensional (2D) CNNs for feature learning on images data generated from driving motions to assist in the automatic detection and classification of driver behaviour. Xing et al., (2019) employed transfer learning to enhance the generalisation capabilities of a pre-trained CNN model for driver behaviour recognition. Even while CNNs are effective at the extraction and learning of deep features and improving detection accuracy, it is still challenging to capture minute behavioural changes. Models based on vision transformers have attracted a lot of attention (Wang et al., 2023). These models, which were first developed for Natural Language Processing (NLP), have excellent modelling abilities for visual tasks, improving convolutional approaches and offering more scalability for both language and vision integration. The Swin transformer (Wu et al., 2024), which is an improved transformer model, offers powerful local and universal feature learning capabilities. Sharma et al. created a model that combines Kervolution and transformer to predict how drivers would behave at roundabout intersections (Sharma et al., 2023). Geng et al., (2023) proposed the application of Dynamic-Learning Spatial-Temporal Transformer Network (DSTTN) for 2D vehicle trajectory prediction.

The driver's actions often fit into a number of different categories. Both enhanced sensitivity to fine-grained features and strong global information processing capacities are necessary for these tasks, which necessitate a detailed analysis of local postures and activities (Xing et al., 2021). In driving scenarios, the driver's actions are often varied and subtle, involving little hand movements on the steering wheel, head rotations, and changes in focus. These subtle characteristics are crucial for accurately identifying driving behaviour (Lotz et al., 2019). Despite the fact that driver behaviour detection algorithms have greatly improved behaviour identification, it is still challenging to accurately identify fine-grained traits. To get over these limitations, the Swin transformer model in this study incorporates the Efficient Channel Attention (ECA) mechanism (Zhou et al., 2023). The application of ECA enhances the model's ability to gather significant fine-grained characteristics while preserving the advantages of the Swin transformer in global feature extraction (Zhang et al., 2024). Additionally, the lightweight channel attention design of ECA reduces processing complexity, making the model more effective, reliable and efficient for real-time driving scenarios. However, the primary objective of this research is to present a driver behaviour detection method based on the Swin transformer and the ECA attention mechanism.

II. RESEARCH METHODOLOGY

In order to enhance the Swin transformer model's feature learning capabilities and raise the precision of driver behaviour recognition, the ECA method was included in this work. The major function of the ECA module is to adjust the channel feature weights dynamically so that the model may focus more efficiently on feature channels that are relevant to the task. Specifically, the ECA module is first used for gathering global spatial information from the input features with the use of Global Average Pooling (GAP) before producing a channel-wise descriptor vector. A One-Dimensional (1D) convolution is applied for further process the vector in order to efficiently capture inter-channel interactions. The weights that are generated are then developed using an activation function, such as Rectified Linear Unit (ReLU) and Sigmoid. Finally, the ECA module performs channel-wise weighted integration by adding these weights to the original channel characteristics, making the relevant channels more significant while suppressing the less significant ones.

By integrating the ECA module following the window-based self-attention approach, the Swin transformer model's feature extraction performance is further enhanced. This updated model keeps the Swin transformer's window-based self-attention architecture while adding the ECA module after each transformer block's output. This enables channel properties to be dynamically changed based on their importance. The enhanced model is able to focus on factors that are relevant to the job by effectively

capturing both local and global information through multi-scale window attention. Additionally, it continues to use the Swin transformer's hierarchical structure.

A. DATA COLLECTION

The dataset, which includes 22,124 images, 25 driver images, and 10 classes of distracted driving behaviours, which was made available by the 2016 State Farm Distracted Driving Detection Challenge and considers a variety of characteristics (Monota et al., 2016). The ten classes of distracted driving behaviours that were selected for training and testing which includes normal driving, texting with the right hand while driving, talking on the phone inclined with the right hand, texting with the left hand while driving, talking on the phone inclined with the left hand, adjusting/tuning the radio while driving, drinking water while driving, looking back while driving forward, adjusting glasses while driving, and talking to other people in different directions while driving. Data augmentation techniques were employed to expand the amount of the dataset after it had been first divided at random. This improved the model's ability to generalise and reduced overfitting during model training. A 7:3 ratio was used to randomly separate the dataset into training and testing sets. Examples of the ten categories of driving habits are shown in Figure 1.

Figure 1: Sample Images of Driving Behaviours
(Source: https://www.kaggle.com/competitions/state-farm-distracted-driver-detection)

B. DATA PREPROCESSING

The State Farm driver dataset, which is openly accessible, is used in this experiment to preprocess the images. The original images are in colour and have a resolution of 640 x 480. The images take up a lot of memory and demand additional processing power during training because of their size. In order to minimise memory usage, the image size is first decreased. The original image should be resized to 224 × 224. Despite offering more precise information, high-resolution images need more memory and processing power, which might cause training to lag and result in excessive resource usage. Reducing the image size improves training performance, optimises memory consumption, and lowers computational expenses. Deep learning frequently uses the 224 × 224 image size because it balances computational efficiency with preserving the quality of input data.

After that, the images are transformed into tensors so that the deep learning framework can analyse and compute them. When images are transformed into tensors, which accurately preserve and express the dimension information of height, width, and channels, models are better able to understand and control the spatial structure and colour information of images. Greyscale images are represented by two-dimensional tensors, whereas colour images are represented by three-dimensional tensors. This tensor structure enables

effective processing and is perfect for parallel computation. It also enables batch processing, which analyses several images at once to expedite the training process. Scaling the images to ensure conformity with the input size of the Swin transformer model further improves training performance. The dataset was expanded by randomly fine-tuning the brightness, contrast, and saturation of the images using the ColorJitter approach to simulate several lighting situations. This process led to an increase in the amount of the dataset.

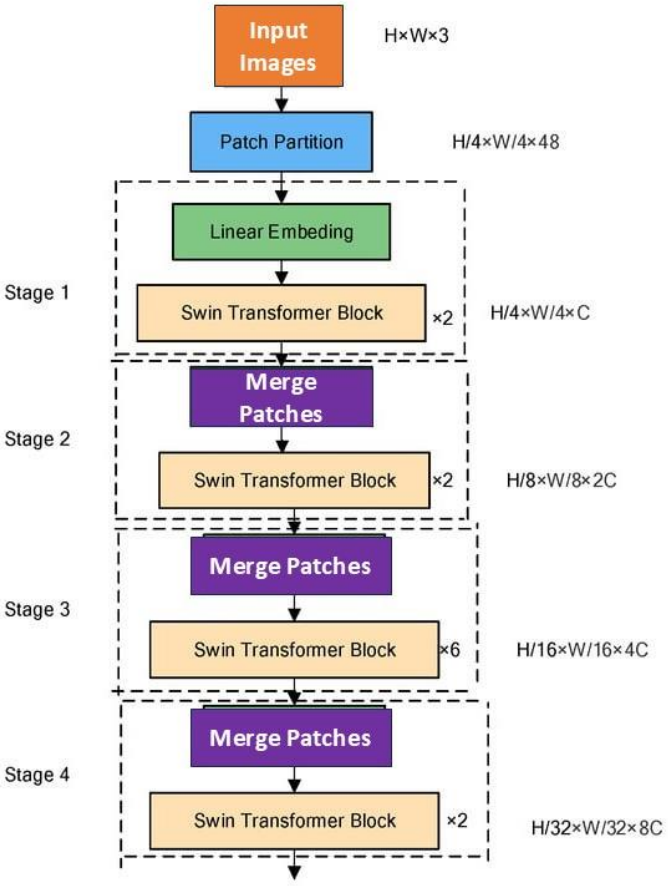

Figure 2: Architecture of the Swin Transformer Network

C. THE SWIM TRANSFORMER MODEL

The transformer model's encoder and decoder modules are designed for feature extraction (Vaswani et al., 2017). Training time is significantly reduced by the encoder's capability for parallel processing. While the CNN relies on advanced layers to extract more features and realise a larger area of receptive field, the transformer model naturally gathers global information without the requirement for significant layer stacking. By using the attention mechanism, it effectively addresses the long-range dependence issue in Recurrent Neural Networks (RNNs). For the driving behaviour recognition task, these transformer advantages enable the model to efficiently gather both local and global information in images. This is particularly important for driver action analysis as it enables more effective cross-regional information collection. The Swin transformer (Li et al., 2022; Song et al., 2024) is an improved transformer model that successfully reduces computing cost by including a sliding window operation. The Swin transformer is primarily composed of a Multilayer Perceptron (MLP), a window-based multi-head self-attention layer, a shifting window multi-head self-attention layer, and normalisation layers (Liu et al., 2021). The network architecture of the Swim transformer is illustrated in Figure 2. Together, these components enhance the model's ability to execute on visual tasks. Specifically, the Swin transformer effectively captures local

information while processing high-resolution images, while the sliding window approach boosts computational efficiency and reduces memory cost.

The network structure of the Swin transformer is used to handle an input image with H × W dimensions and three RGB channels. The image passes through Stages 1, 2, 3, and 4 in order after first passing through the patch partition module. During these stages, the feature map is progressively down sampled. The patch partition module divides the image into smaller patches so that the model can better capture local characteristics, which are crucial for evaluating driver motions. Stage 1 employs a linear embedding structure as opposed to the patch merging structure used in the following stages. The patch merging module handles down sampling, adjusting the number of channels, and lowering the resolution. There is a twofold reduction in resolution with each down sampling step. Reducing image quality and changing the number of channels in driver behaviour detection helps the model better focus on important areas, improving detection performance.

As seen in Figure 3, the block of Swin transformer is made up of two encoder modules that are coupled in series. The attention methods used by these encoders vary. A Shifted Window-Based Multihead Self-Attention Module (SW-MSA) is used in one, while a Window-based Multi-head Self-Attention Module (W-MSA) is used in the other. Accurate local feature extraction is essential for driving behaviour identification. In order to minimize the system's computational complexity, the W-MSA module divides the feature map into fixed-size windows and individually carries out self-attention actions inside each window. This method, however, limits the flow of information between windows. The SW-MSA module, on the other hand, allows communication and information exchange between windows by offsetting the window partitions. This makes it possible for the model to record cross-window relational data, which is very useful for identifying intricate patterns of driver behaviour.

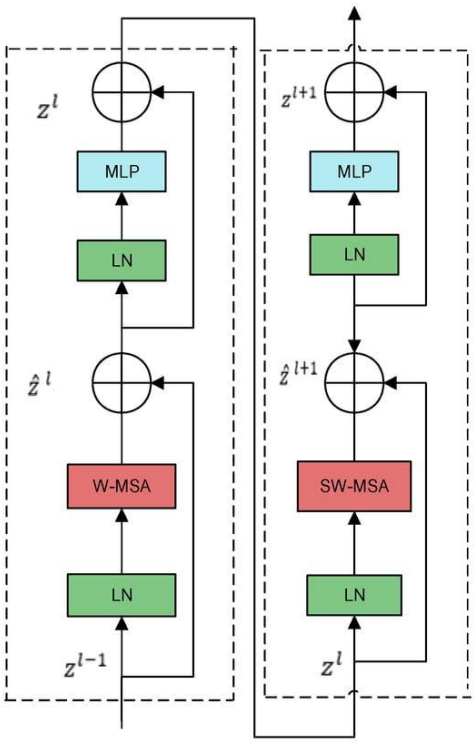

Figure 3: Structure of Swin Transformer Block (Cui et al., 2024)

Despite its notable benefits in handling high-resolution images and intricate visual tasks, the Swin transformer still has several drawbacks. First off, processing large-scale information might cause a bottleneck in CPU resources because to the Swin transformer's comparatively high computational complexity. Furthermore, the constant window size may restrict the model's ability to capture long-range

dependencies, particularly when complex interrelationships need to be recorded, even while the sliding-window self-attention technique efficiently minimises computation. This creates further difficulties for model performance in jobs like driver behaviour detection, because actions entail interactions across several areas. LN is an operation for normalisation. MLPs stand for multi-layer perceptron blocks. To effectively describe both local and global elements, the shifted window approach alternates between W − MSA and SW − MSA layers. To address this, the integration of the ECA mechanism with the Swin transformer model is considered. Through dynamically modifying the attention weights across many channels, the ECA mechanism enhances the model's ability to concentrate on crucial information. Additionally, due of its lightweight channel attention architecture, which ensures little computing cost, it may provide outstanding classification performance in driver behaviour detection tasks.

D. EFFICIENT CHANNEL ATTENTION (ECA) MECHANISM

By enhancing the assessment of the significance of channel characteristics, the ECA mechanism is intended to improve the deep learning models' capacity for feature extraction and, therefore, model performance. ECA reduces computational complexity and inference time by avoiding the need of conventional fully linked layers or matrix operations through efficient computing. The model can better capture important characteristics by learning global feature representations and weighting channels according to local information. Figure 4 depicts the particular structure.

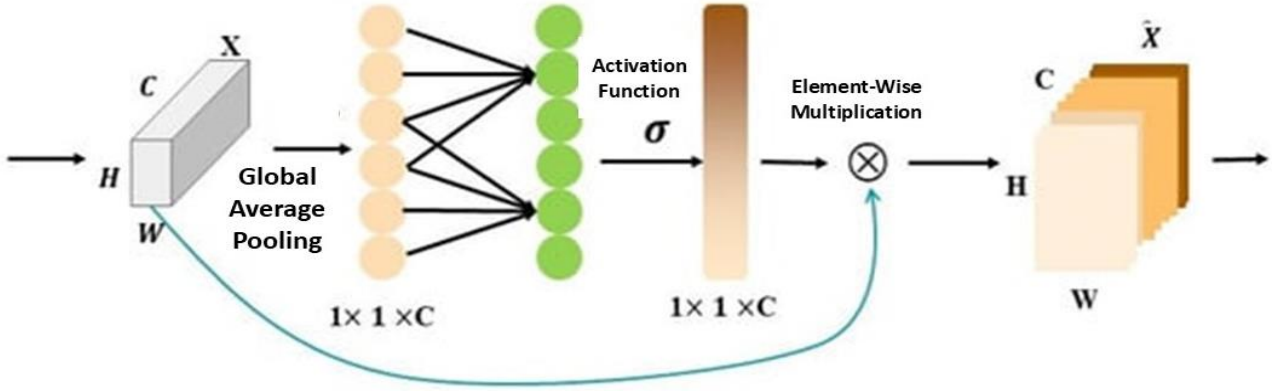

Figure 4: Structure of the ECA Mechanism.

The global characteristics of every channel are extracted from the input feature map using global average pooling, as shown in Figure 4. Inter-channel dependencies are then captured using a one-dimensional convolution with an adaptively chosen kernel size. A Sigmoid function is used to the convolution result in order to provide normalised attention weights. To improve the representation of significant features, these attention weights are then multiplied element-wise with the original input feature map. Based on the number of channels, the ECA module adaptively determines the one-dimensional convolution's kernel size. The formula in Equation 1 can be used to determine the kernel size (Cui et al., 2024):

$$\left[k = \left|\frac{log_2(C)}{\gamma}\right|_{odd}\right] \quad (1)$$

where $|t|_{odd}$ denotes the next odd number of t, γ and b are hyperparameters, and C is the number of input channels. To determine the relative relevance of each channel, a 1-Dimensional (1D) convolution is performed to the input features. The precise computation looks like this (Cui et al., 2024):

$$[out = Conv1D_k(in)] \quad (2)$$

where (in) represents the input features, out for the output features, and Conv1D for 1D convolution.

E. INTEGRATION OF THE SWIM TRANSFORMER MODEL WITH CHANNEL ATTENTION MECHANISM

The ECA module maintains a low computational cost during training and inference since it does not reliant on completely linked layers. Window-based self-attention is the level at which the ECA layer is implemented. The ECA layer applies channel-wise attention weighting to the output features once the self-attention process is completed, therefore dynamically adjusting the weight of each channel. To guarantee that the output of the attention mechanism is completely employed in later phases, the weighted output is then used as input for further processing. Additionally, the ECA module does fine-grained dynamic modification of channel information, whilst the window-based self-attention gathers local spatial information. This makes it possible for the model to successfully combine global channel information with local spatial data. The ECA module optimises the model's capacity to acquire crucial feature information for driver behaviour detection tasks while maintaining the Swin transformer's hierarchical structure.

Figure 5 presents the updated Swin transformer model's structure for detecting driver behaviour. The following describes the specific steps involved in the driver behaviour detection model that uses the Swin transformer and the integrated channel attention mechanism, as shown in Figure 5: The enhancement to the original Swin transformer model is shown in the top portion of Figure 5. At the initial stage, the model receives the driver image and modifies its size to fit the input measurements. The image is then split into many patches using the patch partition technique. After that, a feature vector is created from each patch. Each patch is then subjected to a linear embedding technique, which converts the flattened patches into fixed-dimensional vectors. The Swin transformer model processes these vectors as inputs and uses them to extract features. In every step of the Swin transformer, the ECA attention mechanism is introduced after every Swin block. In order to improve the capacity of the model to extract features from the driver behaviour images and enable precise categorisation of driver behaviour images, this technique dynamically modifies the channel weights.

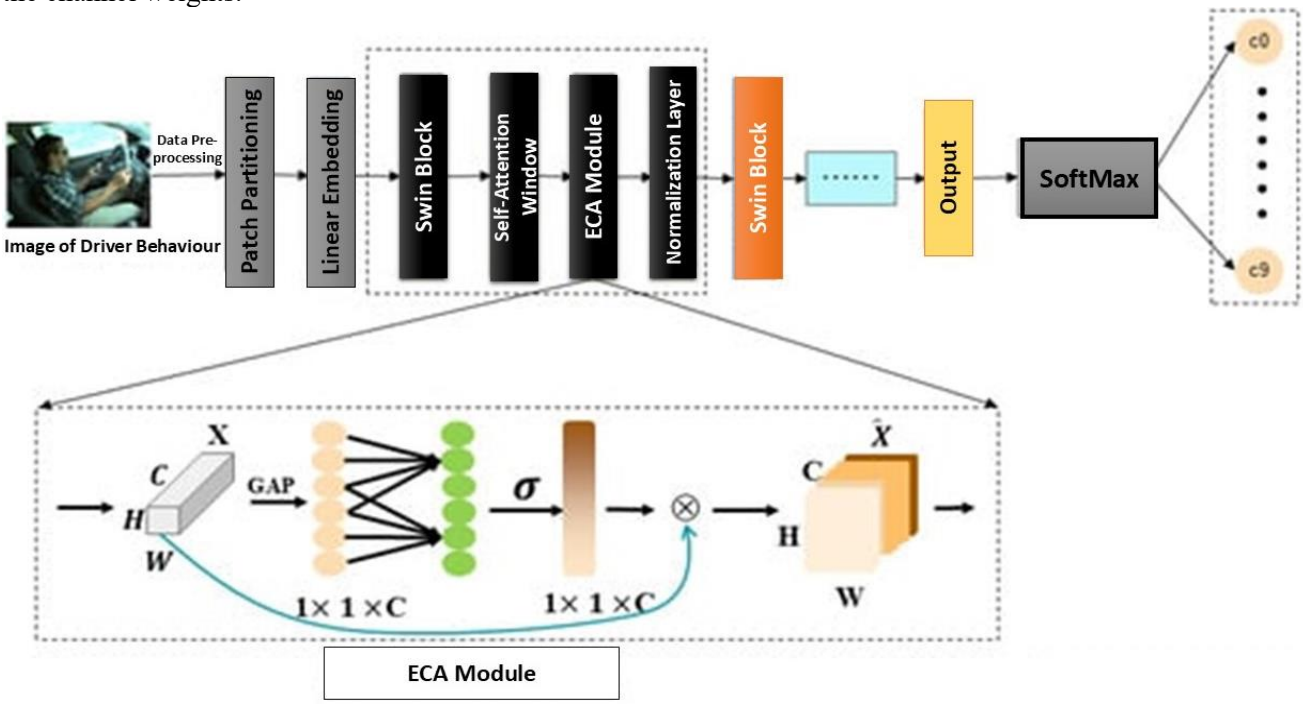

Figure 4: The Complete Architecture of the Improved Swin Transformer Model for the Detection of Driver Behaviour

The ECA module is shown in the bottom section of Figure 4 and is a component of the window self-attention module, which first computes self-attention in each window to extract local characteristics, then improves channel-wise attention and feature representation by highlighting the most important channel properties.

After using residual connections and normalisation procedures, the training process is stabilised, which results in the classification of driving behaviours.

F. SYSTEM IMPLEMENTATION

The Swin Transformer with Efficient Channel Attention (ECA) mechanism implementation for distracted driver classification involves a number of important MATLAB operations. The dataset is initially pre-processed using data augmentation methods, such adjusting contrast and brightness and reducing the size of images to a standard size. The dataset is separated into training and testing subsets using MATLAB's imageDatastore and augmentedImageDatastore functions. A custom ECA layer is built to enhance feature representation by calculating attention weights for channel-wise refinement using sigmoid activation, global average pooling, and a one-dimensional convolution. Combining the ECA layer with the Swin Transformer blocks. which include window-based and shifting multi-head self-attention techniques improves feature extraction. The model architecture starts with an input layer and patch embedding. Next are a number of Swin Transformer blocks that are used in conjunction with ECA to learn hierarchical features. In the end, there are SoftMax, completely connected, and classification layers for the identification of distracted drivers. Training with the appropriate batch sizes, validation data, and learning rates is done using the Adam optimiser. Finally, the test dataset is used to evaluate the trained model's performance and accuracy. This approach ensures precise and effective classification by utilising MATLAB's deep learning and attention techniques.

III. RESULTS AND DISCUSSION

Using a hybrid deep learning model that integrated the Swin Transformer and Enhanced Channel Attention (ECA) processes, the distracted driver classification system was created in MATLAB. The dataset was pre-processed by reducing images, normalising pixel values, and enriching the data in order to boost robustness. A split dataset (80% training, 20% validation) was used to train the model, and trial-and-error was used to optimise parameters including learning rate, batch size, and number of epochs. ECA improved channel-wise feature learning, and the Swin Transformer efficiently handled spatial connections. The model produced encouraging results during training and validation, with important metrics indicating its reliability in identifying distracted driving behaviours. To guarantee robustness, ten-fold cross-validation approach was applied. Ten equal parts of the dataset were created, nine of which were used for training and one for validation. Metrics including accuracy, precision, recall, and F1-score for each fold were calculated to offer a comprehensive evaluation of the system's performance. Table 1 shows the results of the 10-fold validation as well as the average performance metrics of implementing the recommended solution.

Table 1: System Validated Results

Epochs	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	94.8	95.5	94.0	94.7
2	96.2	96.8	95.8	96.3
3	95.7	96.3	95.2	95.7
4	96.0	96.7	95.6	96.1
5	95.5	96.1	95.0	95.5
6	96.1	96.9	95.7	96.3
7	95.9	96.6	95.5	96.0
8	95.4	96.0	94.9	95.4
9	96.3	97.0	95.9	96.4
10	95.8	96.5	95.2	95.8
Average	**95.8**	**96.5**	**95.2**	**95.8**

Table 1 summarises the performance across all 10 folds. The average accuracy of 95.8%, precision of 96.5%, recall of 95.2%, and F1-score of 95.8% all show consistent and high-performance characteristics across folds. The results confirm the robustness and reliability of the hybrid deep learning model in correctly classifying distracted driving behaviours.

IV. CONCLUSION

The study presents a hybrid deep learning model that combines the Efficient Channel Attention (ECA) mechanism with the Swin Transformer to categorise inattentive drivers. By dynamically altering the channel weights, the ECA module enhances the state-of-the-art vision transformer model Swin Transformer, allowing the model to focus on the most relevant components in the incoming images. The key components of the model's architecture are global average pooling, one-dimensional convolution for channel attention, and a window-based multi-head self-attention mechanism. Together, these elements enhance the model's capacity to correctly categorise driver behaviour by successfully capturing both local and global traits. The dataset, which was extracted from the 2016 State Farm Distracted Driving Detection Challenge, consists of 22,124 images showing various distracted driving behaviours. The images undergo preprocessing, which includes data augmentation and scaling, to optimise the model's performance. The application's outcomes demonstrate how effectively the proposed model classifies distracted driving behaviours. The Swin Transformer with ECA greatly improved feature extraction and classification accuracy in comparison to traditional methods. The model was tuned using the Adam optimiser following training phase adjustments to learning rates, batch sizes, and validation data. Recall was 95.2%, accuracy was 95.8%, precision was 96.5%, and F1-score was 95.8%, indicating consistent and strong performance metrics across folds. By efficiently handling spatial relationships between images and honing channel-wise characteristics using the ECA mechanism, the model was able to achieve high classification accuracy. The MATLAB implementation produced encouraging results as the model reliably identified distracted driving behaviours, which might be important for real-time safety applications.

REFERENCES

Bu, L. G., Chen, C. H., Zhang, G., Liu, B. F., Dong, G. J., & Yuan, X. (2020). A hybrid intelligence approach for sustainable service innovation of smart and connected product: A case study. *Advances in Engineering Informatics, 46,* 101163. https://doi.org/10.1016/j.aei.2020.101163

Cui J., Chen Y., Wu Z., Wu H., & Wu A., (2024) A Driver Behavior Detection Model for Human-Machine Co-Driving Systems Based on an Improved Swin Transformer. World Electr. Veh. J. 2025, 16, 7. https://doi.org/10.3390/wevj16010007

Ge, G. (2022). Analysis of dynamic characteristics of man-machine co-driving vehicle during driving right switching. *Automation and Remote Control, 56*(2), 166–179.

Geng, M. S., Chen, Y., Xia, Y. J., & Chen, X. Q. (2023). Dynamic-learning spatial-temporal Transformer network for vehicular trajectory prediction at urban intersections. *Transportation Research Part C: Emerging Technologies, 156,* 104330. https://doi.org/10.1016/j.trc.2023.104330

He, Y. H., Huang, J. Y., & Pan, Y. M. (2024). A novel low-resource consumption and high-speed hardware implementation of HOG feature extraction on FPGA for human detection. *Integration, 97,* 102208. https://doi.org/10.1016/j.vlsi.2024.102208

Hollósi, J., Ballagi, Á., Kovács, G., Fischer, S., & Nagy, V. (2024). Detection of bus driver mobile phone usage using Kolmogorov-Arnold Networks. *Computers, 13*(12), 218. https://doi.org/10.3390/computers13120218

Huang, T., Fu, R., & Chen, Y. X. (2021). Deep driver behavior detection model based on human brain consolidated learning for shared autonomy systems. *Measurement, 179,* 109463. https://doi.org/10.1016/j.measurement.2021.109463

Jo, Y., Jung, A., Oh, C., & Park, J. (2024). Characterizing the driving behavior of manual vehicles following autonomous vehicles and its impact on mixed traffic performance. *Transportation Research Part F: Traffic Psychology and Behaviour, 107,* 69–83. https://doi.org/10.1016/j.trf.2023.09.004

Lahiani, H., & Neji, M. (2018). Hand gesture recognition method based on HOG-LBP features for mobile devices. *Procedia Computer Science, 126,* 254–263. https://doi.org/10.1016/j.procs.2018.07.216

Li, C. J., Xu, C. C., Chen, Y. S., & Li, Z. B. (2024). Development and experiment of an intelligent connected cooperative vehicle infrastructure system based on multiple V2I modes and BWM-IGR method. *Physica A: Statistical Mechanics and its Applications, 635,* 129498. https://doi.org/10.1016/j.physa.2024.129498

Li, J. R., Yu, C., Shi, J. Y., Zhang, C. L., & Ke, T. (2022). Vehicle re-identification method based on Swin-Transformer network. *Array, 16,* 100255. https://doi.org/10.1016/j.array.2022.100255

Li, R. L., Gao, R. B., & Suganthan, P. N. (2023). A decomposition-based hybrid ensemble CNN framework for driver fatigue recognition. *Information Sciences, 624,* 833–848. https://doi.org/10.1016/j.ins.2023.05.020

Lioi, A., & Bassani, M. (2024). The impact of auditory advanced driver distraction warning devices on the behavior of middle-aged drivers along urban roads. *Transportation Engineering, 17,* 100263. https://doi.org/10.1016/j.treng.2023.100263

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* Montreal, QC, Canada, 10–17 October 2021.

Lotz, A., Russwinkel, N., & Wohlfarth, E. (2019). Response times and gaze behavior of truck drivers in time-critical conditional automated driving take-overs. *Transportation Research Part F: Traffic Psychology and Behaviour, 64,* 532–551. https://doi.org/10.1016/j.trf.2019.05.012

Montoya, A., Holman, D., SF_data_science, Smith, T., & Kan, W. (2016). State Farm distracted driver detection. *Kaggle.* San Francisco, CA, USA. Retrieved from https://www.kaggle.com/c/state-farm-distracted-driver-detection

Shahverdy, M., Fathy, M., Berangi, R., & Sabokrou, M. (2020). Driver behavior detection and classification using deep convolutional neural networks. *Expert Systems with Applications, 149,* 113240. https://doi.org/10.1016/j.eswa.2020.113240

Sharma, O., Sahoo, N. C., & Puhan, N. B. (2023). Kernelized convolutional transformer network-based driver behavior estimation for conflict resolution at unsignalized roundabouts. *ISA Transactions, 133,* 13–28. https://doi.org/10.1016/j.isatra.2023.01.001

Song, C., Song, Q., & Cao, F. (2024). Driver distraction detection based on improved Swin Transformer. In *Proceedings of the 2024 43rd Chinese Control Conference (CCC), Kunming, China, 28–31 July 2024* (pp. 2032–8037).

Su, C. R., Yang, H. H., Li, J., & Wu, X. D. (2024). Steering authority allocation strategy for human-machine shared control based on driver take-over feasibility. *Computers & Electrical Engineering, 120,* 109753. https://doi.org/10.1016/j.compeleceng.2024.109753

Tang, X. X., Chen, Y., Ma, Y. F., Yang, W. X., Zhou, H. P., & Huang, J. Z. (2024). A lightweight model combining convolutional neural network and Transformer for driver distraction recognition. *Engineering Applications of Artificial Intelligence, 132,* 107910. https://doi.org/10.1016/j.engappai.2023.107910

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv,* arXiv:1706.03762. https://arxiv.org/abs/1706.03762

Wang, Y. W., Qiu, Y. Y., Cheng, P. T., & Zhang, J. Y. (2023). Transformer-based descriptors with fine-grained region supervisions for visual place recognition. *Knowledge-Based Systems, 280,* 110993. https://doi.org/10.1016/j.knosys.2023.110993

Wu, S., Zhang, G. J., & Liu, X. F. (2024). SwinSOD: Salient object detection using Swin-transformer. *Image and Vision Computing, 146,* 105039. https://doi.org/10.1016/j.imavis.2023.105039

Xing, Y., Lv, C., Cao, D. P., & Hang, P. (2021). Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transportation Research Part C: Emerging Technologies, 128,* 103199. https://doi.org/10.1016/j.trc.2021.103199

Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., & Wang, F.-Y. (2019). Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Transactions on Vehicular Technology, 68*(6), 5379–5390. https://doi.org/10.1109/TVT.2019.2909411

Yang, D. W., Wang, Y., Wei, R., Guan, J. P., Huang, X. H., Cai, W., & Jiang, Z. (2024). An efficient multi-task learning CNN for driver attention monitoring. *Journal of Systems Architecture, 148,* 103085. https://doi.org/10.1016/j.sysarc.2023.103085

Yi, B. L., Cao, H. T., Song, X. L., Zhao, S., Guo, W. F., & Li, M. J. (2022). How to identify the take-over criticality in conditionally automated driving? An examination using drivers' physiological parameters and situational factors. *Transportation Research Part F: Traffic Psychology and Behaviour, 85,* 161–178. https://doi.org/10.1016/j.trf.2021.10.003

Zhang, Y., Zhan, Q., & Ma, Z. (2024). EfficientNet-ECA: A lightweight network based on efficient channel attention for class-imbalanced welding defects classification. *Advanced Engineering Informatics, 62,* 102737. https://doi.org/10.1016/j.aei.2023.102737

Zhao, L., Yang, F., Bu, L. G., Han, S., Zhang, G. X., & Luo, Y. (2021). Driver behavior detection via adaptive spatial attention mechanism. *Advances in Engineering Informatics, 48,* 101280. https://doi.org/10.1016/j.aei.2021.101280

Zhou, X., Yang, W., HaiLaTi, K., & Liao, Y. (2023). The detection of fraudulent smart contracts based on ECA-EfficientNet and data enhancement. *Computers, Materials, and Continua, 77,* 4073–4087. https://doi.org/10.32604/cmc.2023.032210