Comparative study of energy and accuracy in spatio-temporal models for engagement classification

Fahrur Aslami [a,1,*], Peeraya Sripian [a,2]

[a] Shibaura Institute of Technology, 3-7-5 Toyosu, Koto-ku, Tokyo, Japan
[1] nb23506@shibaura-it.ac.jp, [2] peeraya@shibaura-it.ac.jp
* corresponding author

ARTICLE INFO

Article history
Received February 2, 2026
Revised March 16, 2026
Accepted May 17, 2026
Available online August 31, 2026

Keywords
Student engagement
spatio-temporal deep learning
EfficientNet
Green AI
Energy efficiency

ABSTRACT

Energy consumption is a growing concern in deep learning, motivating the Green AI paradigm, where models are evaluated not only based on predictive performance but also on energy efficiency. Most existing evaluations focus on spatial tasks and do not fully capture the computational and temporal costs of video-based workflows. This study presents a systematic energy evaluation of spatio-temporal deep learning for video-based student engagement classification. Using the DAiSEE dataset, we compare EfficientNet variants (B0–B7) as frame-level feature extractors and model temporal dynamics with LSTMs on fixed-length sequences. GPU power consumption is measured during training using nvidia-smi, and energy efficiency is quantified with the Kappa Energy Index (KEI), defined as Cohen's Kappa divided by energy consumption (kWh). The results show a clear trade-off between accuracy and energy: EfficientNet-B7 achieves the highest accuracy (0.62) but incurs the highest energy cost (≈5 kWh), resulting in a low KEI, while EfficientNet-B0 achieves competitive accuracy (0.59) with the highest KEI (2.147) due to its low energy consumption (≈0.38 kWh). EfficientNet-B3 strikes a favorable balance (accuracy ≈ 0.61, KEI = 0.747), outperforming larger models under resource constraints. These findings suggest that deeper models do not always maximize energy efficiency, and the KEI value provides a practical metric to guide the selection of energy-efficient models.

1. Introduction

The rapid growth of machine learning (ML), particularly deep learning (DL), has significantly increased computing needs across various sectors, from industry to public services [1]–[3]. Modern DL models generally require high computing resources for both training and inference, making energy consumption and carbon footprint increasingly important issues in AI system development [4]–[6]. This situation has drawn global attention to Green AI, a paradigm that demands measuring model success not only by accuracy but also by energy efficiency and environmental sustainability [7].

As concerns about the environmental impact of large-scale computing grow, studies have begun to emphasize the importance of measuring energy consumption and carbon emissions in DL systems [8]–[10]. Various approaches to energy consumption estimation have been developed, including software-based energy estimation, hardware meters, and efficiency metrics that combine performance and computational cost [11], [12]. Furthermore, optimization strategies such as pruning, quantization, and neural architecture search (NAS) have been proposed to reduce energy costs without drastically degrading performance [13]–[15]. However, despite the rapid growth of research on DL energy efficiency, evaluations often focus on static model configurations or tasks that do not fully represent realistic spatio-temporal deployment scenarios [16]–[18].

Spatio-temporal models significantly increase power consumption in video-based deep learning systems because they process multiple frames within a temporal window rather than a single input frame [19], [20]. This increases computational complexity with sequence length and frame sampling strategy [20], [21], which directly impacts memory usage, data transfer, inference time, and energy consumption [22], [23]. In continuous, real-time video applications such as student engagement monitoring [24], [25], the overhead becomes a critical obstacle to scalable, sustainable deployment [26]–[28]. Previous energy evaluation studies also show that the choice of architecture and training configuration substantially affects energy consumption during training and inference [29]–[31].

Although energy efficiency has become a significant concern in Green AI [32], comprehensive energy evaluation of spatio-temporal deep learning models in video pipelines remains inadequately addressed [33], [34]. Most existing studies evaluate energy consumption in spatial settings, such as frame-based or single-image classification [35], [36]. Therefore, an explicit and standardized energy performance evaluation framework for spatio-temporal video pipelines is needed, one that systematically accounts for temporal factors while enabling fair and consistent comparisons between spatial and spatio-temporal models [37], [38].

Based on this gap, this study aims to establish a representative evaluation of the energy-performance trade-off between spatial and spatio-temporal deep learning models for video processing. The objectives of this study are as follows: 1) Measure and compare energy consumption and prediction performance between spatial and spatio-temporal pipelines, 2) Analyze the contribution of temporal factors to changes in energy costs relative to performance improvements; and 3) formulate practical recommendations for selecting efficient and sustainable architectures for implementing video-based engagement monitoring systems.

Specifically, this study uses the video-based student engagement classification task as a use case because this application is characterized by continuous monitoring and requires repeated inference during learning sessions, making energy efficiency a limiting factor in real-world deployments [39], [40]. To obtain a systematic and scalable comparison, the evaluation was conducted on the EfficientNet family (B0–B7) as a representative CNN model, ranging from light to large [41], using the DAiSEE public benchmark dataset [42]. The main contributions of this research are summarized as follows:

- A green AI-based energy performance benchmark for spatio-temporal models. This research presents a systematic evaluation of the energy performance trade-off in a video-based student engagement classification task, representing a real-time continuous monitoring scenario in educational environments.
- A spatio-temporal model comparative analysis. We design an evaluation protocol that contributes to a spatio-temporal model, enabling an objective comparison of video pipeline performance.
- Characterizing the efficiency of lightweight to large-scale CNN architectures. By evaluating the EfficientNet architecture family (B0–B7) on the DAiSEE benchmark dataset [13], [15], this research reveals how increasing model capacity and integrating temporal components affect energy consumption, inference latency, and classification performance.

2. Materials

In this study, the DAiSEE dataset was used for video-based student engagement classification, as it is widely adopted in student engagement detection research [43]. Two model pipelines were designed for the analysis: a spatial and a spatio-temporal pipeline, both run under identical experimental conditions. In the spatial pipeline, the EfficientNet model (B0–B7) was used for single-frame inference classification [44]. In the spatial-temporal pipeline, EfficientNet (Table 1) served as a feature extractor for each frame, which was then modeled temporally using LSTM within a specific time window. In the spatial-temporal pipeline, EfficientNet served as a feature extractor for each frame, which was then modeled temporally using LSTM within a specific time window. All experiments used the same hardware and software configuration and followed a consistent energy-measurement protocol [45].

Table 1. Characteristics of the EfficientNet [14]

Model	Params	FLOPs	Feature	Resolution	Accuracy
B7	66M	37B	2560	600 x 600	84.4%
B6	43M	19B	2304	528 x 528	84%
B5	30M	9.9B	2048	456 x 456	83.3%
B4	19M	4.2B	1792	380 x 380	82.6%
B3	12M	1.8B	1536	300 x 300	81.1%
B2	9.2M	1.0B	1408	260 x260	79.8%
B1	7.8M	0.7B	1280	240 x 240	78.8%
B0	5.3M	0.39B	1280	224 x 224	76.3%

2.1. Dataset

The input data for this research comes from DAiSEE (Dataset for Affective States in E-learning Environment), a publicly available dataset designed to analyze affective states in the context of online learning (e-learning). This dataset is widely used in video-based student engagement research, enabling more standardized comparisons with previous studies [46], [47]. DAiSEE contains video recordings of 112 Asian university students (male and female), aged 18–30, collected during various e-learning sessions under diverse recording conditions, including different lighting setups, camera devices, and environments (e.g., laboratories, classrooms, and private rooms) [43].

The dataset provides expert annotations for four affective states: confusion, boredom, engagement, and frustration. Each state is labeled using four intensity levels (0, 1, 2, 3), which is more informative than binary classification and supports fine-grained engagement modeling [43]. DAiSEE consists of thousands of video clips (10 seconds each) recorded at a resolution of 640 × 480 pixels (approximately 300 frames per clip). The engagement label distribution is notably imbalanced, with 61 samples at intensity level 0, 455 at level 1, 4,419 at level 2, and 3,986 at level 3. This imbalance may bias learning toward majority classes and reduce sensitivity to minority classes [48], [49]. This study focuses on the engagement label, as engagement is a key indicator for monitoring learning processes and is relevant for evaluating both accuracy and energy efficiency in video-based measurement systems.

2.2. Spatial Model

CNN (convolutional neural networks) remain widely used due to their computational efficiency and strong representation capabilities despite the development of various CNN-based model derivatives, but EfficientNet introduces a combined scaling strategy that systematically scales the network depth, width, and input resolution designed for computational efficiency [44]. EfficientNet was developed for large-scale image classification tasks. It consists of convolutional layers, Mobile Inverted Bottleneck Convolution (MBConv) blocks, and a fully connected classification head [44]. The MBConv blocks employ inverted residual connections with skip connections, enabling efficient feature extraction while maintaining a low computational cost. EfficientNet also adopts the Swish activation function, which improves gradient flow compared to conventional activation functions such as ReLU. EfficientNet B0–B7 models are obtained through Neural Architecture Search (NAS) [50]. The variants are constructed using a compound scaling strategy that jointly scales network depth, width, and input resolution in a balanced manner. As a result, higher-indexed variants exhibit increased model capacity, parameter counts, and computational complexity.

The architectural characteristics of each variant, including the number of parameters, FLOPs, feature dimensionality, and input resolution, are summarized in Table 1 [44]. In this study, EfficientNet variants from B0 to B7 are used as spatial feature extractors. When employed for feature extraction, the top classification layers are removed, and only the convolutional backbone is retained. Each EfficientNet variant produces a one-dimensional feature vector that encodes spatial visual information from the input image.

2.3. Temporal Model

Recent research in video classification has increasingly focused on models based on transformers and RNN (recurrent neural network) architectures. Both models have demonstrated strong performance in large-scale benchmarks but often require large computational resources and training costs. Vision Transformer (ViT) is used for visual recognition by modeling global dependencies through a self-attention mechanism [51]. Furthermore, RNN models are applied by simultaneously learning spatial and temporal representations from video sequences, which allows for improved performance in complex video understanding tasks [52].

Long Short-Term Memory (LSTM) is a type of RNN designed to model sequential data and capture sequential long-term temporal information. LSTM uses 3 gate mechanisms, namely input, forget, and output gates, to regulate the flow of information and reduce the vanishing gradient problem [53], [54], so that video classification, which is a sequence of image frames is suitable for this approach [55]. In the application of this study, LSTM is used to model the temporal relationship between spatial features that have been extracted for each frame to capture the temporal dynamics in student engagement tasks.

2.4. Hardware and Software

All experiments were conducted on a desktop workstation equipped with an Intel Core i9 processor and an NVIDIA RTX 3090 GPU with 24 GB of VRAM, providing sufficient computational resources for training and evaluating deep learning models. The operating system environment consisted of Windows 10 as the host system, with Ubuntu 24.04 LTS deployed through the Windows Subsystem for Linux (WSL) to facilitate a Linux-based development workflow.

2.5. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

3. Methods

3.1. Experiment Setup

The experimental design in this study employs a pipeline integrating spatial feature extraction, sequential modeling, and multi-metric evaluation shown in Fig. 1, to analyze the behavior of the EfficientNet B0-B7. The workflow begins with the DAiSEE dataset, which undergoes preprocessing before entering the Spatial Feature Extraction unit, which comprises eight model variants ranging from EfficientNet B0 to B7.

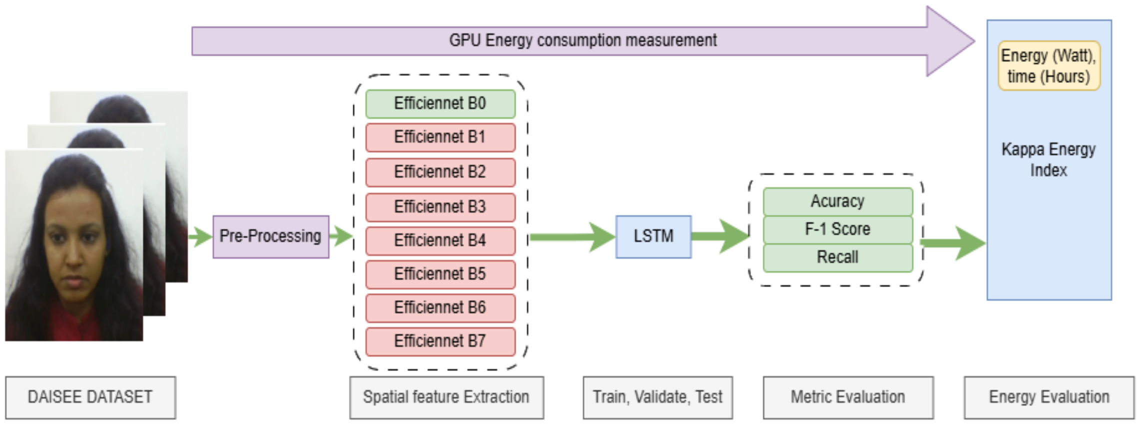


Fig. 1. Experimental setup for video-based student engagement classification.

The spatial feature vectors generated by each backbone variant are then passed to the Long Short-Term Memory (LSTM) layer for training, validation, and testing to capture temporal dependencies in the data. During this process, the system performs parallel GPU energy consumption measurements (in watts and hours), which are recorded using the NVIDIA System Management Interface (nvidia-smi) to generate the Kappa Energy Index as the final efficiency metric [17]. Previous empirical studies show that GPU power consumption dominates total system energy during deep learning training, while CPU and RAM contributions remain relatively stable across models [56]. This indicates that although other system components consume energy, their contributions do not vary significantly between models, and thus GPU-level energy measurements remain a meaningful comparative metric. The final spatiotemporal model's evaluation performance metrics used include Accuracy (Acc), Macro-F1, Weighted-F1, Macro-Recall, Weighted-Recall, and Kappa score.

3.2. Data preprocessing

Before model training, we applied a series of preprocessing steps to prepare the raw video data for spatial and temporal feature extraction. Each video clip in the DAiSEE dataset, originally recorded at 640 × 480 pixels for 10 seconds, was first decomposed into individual frames treated as static images. To ensure compatibility with the EfficientNet spatial models, all frames were resized according to the input resolution required by each EfficientNet variant, as summarized in Table 1. Pixel values were normalized to stabilize training and reduce sensitivity to illumination variations across recording environments. Given the pronounced class imbalance in the engagement labels, data augmentation was applied selectively to underrepresented engagement levels (levels 0 and 1). The augmentation techniques include salt noise, pixel transformation for brightness reduction, pepper noise, blurring effects, directional transformations (e.g., flipping), and color manipulation. As a result of augmentation, the number of videos in engagement level 0 increased to 800, and engagement level 1 to 1.000. For temporal modeling, each video was segmented into 300 temporal segments. From each segment, a predefined number of frames was uniformly sampled to represent the temporal evolution of visual features.

3.3. Training, Validation, Test Configuration

The pre-processed data sequences were split into training, validation, and testing sets to ensure robust model evaluation. Specifically, 70% of the video clips were allocated for training, 15% for validation, and the remaining 15% for testing, following a subject-independent splitting strategy to prevent data from the same individual appearing in multiple sets. The training was carried out over 7 epochs with a batch size of 16. Each video was decomposed into individual frames, and each frame was converted into a feature vector using EfficientNetB7 with ImageNet pretrained weights. EfficientNetB7 was used as a fixed backbone to extract these frame-level features, which were then arranged into temporal sequences to serve as inputs to the LSTM model. Feature extraction was performed prior to LSTM training, and only the LSTM and classification layers were trained. The LSTM was configured with 1,024 hidden units to capture temporal dependencies effectively. To mitigate overfitting, we applied a dropout rate of 0.1. Model optimization used the Root Mean Square Propagation (RMSProp) optimizer with a learning rate of 0.001. The output layer used the Softmax activation function, and the categorical cross-entropy loss guided training. Following training, the model was evaluated on a validation dataset, during which the parameter update process was temporarily paused. Predictions generated on the validation set were compared to the actual labels to compute evaluation metrics. Upon completion of the entire training and validation process, a final evaluation was performed using the test dataset, including metrics such as accuracy, precision, recall, and F1-score. The summary of the configuration is shown in Table 2.

Table 2. Experimental Setup and Hyperparameters

Parameter	Details
Epochs	7
Batch size	16
Hidden units	1024
Dropout	0.1
Optimizer	RMSProp
Learning rate	0.001

4. Data Analysis

The data analysis was assessed using standard metrics, including accuracy, F1-score, recall, and Kappa Index, or Cohen's Kappa *(k)* [6]. The k is used in this work because it evaluates agreement between the model's predictions and the actual labels and applies to both binary and multiclass classification problems. Unlike accuracy, k adjusts for chance agreement, offering a more precise evaluation of the model's performance. It is defined as:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

where *p0* is the proportion of observed agreements and *pc* is the proportion of agreements expected by chance. The interpretation k is shown in Table 3, based on a standardized scale with established benchmarks indicating various levels of agreement. The model's effectiveness was evaluated using the Kappa Energy Index (KEI), which quantifies the trade-off between predictive accuracy and energy efficiency as shown in Equation (2).

$$KEI = \frac{K}{E} \tag{2}$$

It is defined as the ratio of Cohen's Kappa coefficient (k), as shown in Equation (1), to the energy consumption (E), measured in kilowatt-hours (kWh). This metric combines accuracy and resource efficiency; models that achieve high accuracy with low energy consumption and training time receive higher KEI scores.

Table 3. Cohen's Kappa agreement level

Cohen's Kappa score	Params
k<0	No agreement
0.01<k<0.20	Slight
0.21<k<0.40	Fair
0.41<k<0.60	Moderate
0.61<k<0.80	Substantial
0.81<k<1.00	Almost Perfect

4. Results and Discussion

4.1. Performance Metrics for Each Model

The results of the model performance evaluation are presented in Fig. 2, which compares various evaluation metrics, including Accuracy, Macro-F1 (M-F1), Weighted-F1 (W-F1), Macro Recall (M-Recall), Weighted Recall (W-Recall), and Kappa. Analysis of the accuracy performance graph shows that model performance does not strictly follow EfficientNet's default complexity order on the ImageNet classification task, as indicated by the pink dashed line. Although model complexity increases from B0 to B7, the accuracy graph (blue line) shows an observed performance order of B7, B3, B6, B4, B2, B0, B5, and B1. Model B3 improves significantly, rising from fifth to second place, while B0 also improves from eighth to sixth. Conversely, B5 declines significantly, dropping from third to seventh place, while B6 and B1 decline slightly, and B7 and B4 maintain their positions.

In general, the graph shows that most models have higher M-F1 and M-Recall scores than W-F1 and W-Recall, indicating they maintain relatively stable performance across classes and handle potential class imbalances reasonably well. The F1-score and Recall scores range from 0.6–0.7, indicating that the models generally recognize the average class with adequate performance, while the weighted metrics range from 0.55–0.62, suggesting the class distribution is relatively balanced and not overly imbalanced. In terms of individual performance, models B7 and B3 stand out, achieving the highest accuracy, consistent with high M-F1 and M-Recall scores. This suggests that both models' accuracy is not solely due to recognizing the majority class but also reflects strong generalization across classes. Conversely, model B1 performed worst across both accuracy and class-average-based metrics, indicating limitations in capturing comprehensive data patterns. Interestingly, although models B4, B2, and B0 have identical accuracy values (0.594), the M-F1 and M-Recall graphs show quite clear differences in performance. This confirms that equal accuracy does not necessarily imply equal classification quality, underscoring the importance of macro metric-based evaluation. Based on the Kappa score graph, most models fall into the fair-to-moderate agreement category, with values ranging from 0.38 to 0.45. Model B7 achieved the highest score (0.445) and fell into the moderate agreement category. However, the Kappa difference between B7 (0.445) and B3 (0.437) is minimal, despite B7's significantly greater complexity and number of parameters. This suggests that B3 is a strong candidate model in terms of the balance between performance and complexity. Model B0 also has a Kappa value comparable to B4 and B2, although lower than B7 and B3. Considering its significantly lower computational requirements, B0 remains a model worthy of further exploration.

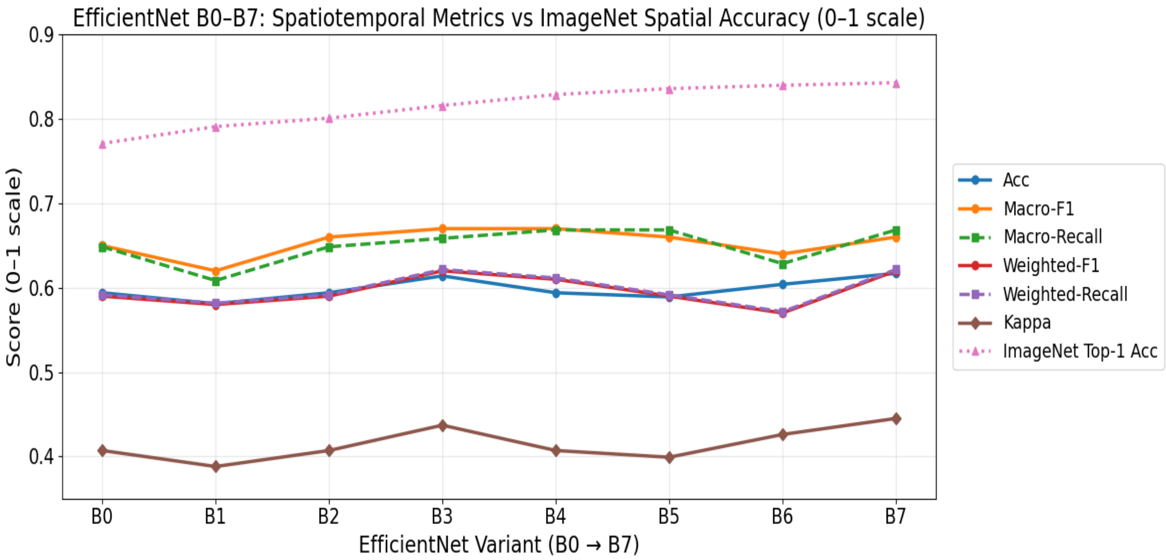


Fig. 2. Metric results of EfficientNet B0-B7 + LSTM models

4.2. Energy Consumption

Energy consumption monitoring results for eight workload profiles (B0–B7) show a clear correlation between execution duration and the resulting power stability. Based on Fig. 3, the power profiles fall into two main patterns: a degradative pattern (B0, B1, B2) and a constant pattern (B3, B5, B6, B7). In profiles B0 and B1, peak power ranges from 140W to 220W early in the duration, then gradually decreases (steps down) over time, stabilizing at a lower level before finally stopping. A similar trend is observed in B2, where power drops from 240W to 150W midway through the 132.79-minute run. The power values drop dramatically mid-execution for models B0, B1, and B2, likely due to a reduction in computational workload shortly after the models complete their computationally intensive preprocessing. This decrease is further influenced by the GPU's large memory capacity, which allows most data to reside in GPU RAM at the start of the computation.

Conversely, in profiles with longer durations, such as B5 (352.52 m), B6 (545.29 m), and B7 (894.02 m), the graphs show very high-power stability (steady-state). On B7, power consumption remained consistently near 370W for nearly the entire 14.9-hour runtime. This consistency indicates that the workload in this profile has a uniform computing intensity, with minimal load fluctuations. Some noise or minor variations are visible at the peaks of graphs B5 and B6, but overall, the lines remain horizontal, indicating a continuous power consumption at maximum load. A different characteristic is observed in profile B4 (402.6 m), which shows a low baseline power consumption of 80W but has two sharp spikes exceeding 110W between the 200th and 300th minutes. This pattern indicates a heterogeneous workload, with brief, computationally intensive phases amidst periods of low load. At the end of each test, all profiles exhibit a uniform power drop toward the idle level (20W).

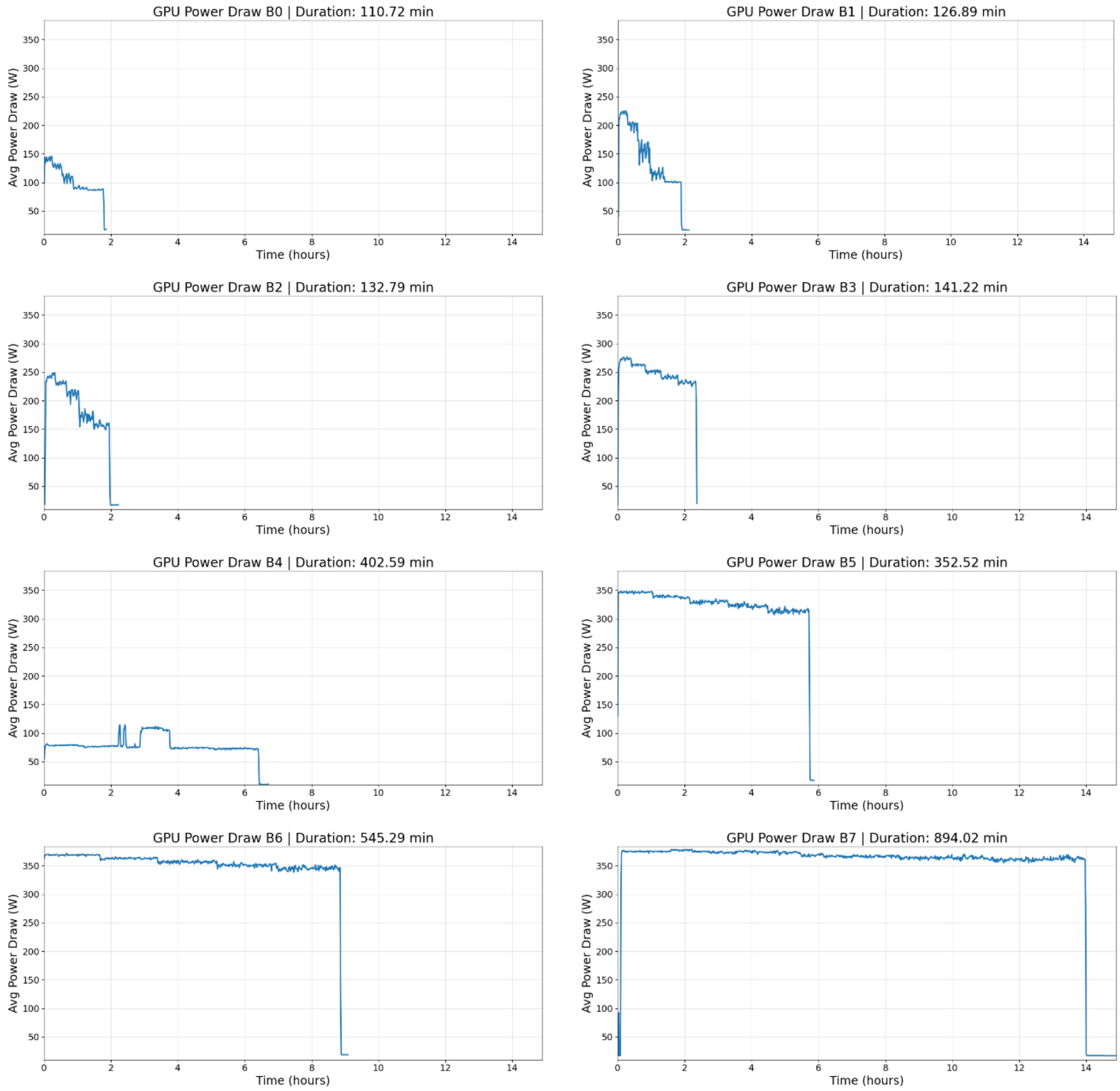


Fig. 3. GPU energy consumption over time for EfficientNet B0–B7 + LSTM models

Based on the duration summary graph shown in Fig. 4, there is a clear correlation between the EfficientNet model variants and the required execution time. The baseline variants B0 through B3 show relatively stable durations of 132.8 to 152.3 minutes, with the shortest recorded for variant B2. However, there is a significant spike in duration from variant B4 (495.1 minutes) to B7 (894.0 minutes). Interestingly, variant B4 shows a longer logging time (352.5 minutes) than B5, even though B5 is a larger model in the hierarchy. Overall, this graph shows an exponential increase in workload at high variants.

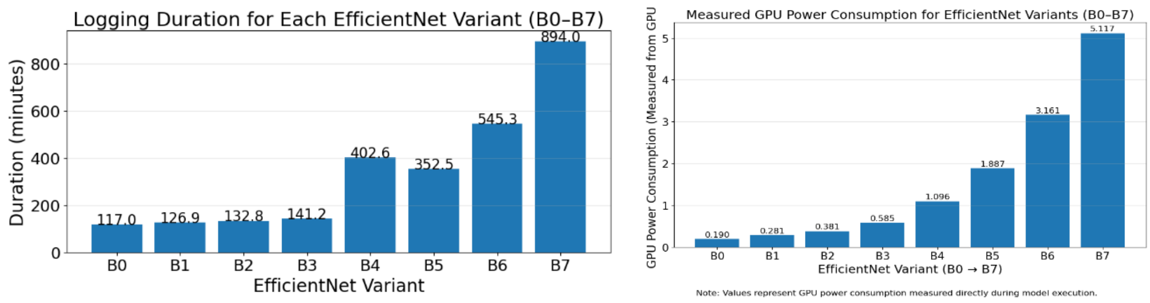


Fig. 4. GPU power consumption and KEI comparison

4.3. Kappa Energy Index and Training Time

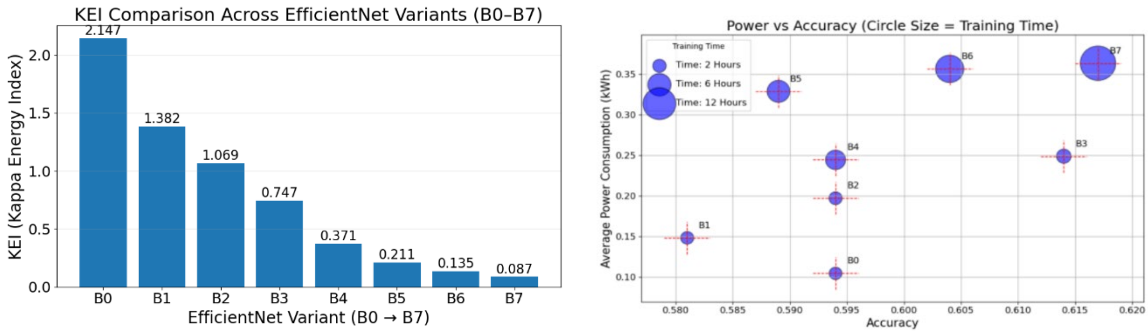
To assess computational efficiency, Kappa Energy Index (KEI) computed using GPU power consumption data. Figure 4 shows KEI values for all models (B0–B7). Highly complex models such as B7 tend to achieve the highest accuracy; however, they also require significantly more energy and longer training times, resulting in lower KEI values. In contrast, simpler models like B0 demonstrate much better energy efficiency. B0 achieves the highest KEI of 2.147, indicating it is the most energy-efficient model for predictive performance. However, its accuracy (59%) is slightly lower than that of the best-performing model, B7 (62%). Meanwhile, model B3, with a KEI value of 0.747, offers a well-balanced compromise between efficiency and performance. It achieves relatively high accuracy (around 61%) and even outperforms more complex models such as B4, B5, and B6. With an energy consumption of 0.586 kWh, both considered reasonable, B3 emerges as a strong candidate for systems that require efficiency without significantly reducing accuracy. However, B7 remains the preferred choice when maximum accuracy is critical, especially in high-stakes applications such as medical diagnosis or security systems. The accuracy achieved by EfficientNet-B7 (62%) on the DAiSEE dataset implies that, given the challenges of this dataset. For comparison, current research done by Selim et al. [57] shows the highest accuracy of around 67% on the same dataset as a benchmark. While these results may not be sufficient for real-world scenarios, they still provide valuable insights into temporal entanglement modeling and the relative capabilities of different EfficientNet architectures in terms of energy efficiency in the future. Further improvements in accuracy could include more aggressive data augmentation, ensemble modeling, or adding other modalities, such as physiological signals, to enhance practical applications.

4.4. Accuracy, Energy and Time Trade-off

To examine trade-offs among accuracy, training duration, and power consumption, Fig. 5 visualizes the relationships among these three factors across eight model variants, B0-B7. The scatter plot represents this multidimensional relationship, with classification accuracy on the horizontal axis, power consumption (kWh) on the vertical axis, and circle size indicating each model's required training time. In terms of training duration, models that required longer training times tended to exhibit higher classification accuracy. For example, model B6 required approximately 8.8 hours of training, while B7 took around 14 hours, both of which were among the most complex models and achieved high accuracy.

However, this trend is not entirely consistent. Model B3, for instance, required only 2.35 hours of training yet demonstrated competitive accuracy, even outperforming several more complex models. Regarding power consumption, models that consume more energy generally achieved better accuracy. Model B7, for instance, recorded the highest power consumption at approximately 5 kWh and also attained the highest accuracy among all models. In contrast, B0 consumed only about 0.380 kWh and had a very short training time of roughly 1.9 hours, making it suitable for real-time applications and devices with limited computational resources, such as mobile or IoT-based systems. Nonetheless, the relationship between power consumption and accuracy is not strictly linear. Regarding accuracy, there is no direct correlation between model complexity and classification perfor- mance. For instance, B3 outperformed several more

complex models such as B4, B5, and B6. This finding indicates that increasing the number of parameters or network depth does not necessarily lead to better accuracy, and that factors such as architectural suitability to the data and training strategies also play a significant role. When considering all three aspects simultaneously, a clear trade-off emerges. While B7 excels in accuracy, it comes at a high cost in terms of power consumption and training time. In contrast, B3 offers a well-balanced trade-off between efficiency and effectiveness, maintaining high accuracy while consuming low power and requiring short training time. Lightweight models such as B0 to B2, with training durations of approximately 1.8 to 1.9 hours, are well-suited for resource-constrained applications, even though with some compromise in accuracy. Overall, the graph underscores the complex, non-linear relationship among accuracy, training duration, and power consumption. Therefore, model performance evaluation should not rely on a single metric but instead consider a balanced trade-off between efficiency and effectiveness.



(a) (b)

Fig. 5. (a) Kappa Energy Index (KEI) computed from GPU energy consumption for EfficientNet variants (B0-B7). (b) Energy-accuracy trade-off in terms of training power consumption

5. Conclusion

This study highlights the effectiveness of deep learning models in classifying student engagement in online learning environments using facial video recordings. This research evaluated a series of EfficientNet models combined with LSTM. The proposed method utilized the DAiSEE dataset, which contains video clips annotated with 4 levels of student engagement. Each video was decomposed into individual frames to extract spatial features using a range of B0-B7 models, which were then processed temporally using an LSTM network. The evaluation was conducted using metrics such as accuracy, F1-score, and Cohen's Kappa, while accounting for energy consumption and training time. The results indicate that B7 achieved the highest classification performance, although it required substantial computational resources. In contrast, B0 demonstrated the best energy efficiency, with the highest Kappa Energy Index, making it more suitable for deployment in resource-constrained environments. Additionally, B3 offered a well-balanced trade-off between accuracy and computational efficiency. Sánchez-Mompó et al. [58] also analyzed energy consumption across various AI models, showing that the choice of architecture and hyperparameters can substantially impact energy consumption without necessarily degrading performance, in line with the trends observed in our KEI evaluation across different EfficientNet variants, where attempts to choose architectures that minimize weights or layer depth alone do not guarantee lower energy consumption. Overall, the findings underscore the importance of balancing effectiveness and energy efficiency in designing practical and sustainable AI systems for educational applications. The limitation of this work is that it uses only a single dataset, which may result in different results when applied to different environmental scenarios with different setups. Furthermore, the model applied in this study also focuses on the CNN-RNN architecture and does not compare non-deep learning models such as classical machine learning methods (e.g., SVM or Random Forest). Future work will be conducted by comparing these models. However, this result contributes to the development of Green AI and provides a guidance framework for selecting models that align with both performance requirements and infrastructure constraints.

Declarations

Author contribution. Conceptualization, P.S. and F.A.; methodology, P.S. and F.A.; software, F.A.; validation, P.S. and F.A.; writing original draft preparation, F.A.; writing review and editing, P.S.; visualization, F.A.; supervision, P.S.; project administration, P.S.; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding statement. None of the authors received funding or grants from any institution or funding body for this research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," *Neurocomputing*, vol. 599, p. 128096, Sep. 2024, doi: 10.1016/j.neucom.2024.128096.

[2] I. Yoon, J. Mun, and K.-S. Min, "Comparative Study on Energy Consumption of Neural Networks by Scaling of Weight-Memory Energy Versus Computing Energy for Implementing Low-Power Edge Intelligence," *Electronics*, vol. 14, no. 13, p. 2718, Jul. 2025, doi: 10.3390/electronics14132718.

[3] A. S. Luccioni And A. Hernandez-Garcia, "Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning," *arxiv Artif. Intell.*, pp. 1–19, 2023, [Online]. Available at: https://arxiv.org/pdf/2302.08476.

[4] S. M. Hasan, T. Islam, M. Saifuzzaman, K. R. Ahmed, C.-H. Huang, and A. R. Shahid, "Carbon Emission Quantification of Machine Learning: A Review," *IEEE Trans. Sustain. Comput.*, vol. 10, no. 6, pp. 1085–1102, Nov. 2025, doi: 10.1109/TSUSC.2025.3578834.

[5] P. Jiang, C. Sonne, W. Li, F. You, and S. You, "Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots," *Engineering*, vol. 40, pp. 202–210, Sep. 2024, doi: 10.1016/j.eng.2024.04.002.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Modern Deep Learning Research," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 09, pp. 13693–13696, Apr. 2020, doi: 10.1609/aaai.v34i09.7123.

[7] S. Dash, "Green AI: Enhancing Sustainability and Energy Efficiency in AI-Integrated Enterprise Systems," *IEEE Access*, vol. 13, pp. 21216–21228, 2025, doi: 10.1109/ACCESS.2025.3532838.

[8] Y. Yu *et al.*, "Revisit the environmental impact of artificial intelligence: the overlooked carbon emission source?," *Front. Environ. Sci. Eng.*, vol. 18, no. 12, p. 158, Dec. 2024, doi: 10.1007/s11783-024-1918-y.

[9] C. He, F. Yue, L. Li, Y. Tang, Q. Wu, and W. Chen, "AI carbon footprint: The non-negligible hidden emission source," *Eco-Environment Heal.*, vol. 4, no. 4, p. 100197, Dec. 2025, doi: 10.1016/j.eehl.2025.100197.

[10] S.-Y. Wang and N.-Z. Ye, "Invisible footprints, visible insights: machine learning reveals Scope 3 emissions," *Front. Sustain.*, vol. 6, pp. 1–13, Sep. 2025, doi: 10.3389/frsus.2025.1649150.

[11] A. de Vries, "The growing energy footprint of artificial intelligence," *Joule*, vol. 7, no. 10, pp. 2191–2194, Oct. 2023, doi: 10.1016/j.joule.2023.09.004.

[12] M. F. Argerich and M. Patiño-Martínez, "Measuring and Improving the Energy Efficiency of Large Language Models Inference," *IEEE Access*, vol. 12, pp. 80194–80207, 2024, doi: 10.1109/ACCESS.2024.3409745.

[13] R. Zhang and A. C. S. Chung, "EfficientQ: An efficient and accurate post-training neural network quantization method for medical image segmentation," *Med. Image Anal.*, vol. 97, p. 103277, Oct. 2024, doi: 10.1016/j.media.2024.103277.

[14] H. Cheng, M. Zhang, and J. Q. Shi, "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10558–10578, Dec. 2024, doi: 10.1109/TPAMI.2024.3447085.

[15] Q. Li, Z. Chen, Y. Li, Z. Jiang, and S. Cao, "Efficient network compression via gradient-score aware pruning," *Neurocomputing*, vol. 660, p. 131870, Jan. 2026, doi: 10.1016/j.neucom.2025.131870.

[16] X. Zhuang *et al.*, "Multi-objective optimization of reservoir development strategy with hybrid artificial intelligence method," *Expert Syst. Appl.*, vol. 241, p. 122707, May 2024, doi: 10.1016/j.eswa.2023.122707.

[17] B. Tian, Y. Pang, M. Huzaifa, S. Wang, and S. Adve, "Towards Energy-Efficiency by Navigating the Trilemma of Energy, Latency, and Accuracy," in *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2024, pp. 913–922, doi: 10.1109/ISMAR62088.2024.00107.

[18] Y. Shang, S. Zhou, D. Zhuang, J. Żywiołek, and H. Dincer, "The impact of artificial intelligence application on enterprise environmental performance: Evidence from microenterprises," *Gondwana Res.*, vol. 131, pp. 181–195, Jul. 2024, doi: 10.1016/j.gr.2024.02.012.

[19] S. Lu *et al.*, “Spatiotemporal Feature Learning for Daily-Life Cough Detection Using FMCW Radar,” *Bioengineering*, vol. 12, no. 10, p. 1112, Oct. 2025, doi: 10.3390/bioengineering12101112.

[20] M. Wang, J. Xing, J. Su, J. Chen, and L. Yong, “Learning SpatioTemporal and Motion Features in a Unified 2D Network for Action Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2022, doi: 10.1109/TPAMI.2022.3173658.

[21] Y. Ma, H. Lou, M. Yan, F. Sun, and G. Li, “Spatio-temporal fusion graph convolutional network for traffic flow forecasting,” *Inf. Fusion*, vol. 104, p. 102196, Apr. 2024, doi: 10.1016/j.inffus.2023.102196.

[22] W. Chen *et al.*, “Spatio-temporal characteristics and influencing factors of traditional villages in the Yangtze River Basin: a Geodetector model,” *Herit. Sci.*, vol. 11, no. 1, p. 111, May 2023, doi: 10.1186/s40494-023-00948-x.

[23] B. Olczak, M. Wilkosz-mamcarczyk, B. Prus, K. Hodor, and R. Dixon-gough, “Application of the building cohesion method in spatial planning to shape patterns of the development in a suburban historical landscape of a ‘village within Kraków,’” *Land use policy*, vol. 114, p. 105997, Mar. 2022, doi: 10.1016/j.landusepol.2022.105997.

[24] R. Das and S. Dev, “Optimizing student engagement detection using facial and behavioral features,” *Neural Comput. Appl.*, vol. 37, no. 23, pp. 19063–19085, Aug. 2025, doi: 10.1007/s00521-025-11317-z.

[25] I. Qarbal, N. Sael, and S. Ouahabi, “Student’s Engagement Detection Based on Computer Vision: A Systematic Literature Review,” *IEEE Access*, vol. 13, pp. 140519–140545, 2025, doi: 10.1109/ACCESS.2025.3596885.

[26] T. Shan, S. Feng, K. Li, R. Chang, and R. Huang, “Unveiling the effects of artificial intelligence and green technology convergence on carbon emissions: An explainable machine learning-based approach,” *J. Environ. Manage.*, vol. 373, p. 123657, Jan. 2025, doi: 10.1016/j.jenvman.2024.123657.

[27] R. Ranpara, “Energy-efficient green AI architectures for circular economies through multi-layered sustainable resource optimization framework,” *Discov. Sustain.*, vol. 6, no. 1, p. 1031, Oct. 2025, doi: 10.1007/s43621-025-01846-x.

[28] S. Khan *et al.*, “Green AI techniques for reducing energy consumption in AI systems,” *Array*, vol. 29, p. 100652, Mar. 2026, doi: 10.1016/j.array.2025.100652.

[29] T. Yigitcanlar, R. Mehmood, and J. M. Corchado, “Green Artificial Intelligence: Towards an Efficient, Sustainable and Equitable Technology for Smart Cities and Futures,” *Sustainability*, vol. 13, no. 16, p. 8952, Aug. 2021, doi: 10.3390/su13168952.

[30] R. Verdecchia, J. Sallou, and L. Cruz, “A systematic review of Green AI,” *WIREs Data Min. Knowl. Discov.*, vol. 13, no. 4, p. e1507, Jul. 2023, doi: 10.1002/widm.1507.

[31] R. Verdecchia, P. Lago, C. Ebert, and C. de Vries, “Green IT and Green Software,” *IEEE Softw.*, vol. 38, no. 6, pp. 7–15, Nov. 2021, doi: 10.1109/MS.2021.3102254.

[32] O. O. Oyewole and J. F. Joseph, “Sustainable AI and Green Computing: Reducing the Environmental Impact of Large-Scale Models with Energy-Efficient Techniques,” *Int. J. Sci. Res. Netw. Secur. Commun.*, vol. 13, no. 3, pp. 19–26, Jun. 2025, doi: 10.26438/ijsrnsc.v13i3.276.

[33] R. Z. Khudhur and M. A. Mohammed, “A Spatio-Temporal Deep Learning Approach for Efficient Deepfake Video Detection,” *ARO-THE Sci. J. KOYA Univ.*, vol. 13, no. 2, pp. 75–82, Aug. 2025, doi: 10.14500/aro.12190.

[34] Z. Zhang *et al.*, “PromptST: Prompt-Enhanced Spatio-Temporal Multi-Attribute Prediction,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Oct. 2023, pp. 3195–3205, doi: 10.1145/3583780.3615016.

[35] J. Wu, Z. Niu, X. Li, L. Huang, P. S. Nielsen, and X. Liu, “Understanding multi-scale spatiotemporal energy consumption data: A visual analysis approach,” *Energy*, vol. 263, p. 125939, Jan. 2023, doi: 10.1016/j.energy.2022.125939.

[36] G. Zhou, O. Soufan, J. Ewald, R. E. W. Hancock, N. Basu, and J. Xia, "NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W234–W241, Jul. 2019, doi: 10.1093/nar/gkz240.

[37] C.-W. Shen, J.-W. Jiang, and H.-P. Hsieh, "STGym: A Modular Benchmark for Spatio-Temporal Networks With a Survey and Case Study on Traffic Forecasting," *IEEE Trans. Big Data*, vol. 12, no. 1, pp. 15–33, Feb. 2026, doi: 10.1109/TBDATA.2025.3618482.

[38] Z. Ma, C. Chen, J. Chen, and Y. Jiang, "STC-SORT: A Dynamic Spatio-Temporal Consistency Framework for Multi-Object Tracking in UAV Videos," *Appl. Sci.*, vol. 16, no. 2, p. 1062, Jan. 2026, doi: 10.3390/app16021062.

[39] T. Fütterer *et al.*, "Artificial intelligence in classroom management: A systematic review on educational purposes, technical implementations, and ethical considerations," *Comput. Educ. Artif. Intell.*, vol. 9, p. 100483, Dec. 2025, doi: 10.1016/j.caeai.2025.100483.

[40] A. T. S. and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1387–1415, Mar. 2020, doi: 10.1007/s10639-019-10004-6.

[41] M. Alruwaili and M. Mohamed, "An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification," *Diagnostics*, vol. 15, no. 5, p. 551, Feb. 2025, doi: 10.3390/diagnostics15050551.

[42] W. Lu, Y. Yang, R. Song, Y. Chen, T. Wang, and C. Bian, "A Video Dataset for Classroom Group Engagement Recognition," *Sci. Data*, vol. 12, no. 1, p. 644, Apr. 2025, doi: 10.1038/s41597-025-04987-w.

[43] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," *J. LATEX Cl. FILES*, vol. 14, no. 8, pp. 1–14, 2015, [Online]. Available at: https://arxiv.org/abs/1609.01885.

[44] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019. [Online]. Available at: https://arxiv.org/abs/1905.11946v5.

[45] S. R. Iyer, "Green AI: Energy-Efficient Training of Large-Scale Models," *Int. J. Res. Appl. Innov.*, vol. 8, no. 2, pp. 11947–11951, Mar. 2025. [Online]. Available at: https://www.ijrai.org/index.php/ijrai/article/view/85.

[46] R. Su, L. He, and M. Luo, "Leveraging part-and-sensitive attention network and transformer for learner engagement detection," *Alexandria Eng. J.*, vol. 107, pp. 198–204, Nov. 2024, doi: 10.1016/j.aej.2024.06.074.

[47] R. Das and S. Dev, "Enhancing frame-level student engagement classification through knowledge transfer techniques," *Appl. Intell.*, vol. 54, no. 2, pp. 2261–2276, Jan. 2024, doi: 10.1007/s10489-023-05256-2.

[48] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[49] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–36, Jul. 2020, doi: 10.1145/3343440.

[50] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv Comput. Vis. Pattern Recognit.*, pp. 1–9, Apr. 2017. [Online]. Available at: https://arxiv.org/abs/1704.04861v1.

[51] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, pp. 1–22, 2021, [Online]. Available at: https://arxiv.org/pdf/2010.11929.

[52] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[53] M. Krichen and A. Mihoub, "Long Short-Term Memory Networks: A Comprehensive Survey," *AI*, vol. 6, no. 9, p. 215, Sep. 2025, doi: 10.3390/ai6090215.

[54] M. M. Taye, “Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions,” *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.

[55] T. T. Kieu Tran, T. Lee, J. Y. Shin, J. S. Kim, and M. Kamruzzaman, “Deep learning-based maximum temperature forecasting assisted with meta-learning for hyperparameter optimization,” *Atmosphere (Basel).*, vol. 11, no. 5, pp. 1–21, 2020, doi: 10.3390/ATMOS11050487.

[56] S. Aquino-Brítez, P. García-Sánchez, A. Ortiz, and D. Aquino-Brítez, “Towards an Energy Consumption Index for Deep Learning Models: A Comparative Analysis of Architectures, GPUs, and Measurement Tools,” *Sensors*, vol. 25, no. 3, p. 846, Jan. 2025, doi: 10.3390/s25030846.

[57] T. Selim, I. Elkabani, and M. A. Abdou, “Students Engagement Level Detection in Online e-Learning Using Hybrid EfficientNetB7 Together With TCN, LSTM, and Bi-LSTM,” *IEEE Access*, vol. 10, pp. 99573–99583, 2022, doi: 10.1109/ACCESS.2022.3206779.

[58] A. Sánchez-Mompó, I. Mavromatis, P. Li, K. Katsaros, and A. Khan, “Green MLOps to Green GenOps: An Empirical Study of Energy Consumption in Discriminative and Generative AI Operations,” *Information*, vol. 16, no. 4, p. 281, Mar. 2025, doi: 10.3390/info16040281.