

Scene-Level Recognition of Classroom Operational States Using VGG16 Transfer Learning

Ayesha Khattak^{*1}, Syed Asim Jalal², Laiba Awan³, Muhammad Haseeb⁴, Muhammad Sadeequllah⁵, Saif Ur Rehman⁶

¹ School of Computer Science and IT, Institute of Management Science, Peshawar, Pakistan
^{2,3,4,5,6} Department of Computer Science, University of Peshawar, Peshawar, Pakistan

¹ ayesha.khattak@imsciences.edu.pk

DOI: <https://doi.org/>

Keywords

Classroom Operational State Recognition; Computer Vision; Deep Learning; VGG16; Transfer Learning; Smart Classroom; Educational Technology

Article History

Received on 20 Dec 2025

Accepted on 20 Jan 2026

Published on 30 Jan 2026

Copyright @Author

Corresponding Author: *

Ayesha Khattak

School of Computer Science and IT, Institute of Management Science, Peshawar, Pakistan

ayesha.khattak@imsciences.edu.pk

Abstract

Automatic recognition of classroom operational states can be useful for classroom availability, room-use monitoring, and academic administration. This paper investigates whether three general classroom states can be recognized using a deep learning-based computer vision approach. The three states are lecture in progress, students present but the teacher absent, and empty classroom. The dataset was collected locally and contains about 1,200 images, with nearly 400 images in each class. The images were preprocessed and then used to train a VGG16 convolutional neural network. The network was initialized with ImageNet weights and adapted for three output classes. The main experiment used an 80:20 training and testing split. On 240 test images, the model obtained an accuracy of 97.08%. The F1-score was 0.96 for lecture in progress, 0.96 for students present without the teacher, and 0.99 for the empty classroom. Most errors occurred between the two occupied classroom states. The empty classroom was identified more accurately. A small web application was also developed to demonstrate how a classroom image can be submitted to the model and a state prediction returned. The results are promising for the dataset, but they should be considered only as a proof of concept. The images were collected from more than one classroom, but no independent external classroom dataset was used for validation. Therefore, the study gives an initial basis for future work on classroom availability, room-use monitoring, and simple operational monitoring.

INTRODUCTION

Digital technology is now widely used in teaching environments. Smart classrooms may use cameras, sensors, artificial intelligence, and data analysis to observe different conditions inside a classroom. Such systems can support attendance, classroom management, engagement analysis, environmental control, and better use of university resources. Recent reviews also show that computer vision is becoming common in smart classroom research because visual data can provide information about people, movement, posture, and the general classroom condition [1]. For this reason, computer vision can support both learning-related analysis and practical classroom management.

A basic but useful task is to identify the current operational condition of a classroom. A timetable can show that a room has been assigned for a class, but it cannot always show what is happening inside the room at that moment. The room may be empty, students may be waiting for the teacher, or teaching may already be taking place. This status information can provide useful data that can be analyzed for different purposes. This can help identify the availability of a room in real time. The data can be used to support more effective management of limited teaching spaces. The same information can provide data on scheduled lectures that have not taken place. In many institutions this information is gathered manually by staff. For large educational institutions, a simple camera-based system can provide a lightweight alternative, especially using existing cameras if already installed.

Most educational computer vision studies have focused on more detailed tasks. Common examples are recognition of teacher and student activities, student engagement, attendance, facial expressions, and body posture. The EduNet dataset was introduced for classroom activity recognition and contains 20 teacher and student activity classes taken from classroom

videos [2]. Sharma et al. later proposed the STAR-3D method and reported 83.5% accuracy on EduNet [3]. Rajamanickam et al. used hybrid 3D convolutional models and reported an average accuracy of 88.17% on EduNet. Their study also included testing on another classroom video set [4]. These studies show that automatic understanding of classroom activity is receiving increasing attention.

Detailed activity recognition usually depends on video because many actions can only be understood over time. It may also need larger labelled datasets and more complex models. This is useful when the aim is to study a particular teacher or student behaviour. However, such detail is not always necessary for a simple classroom management task. A room may only need to be classified as empty, occupied without an active lecture, or currently used for teaching. This can be treated as a scene-level image classification problem, where the output describes the operational state of the whole classroom. In such a case, one classroom image may provide enough information for a useful decision, and there is no need to identify individual students or teachers.

Another related direction is the automatic analysis of student engagement from classroom images and videos. Ikram et al. used a transfer-learning-based VGG16 model to classify engagement states in a real classroom environment and reported 90% accuracy [5]. This work is important for the present study because it shows that a pretrained image model can be adapted to classroom visual data. At the same time, engagement recognition and classroom-state recognition have different purposes. Engagement studies try to infer the behaviour or affective condition of students. The present work instead focuses on the overall operational state of the classroom. The required output is therefore simpler and more suitable for applications related to classroom availability and room-status monitoring.

Deep learning models normally perform better when a large amount of labeled data is available. In small applied projects, collecting and labeling a very large image dataset may not be possible. Transfer learning provides one practical way to work under this limitation. A model that has already learned general visual features from a large source dataset can be adapted to a smaller target task. Recent reviews continue to identify transfer learning as an important approach for image classification when target data are limited [6], [7]. In the present study, VGG16 was selected as the base convolutional neural network. VGG16 is a well-established deep architecture that demonstrated the value of increasing convolutional network depth using small 3×3 filters and has been widely reused as a pretrained visual feature model [8]. Although newer architectures are available, VGG16 remains suitable for a proof-of-concept study because it is well understood and can be directly adapted to a three-class image classification task.

Based on this background, this study investigates whether a VGG16-based model can distinguish three broad classroom states from still images collected in university classrooms. The three states are: (1) lecture in progress, (2) students present while the teacher is absent, and (3) empty classroom. The dataset contains approximately 1,200 images, with about 400 images representing each state. The main evaluation uses an 80:20 training and testing division. Performance is examined using overall accuracy and class-level precision, recall, and F1-score. A confusion matrix is also used to examine which classroom states are more difficult for the model to distinguish.

The study is guided by the following research questions.

RQ1: To what extent can a VGG16-based vision model distinguish three classroom operational states from still images collected in university classrooms?

RQ2: How does classification performance vary across the three classroom operational states?

The aim is not to propose a new deep learning architecture or to claim a general solution for classroom activity recognition. The aim is to examine the feasibility of a simple vision-based approach for coarse classroom-state recognition in the studied setting.

This work has a practical focus. It treats classroom monitoring as a three-class image-classification problem based on common classroom situations. The VGG16 model was initialized with ImageNet weights and tested on the locally collected images. Results are reported for the complete test set and also for each class. Furthermore, a lightweight web interface was linked with the trained model so that an image can be submitted and the predicted classroom state can be shown.

The remainder of the paper is organized as follows. Section 2 presents related work on classroom activity recognition and the use of transfer learning for images classification. Section 3 describes the research methodology. Section 4 presents the results of the experiments, while section 5 discusses the findings and identifies the major limitations of the study. Section 6 identifies some future directions and concludes the paper.

Related Work

Research on smart classroom technologies is no longer limited to basic automation in classrooms. Recent smart classroom research involves cameras, sensors, connected learning platforms, and the use of artificial intelligence to observe different contexts and activities in the classroom. Zhang et al. [1] reported uses of sensors and AI for attendance, engagement, environmental monitoring, and personalized learning. Llurba et al. [13] reviewed studies published from 2021 to 2026 and noted that most work gives stronger attention to technical aspects than to pedagogy or the physical classroom setting. These studies show that smart classroom research is expanding, but the aims of individual systems can be very different.

The present study is related to this wider area, but its purpose is narrower. It does not attempt

to measure learning quality, emotion, or detailed student behaviour. It focuses on the overall operational state of a classroom. For this reason, the most relevant literature can be grouped into three areas: computer vision in smart classrooms, classroom activity and behaviour recognition, and transfer learning for image-based classroom analysis.

Computer Vision in Smart Classrooms

Computer vision has become one of the main technical approaches for automatic classroom analysis. A camera can provide information about the number of people, their positions, movements, gestures, and the general condition of the room. This has encouraged work on attendance, student attention, engagement, teacher behaviour, and classroom interaction. A recent systematic review by Liu et al. [9] examined 80 studies on computer-vision-based classroom behaviour recognition. The review grouped the main recognition targets into physical actions, learning engagement, attention, and emotion. It also reported that deep learning has become the main recognition approach in this area. At the same time, the review identified continuing problems related to experimental design, practical deployment, and the connection between recognized behaviour and educational meaning.

Fütterer et al. [12] reviewed 104 studies on AI for classroom management. Attention, behaviour, and engagement were among the common topics. Many of the reviewed studies used image or video data as input to deep learning models. This shows that computer vision has an important place in classroom management research. The review also reported that ethical issues were not always discussed in sufficient detail. This is important because classroom images may include identifiable students and teachers even when the system does not perform identity recognition. Therefore, practical use of such systems requires clear rules for collecting, storing, accessing, and using visual data.

Many recent computer vision studies in classrooms focus on the behaviour or learning

state of individual persons. This is suitable when the aim is to study engagement, attention, posture, or teaching behaviour. In some cases, however, a university only needs simple room-level information. It may need to know whether a room is empty, whether students are waiting for a teacher, or whether a lecture is taking place. This is closer to classroom-state or scene classification than detailed behaviour recognition. The present study is based on this simpler problem.

Classroom Activity and Behaviour Recognition

A major direction in classroom computer vision is the recognition of teacher and student activities from video. The EduNet dataset introduced by Sharma et al. [2] has supported much of the recent work in this area. EduNet contains 7,851 manually annotated video clips covering 20 teacher and student activities. The clips were collected from online videos and real classroom settings, with a total duration of about 12 hours. Using an I3D-ResNet-50 baseline, the authors reported an accuracy of 72.3%. The importance of EduNet is not only its size. It also provides a common classroom-specific dataset for testing models on actions such as hand raising, writing, reading, talking, walking, and explaining a subject.

Later studies improved activity recognition on this dataset using more complex spatiotemporal models. Sharma et al. [3] proposed STAR-3D, which uses 3D convolutional processing for student and teacher action recognition. The model reported an average accuracy of 83.5% on the 20 EduNet activity classes. Yuvaraj et al. [11] used a long-term recurrent convolutional network to model classroom video. Their system reported a mean accuracy of 93.17% on an EduNet subset under five-fold cross-validation and 83.33% on an independent set of classroom videos. Rajamanickam et al. [4] later combined 3D-CNN features with different machine learning classifiers. Their 3D-CNN with an extreme learning machine achieved an average accuracy of 88.17% on EduNet and 80.00% on an independent video set. They also

used Grad-CAM to examine which visual regions contributed to the predictions.

Other work has examined broader instructional activities instead of individual actions. Foster et al. [10] analyzed nearly 250 hours of elementary classroom videos. Their neural-network-based approach was used to identify instructional activity types such as whole-class activity, small-group activity, individual activity, and transition. For the four common activity types, one of their models achieved more than 80% accuracy. This work shows that classroom computer vision can also be used at a higher level than individual gestures. However, the analysis still depends on classroom video and is designed to identify instructional activity over time.

These studies provide useful evidence that deep learning can recognize classroom actions and instructional patterns. They also show why video is commonly used for fine-grained activity recognition. Actions such as walking, writing, explaining, and hand raising contain temporal information. A sequence of frames can therefore provide information that may not be available in one image. The cost is that video models are normally more complex and require larger annotated datasets and greater computation. For an operational room-state problem, this level of temporal modeling may not always be necessary.

Table 1. Representative recent studies on vision-based classroom analysis

Study	Focus / Dataset	Approach	Main result and relevance
Sharma et al. [2], 2021	20 teacher/student activities	3D-ResNet-50, EduNet, 7,8!	72.3% accuracy; established a classroom specific benchmark dataset.

Study	Focus / Dataset	Approach	Main result and relevance
Ikram et al. [5], 2023	Student engagement in a classroom; engaged & non-engaged	VGG16 transfer learning	90% accuracy; showed that pretrained image models can be adapted to classroom visual data.
Sharma et al. [3], 2024	20 student activities	STAR-3D, EduNet	83.5% average recognition accuracy.
Foster et al. [10], 2024	Instructional activity from nearly 250 hours of classroom video	Video-based neural network	More than 80% accuracy for common instructional activity types.
Yuvaraj et al. [11], 2024	Teacher and student activities using EduNet subset as independent videos	LRCN	93.17% mean accuracy on EduNet subset; 83.33% on independent videos.
Jia and Li [14], 2024	Student classroom behaviours in complex scenes	YOLOv8 + attention + OpenPose	Maximum mAP 82.1%; focused on individual student behaviour and posture.
Rajamanickam et al. [4], 2025	Teacher and student activities; EduNet as independent videos	3D-CNN + ELM	88.17% on EduNet; 80.00% on independent videos.

Study	Focus / Data	Approach	Main result and relevance
Sheng et al. [15], 2025	Student behaviours including hand raising, reading, and writing	Improved YOLOv8s	mAP 76.5% and 95.0% on two classroom behaviour datasets.

Image-Based Behaviour Recognition and Transfer Learning

Not all classroom analysis requires video. Several studies use individual frames or still images to detect behaviour, posture, or engagement. Ikram et al. [5] used a VGG16 transfer-learning model to classify students as engaged or non-engaged in a real classroom setting. Their dataset contained 3,000 images, and the reported classification accuracy was 90%. This study is especially relevant because it demonstrates that a pretrained VGG16 model can learn useful information from classroom images when the target problem is defined at the image level. However, the output in that study was an engagement state of students rather than the operational state of the whole classroom.

Recent behaviour-detection studies have also used object-detection and pose-estimation models. Jia and He [14] combined YOLOv5, a contextual attention mechanism, and OpenPose to recognize student classroom behaviour. Their model reported a maximum mean average precision of 82.1%. Sheng et al. [15] proposed an improved YOLOv8s framework for real-time classroom behaviour detection. Their experiments reported mAP values of 76.5% and 95.0% on two classroom datasets. These models are useful when the system must locate students and identify behaviours such as reading, writing, or hand raising. They also show the technical difficulty caused by occlusion, small targets, similar poses, and crowded classroom scenes.

Transfer learning is useful in applied image-classification problems because collecting a very large labeled dataset is often difficult.

Iman et al. [7] reviewed deep transfer learning and described its main purpose as reusing knowledge learned from a source task to support a target task with less data or lower training cost. A recent survey focused specifically on image classification reached a similar conclusion. Plested et al. [6] noted that the success of transfer learning depends on factors such as the size of the target dataset and the relationship between the source and target domains. These points are relevant for classroom image analysis because classroom datasets are often locally collected and smaller than general-purpose vision datasets.

VGG16 remains a common pretrained backbone in applied computer vision. The original VGG work by Simonyan and Zisserman [8] showed that deeper convolutional networks using small 3 x 3 filters could learn strong visual representations. Newer architectures can be more efficient or can achieve better performance on large benchmark datasets. However, the purpose of the present work is not to compare modern architectures. VGG16 was used as a classification model in this study. The paper therefore evaluates the practical result of that implementation and treats VGG16 as an established model rather than as a methodological contribution.

Research Gap and Position of the Present Study

The reviewed studies show strong progress in computer vision for classrooms. However, most recent work is directed towards one of three tasks. The first is fine-grained activity recognition from video, where temporal models identify actions of teachers and students [2]-[4], [11]. The second is detection of individual

student or teacher behaviours, often using YOLO-based methods and pose information [14], [15]. The third is recognition of learning-related states such as engagement [5]. These tasks are important, but they are more detailed than the problem considered in this paper.

Comparatively less attention has been given to coarse classroom-state recognition from a single image. In many operational situations, the required information may simply be whether a room is empty, whether students are present but teaching has not started, or whether a lecture is in progress. These classes describe the state of the whole classroom rather than a specific action of one person. They may therefore be recognized from scene-level visual information such as occupancy, presence of a teacher, arrangement of students, and the overall teaching context. A still-image classifier also avoids the need to model temporal motion when that information is not required by the task.

The present study addresses this narrower problem using a locally collected image dataset containing three classroom operational states and applies a VGG16-based classifier to

distinguish them. The contribution is therefore application focused. It does not claim a new deep learning architecture and it does not attempt to compete with recent video-based activity-recognition systems. Instead, it examines whether a standard pretrained vision model can provide useful classroom-state classification in a limited real-world setting. This position is consistent with the proof-of-concept nature of the study and with the evaluation.

This gap also defines the limits of the study. Recognition of three broad states cannot provide information about student attention, engagement, learning quality, or individual behaviour. The images were collected from more than one classroom, but the study does not provide a controlled test on a completely unseen classroom environment. Generalization to different room layouts, camera positions, lighting conditions, and institutions therefore remains uncertain. These limitations are considered directly in the later discussion. The value of the work is in demonstrating a simple and practical form of classroom visual classification that can provide a basis for more extensive future systems.

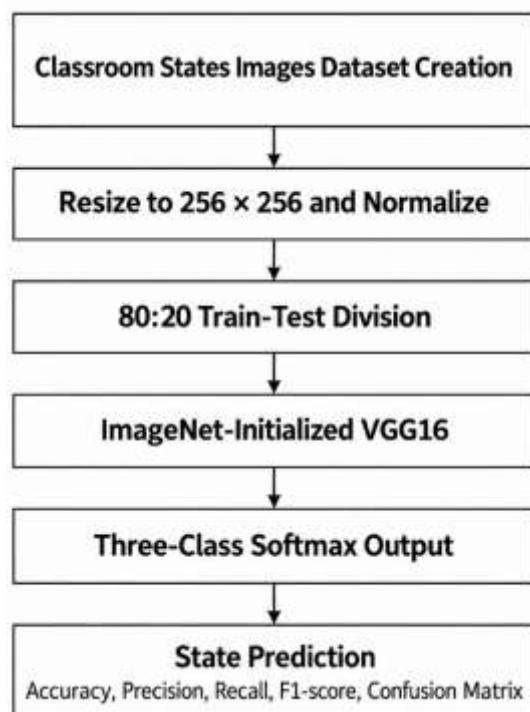


Figure 1. Workflow of the classroom-state

Recognition Approach

Methodology

This paper follows an applied proof-of-concept design for classroom-state recognition. It checks whether a deep image classifier can distinguish three broad classroom conditions from still images. The workflow follows a simple image-classification process. Classroom images are assigned to one of three states. Each image is resized and normalized before it is used by the model. The model is VGG16 initialized with ImageNet weights and changed to produce three output classes. Its output describes the state of the classroom as a whole, not the identity or action of one person. The trained model was also linked with a small web interface that accepts an image and displays the predicted state.

Figure 1 shows the main steps of the classification approach. The approach is based

on whole-scene classification. It does not use face recognition, person identification, or detailed recognition of individual student actions. This matches the purpose of the study, which is only to identify whether a lecture is taking place, students are waiting for the teacher, or the room is empty.

The next step was the preparation of the classroom image dataset used in this workflow. The images were collected manually in university classrooms using a camera. Different classroom conditions were recorded before, during, and between teaching sessions. The images were then divided into three classes. The first class is lecture in progress, where the teacher is present and teaching is taking place. The second class contains students in the classroom while the teacher is absent. The third class represents a fully empty classroom. Figure 2 one representative image for each of the three states.

Each class contained approximately 400 images. The complete dataset therefore contains 1,200 images, and the three classes are nearly balanced. Table 2 gives the meaning of each class. The images were collected from more than one university classroom. The evaluation was performed on the combined dataset rather than separately for each classroom.

After organizing the images into the three classes, the images were prepared for model training and testing. The implementation reads the images using OpenCV and resizes every image to 256×256 pixels. The image arrays are then converted to numerical form and normalized by dividing pixel values by 255. This scales the pixel values to the range from 0 to 1. The class labels are converted to one-hot encoded vectors because the network uses a three-class softmax output and categorical cross-entropy loss.

For the main evaluation, the dataset was divided into an 80% training portion and a 20% held-out test portion using the train-test split procedure. The test set contained 240 images. During model fitting, 20% of the training portion was further used as validation data. The evaluation used a single holdout split rather

than repeated cross-validation. For this reason, the results in this paper are interpreted as proof-of-concept results for the studied dataset and not as a precise estimate of performance across different classrooms.



Figure 2. Representative samples from the classroom image dataset: (a) lecture in progress, (b) lecture not in progress, and (c) empty classroom.

Table 2. Classroom operational-state classes used in the study

Class	Operational meanin	Images
Lecture progress	Teacher and studer are present and t lecture is taking place	400
Students prese / teacher absen	Students are present the classroom and a waiting for the teach or for teaching to star	400
Empty classroom	No teacher or studer are present in t classroom.	400

After preprocessing and data division, the VGG16 model was configured for the three-class classification task. The classifier uses VGG16 as its convolutional base. VGG16 is a deep convolutional neural network originally developed for large-scale image recognition [8]. In this implementation, the network is initialized with weights learned from ImageNet. The original VGG16 classification head is not included. The output feature map of the convolutional network is flattened and connected to a dense output layer with three neurons. A softmax activation is used to produce a probability for each classroom state.

During model fitting the VGG16 layers were kept trainable. The model therefore used ImageNet initialization followed by fine-tuning on the classroom images rather than using VGG16 only as a fixed feature extractor. This is a form of transfer learning in which knowledge from a large source dataset is used as the starting point for a smaller target task [6], [7]. The resulting model had 14,812,995 trainable parameters.

The network was compiled using categorical cross-entropy as the loss function and Adam as the optimizer. The model was trained for five epochs. Validation accuracy increased from 31.77% after the first epoch to 96.35% after the fifth epoch. The purpose of the present study is not to optimize VGG16 or compare it with newer architectures. The model is treated as the implemented classifier whose feasibility is macro and weighted averages. The confusion matrix shows which classroom states were confused with each other. Looking at these class-level values is useful because overall accuracy does not show where the model performs well or where it makes errors.

Finally, the trained classifier was connected to a simple Flask-based web application to demonstrate its use. A classroom image can be

Table 3. Main configuration of the VGG16 classifier

Component	Configuration
Input size	$256 \times 256 \times 3$
Backbone	VGG16
Initialization	ImageNet pretrained weights
Original VGG	Removed (include_top
Task-specific	Flatten layer followed

Results

The classifier was tested on the held-out part of the 80:20 data split. This test set contained 240 images from the three classroom classes. The reported measures are accuracy, precision, recall, F1-score, and the confusion matrix. The findings are presented descriptively because the evaluation used one holdout test and did not include repeated runs or an external validation dataset.

evaluated for the three-state classroom problem. Table 3 summarizes the main configuration and training parameters of the VGG16 transfer learning model used for classroom state classification.

After model training, performance was evaluated on the held-out test images using standard classification measures. The main measures were accuracy, precision, recall, and F1-score. Accuracy describes the proportion of all test images that were assigned to the correct classroom state. Precision describes how many images predicted as a particular class actually belonged to that class. Recall describes how many images belonging to a class were correctly identified. The F1-score provides a combined measure of precision and recall.

For each class, the classification report shows precision, recall, and F1-score. It also gives given to the application, and the predicted classroom state is then shown to the user. The developed interface displays the predicted classroom state as text. This interface is included only as a simple demonstration of model use. The study does not evaluate usability, processing speed, or real-time operation of the web application.

Component	Configuration
Trainable	14,812,995
Loss function	Categorical cross-entropy
Optimizer	Adam
Training epochs	5
Test division	20% of the full dataset
Validation	20% of the training porti

Overall Classification Performance

The VGG16-based classifier correctly predicted 233 of the 240 test images. This gives an overall test accuracy of 97.08%. The macro-average precision, recall, and F1-score were all 0.97. The weighted averages were also 0.97. These results show that the classifier performed consistently across the three classes in the test split.

Table 4 presents the class-level results. The strongest performance was observed for the

empty-classroom state, with a precision of 1.00, recall of 0.99, and F1-score of 0.99. The two occupied classroom states both obtained an F1-score of 0.96. This indicates that most

Table 4. Classification performance on the 80:20 held-out test set

Classroom state	Precision	Recall	F1-score	Support
Lecture in progress	0.93	0.99	0.96	76
Students present / teacher absent	0.99	0.94	0.96	81
Empty classroom	1.00	0.99	0.99	83
Macro average	0.97	0.97	0.97	240
Weighted average	0.97	0.97	0.97	240

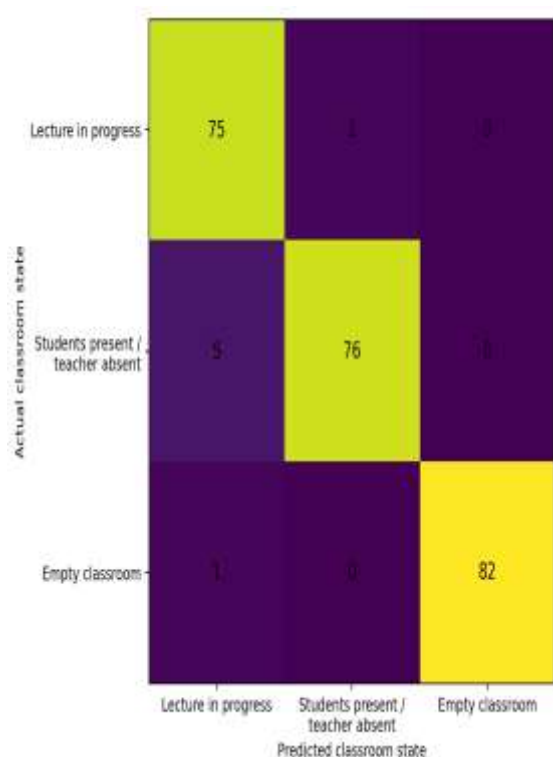


Figure 3. Confusion matrix for the 80:20 held-out test set

Confusion Matrix and Error Pattern

The confusion matrix gives a more direct view of the prediction errors. Figure 3 shows that 75 of the 76 lecture-in-progress images were correctly classified. One image from this class was classified as students present with the teacher absent. For the students-present/teacher-

classification errors occurred between the two visually similar occupied states rather than between occupied and empty classrooms.

absent state, 76 of 81 images were correctly classified, while five images were classified as lecture in progress. For the empty-classroom state, 82 of 83 images were correctly classified, while one image was classified as lecture in progress.

In total, seven test images were misclassified. Five of these errors occurred when the students-present/teacher-absent state was predicted as lecture in progress. This was the largest single error pattern in the test set. The result is reasonable for a scene-level classifier because both states contain students and share many similar classroom features. The presence and role of the teacher are the main visual difference between these two conditions. A fuller interpretation of this pattern is provided in the Discussion section.

4.3 Training and Validation Behaviour

Table 5 shows the training and validation performance across the five epochs. Training accuracy was 35.16% in the first epoch and increased to 96.09% by the fifth epoch. Validation accuracy increased from 31.77% to 96.35%, while validation loss decreased from 1.1000 to 0.1464. The largest improvement occurred between the second and fourth epochs. These values describe the training run only and are not treated as another test result.

Table 5. Recorded training and validation performance across five epochs

Epoch	Training accuracy	Validation accuracy	Training loss	Validation loss
1	0.3516	0.3177	1.4134	1.1000
2	0.3490	0.4010	1.0969	1.0933
3	0.5039	0.9062	0.9351	0.3228
4	0.8737	0.9271	0.3769	0.1719
5	0.9609	0.9635	0.1262	0.1464

Results in Relation to the Research Questions

RQ1 asked how well the VGG16-based model could separate the three classroom states in the dataset. The test accuracy was 97.08%, and the macro F1-score was 0.97. This provides encouraging proof-of-concept evidence for the dataset. RQ2 asked whether performance was different across the three states. The results show that the empty classroom was recognized most accurately. Most of the confusion was between lecture in progress and students present without the teacher. These findings apply to the dataset and the recorded train-test split. They should therefore be considered evidence of feasibility, not an estimate of performance for other classroom environments.

Discussion

The study tested whether VGG16 could recognize three broad classroom states from still images. On the test set, the model reached 97.08% accuracy and a macro F1-score of 0.97. These results provide positive evidence that the three classroom states can be distinguished from the images. However, the findings should still be treated as proof-of-concept results because the evaluation used one train-test split and no independent external dataset.

The class-level results also provide useful information about the performance of the model. The empty-classroom state achieved the highest F1-score of 0.99, while both occupied states achieved an F1-score of 0.96. Most errors occurred between lecture in progress and students present without the teacher. These two states can contain many similar visual features, such as students, desks, boards, and normal classroom arrangements. The main difference is whether the teacher is present and whether teaching is taking place. In comparison, an empty classroom has a clearer visual difference from the two occupied states.

The confusion between the two occupied states also shows one limitation of using a single image. A still image represents only one moment. In some images, the teacher may not be clearly visible or may not be standing at the front of the room. Similarly, students waiting

for a teacher may appear similar to students attending a lecture. A detailed manual analysis of the misclassified images was not performed. Therefore, these explanations should be treated as possible reasons for the observed errors rather than confirmed causes.

The findings support the main position of this paper that not every classroom vision problem requires detailed recognition of individual activities. Previous studies have mainly examined actions such as writing, walking, reading, hand raising, and explaining [2]-[4], [11]. Such activities often require video because the action develops over time. The present study addresses a simpler operational problem. It predicts the overall state of the classroom rather than the activity of an individual person. For this type of problem, a still image may provide sufficient information for an initial classification.

The study is also related to the work of Ikram et al. [5], who used VGG16 transfer learning for image-based student engagement recognition. Their study showed that a pretrained VGG16 model can be adapted to classroom images. However, engagement recognition and classroom-state recognition have different purposes. Engagement recognition focuses on the learning state of students, while the present study predicts the overall condition of the classroom. Therefore, the reported accuracy values should not be compared directly. Similarly, the higher accuracy obtained in the present study does not mean that the model performs better than video-based classroom activity-recognition approaches because the classification tasks are different.

From a practical point of view, the model may be useful for simple classroom-status applications. A future system could indicate whether a classroom appears empty, whether students are present while waiting for a teacher, or whether a lecture appears to be in progress. The existing Flask interface provides a basic example of how the classifier can be connected to an application. However, the present study did not evaluate continuous camera input, real-time processing, system speed, user acceptance,

or integration with institutional systems. Therefore, the prototype should be considered only as a demonstration of the classification approach and not as a fully evaluated smart-classroom system.

Several limitations need to be considered. The images were collected from more than one classroom, but the study does not include a controlled test on a completely unseen classroom. The number of sessions from which the images were extracted is also limited. The study also used one 80:20 holdout split, did not compare VGG16 with other models, and included only a limited number of classroom states. The reported accuracy should therefore be treated as the result for this dataset and not as an expected level of performance for every classroom.

Overall, the results show that the three broad classroom states can be handled as a scene-classification task. The empty-classroom state was recognized most accurately, while the two occupied states were more difficult to separate. This shows that an established pretrained vision model can be useful for this limited problem. Future evaluation should use clearly separated recording sessions and independent test data from additional classrooms. If more complex classroom states are added, methods such as object detection, pose information, or short video clips may also be examined.

Conclusion

This study investigated a simple camera-based method for identifying the overall state of university classrooms. Only three states were considered: lecture in progress, students present while the teacher is absent, and empty

classroom. The study did not aim to introduce a new deep learning architecture. Its purpose was to check whether an established pretrained model can perform this practical classification task using still classroom images.

The dataset contained about 1,200 locally collected images, with approximately 400 images in each state. A VGG16-based classifier was adapted for the three classes. Using the 80:20 train-test split, the model obtained 97.08% test accuracy. Macro precision, recall, and F1-score were all 0.97. These findings answer RQ1 by showing that the three defined classroom states were separated with high accuracy in the dataset.

The class-level results answer RQ2. The empty-classroom state achieved the strongest class-level performance, with an F1-score of 0.99. The other two states both had an F1-score of 0.96. Most errors were between lecture in progress and students present without the teacher. This is understandable because both classes contain students and use the same classroom background. The main difference is the presence of the teacher and whether active teaching is taking place. This makes the two occupied states more difficult to distinguish than an empty room.

The contribution of the study is practical and intentionally limited. It shows that a broad classroom condition can be handled as a scene-classification task rather than as detailed human activity recognition. The findings should be treated as proof-of-concept results because the evaluation was based on one recorded holdout split. Further work should test the approach using independent data from different classroom environments.

Classroom Environment,” *Sensors*, vol. 21, no. 17, Art. no. 5699, 2021, doi: 10.3390/s21175699.

References

- [1] X. Zhang, Y. Ding, X. Huang, W. Li, L. Long, and S. Ding, “Smart Classrooms: How Sensors and AI Are Shaping Educational Paradigms,” *Sensors*, vol. 24, no. 17, Art. no. 5487, 2024, doi: 10.3390/s24175487.
- [2] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, “EduNet: A New Video Dataset for Understanding Human Activity in the
- [3] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, “STAR-3D: A Holistic Approach for Human Activity Recognition in the Classroom Environment,” *Information*, vol. 15, no. 4, Art. no. 179, 2024, doi: 10.3390/info15040179.

- [4] Y. Rajamanickam, J. S. Fogarty, R. Rajalakshmi, and R. Sarkar, "Classroom Activity Recognition Using Hybrid 3D-CNNs and Visualization of Action Features with Grad-CAM," *Neurocomputing*, vol. 657, Art. no. 131694, 2025, doi: 10.1016/j.neucom.2025.131694.
- [5] S. Ikram, H. Ahmad, N. Mahmood, C. M. N. Faisal, Q. Abbas, I. Qureshi, and A. Hussain, "Recognition of Student Engagement State in a Classroom Environment Using Deep and Efficient Transfer Learning Algorithm," *Applied Sciences*, vol. 13, no. 15, Art. no. 8637, 2023, doi: 10.3390/app13158637.
- [6] J. Plested, M. Phiri, and T. Gedeon, "Deep Transfer Learning for Image Classification: A Survey," *Artificial Intelligence Review*, vol. 59, Art. no. 100, 2026, doi: 10.1007/s10462-026-11491-z.
- [7] M. Iman, H. R. Arabnia, and K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," *Technologies*, vol. 11, no. 2, Art. no. 40, 2023, doi: 10.3390/technologies11020040.
- [8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [9] Q. Liu, X. Jiang, and R. Jiang, "Classroom Behaviour Recognition Using Computer Vision: A Systematic Review," *Sensors*, vol. 25, no. 2, Art. no. 373, 2025, doi: 10.3390/s25020373.
- [10] J. K. Foster, M. Korban, P. Youngs, G. S. Watson, and S. T. Acton, "Automatic Classification of Activities in Classroom Videos," *Computers and Education: Artificial Intelligence*, vol. 6, Art. no. 100207, 2024, doi: 10.1016/j.caeai.2024.100207.
- [11] Y. Rajamanickam, A. A. Prince, and M. Murugappan, "Automated Recognition of Teacher and Student Activities in the Classroom: A Deep Learning Framework," *IEEE Access*, vol. 12, pp. 192159–192171, 2024, doi: 10.1109/ACCESS.2024.3518577.
- [12] T. Fütterer, P. Goldberg, B. Bühler, V. Sikimić, U. Trautwein, P. Gerjets, K. Stürmer, and E. Kasneci, "Artificial Intelligence in Classroom Management: A Systematic Review on Educational Purposes, Technical Implementations, and Ethical Considerations," *Computers and Education: Artificial Intelligence*, vol. 9, Art. no. 100483, 2025, doi: 10.1016/j.caeai.2025.100483.
- [13] C. Llurba, G. Fretes, A. Martínez-Ballesté, and R. Palau, "Artificial Intelligence in Smart Classrooms: A Systematic Literature Review of Applications, Dimensions, and Teacher Roles," *Encyclopedia*, vol. 6, no. 6, Art. no. 138, 2026, doi: 10.3390/encyclopedia6060138.
- [14] Q. Jia and J. He, "Student Behaviour Recognition in Classroom Based on Deep Learning," *Applied Sciences*, vol. 14, no. 17, Art. no. 7981, 2024, doi: 10.3390/app14177981.
- [15] X. Sheng, S. Li, and S. Chan, "Real-Time Classroom Student Behaviour Detection Based on Improved YOLOv8s," *Scientific Reports*, vol. 15, Art. no. 14470, 2025, doi: 10.1038/s41598-025-99243-x.