*electronics*

MDPI

*Article*

# Hybrid Feature Extraction Model to Categorize Student Attention Pattern and Its Relationship with Learning

**Sujan Poudyal *, Mahnas J. Mohammadi-Aragh and John E. Ball**

Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA; jean@ece.msstate.edu (M.J.M.-A.); jeball@ece.msstate.edu (J.E.B.)
* Correspondence: sp2115@msstate.edu

**Abstract:** The increase of instructional technology, e-learning resources, and online courses has created opportunities for data mining and learning analytics in the pedagogical domain. A large amount of data is obtained from this domain that can be analyzed and interpreted so that educators can understand students' attention. In a classroom where students have their own computers in front of them, it is important for instructors to understand whether students are paying attention. We collected on- and off-task data to analyze the attention behaviors of students. Educational data mining extracts hidden information from educational records, and we are using it to classify student attention patterns. A hybrid method is used to combine various techniques like classifications, regressions, or feature extraction. In our work, we combined two feature extraction techniques: principal component analysis and linear discriminant analysis. Extracted features are used by a linear and kernel support vector machine (SVM) to classify attention patterns. Classification results are compared with linear and kernel SVM. Our hybrid method achieved the best results in terms of accuracy, precision, recall, F1, and kappa. Also, we correlated attention with learning. Here, learning corresponds to tests and a final course grade. For determining the correlation between grades and attention, Pearson's correlation coefficient and *p*-value were used.

**Keywords:** hybrid feature extraction; educational data mining; correlation; learning analytics; attention pattern



**Citation:** Poudyal, S.; Mohammadi-Aragh, M.J.; Ball, J.E. Hybrid Feature Extraction Model to Categorize Student Attention Pattern and Its Relationship with Learning. *Electronics* **2022**, *11*, 1476. https://doi.org/10.3390/electronics11091476

Academic Editors: Georgios Kostopoulos and Sotiris Kotsiantis

Received: 24 March 2022
Accepted: 29 April 2022
Published: 5 May 2022



## 1. Introduction

Today's classrooms are rich with opportunities to examine learning behaviors and generate data-driven improvements to educational systems by using data mining. This opportunity exists, in part, due to the incorporation of laptops, tablets, mobile devices, and virtual reality (VR) systems into classrooms, which allow students to experience the outside world without leaving the classroom. This is one of the fastest growing trends in education [1–4], and these devices provide numerous opportunities for data collection. Thanks to technological growth in education, a large amount of educational data is produced [5]. The area of data mining that explores data generated within the educational setting is called educational data mining (EDM). Datasets include information about student academic records and the patterns of student interaction with classroom technology [6]. EDM has been used on student record data to predict academic performance [7–11] and dropout prediction [12–14]. EDM has also been used to detect undesirable students' behavior in the classroom. By using computer vision techniques and machine-learning algorithms [15], Thomas et al. analyzed the engagement or attention level of only ten students in the classroom from their facial expressions, eye gaze, and head pose. Zaletelj et al. used only 22 undergraduate engineering students in a small classroom to determine the attention of students by using machine-learning algorithms such as decision trees and K-Nearest Neighbor [16]. A few other studies were performed to detect the students' behavior in a small classroom, but they similarly used a small participant pool (less than 30 students) [17–20].

In this study, we are examining the use of data mining to detect students' behavior in a large classroom.

One example in which data-mining techniques may provide especially useful insights is the examination of learning behaviors within large lecture-based courses. Typically, these courses are held in large auditoriums where classroom management is challenging. In smaller classes, it is easier for the instructor to observe individual students, and any students who are performing off-task can be warned by the instructor [21]. However, in large classes, it becomes difficult for instructors to keep track of individual students [20]. Technology exacerbates classroom management difficulties in large lectures, as technology often pulls students off-task. For example, although students bring laptops to the classroom for academic tasks such as taking notes [22], they also use laptops for non-academic tasks like browsing the internet [23], playing games [24], and checking email [25]. They switch back and forth between academic and nonacademic tasks [26], and such multitasking hinders learning in the classroom environment. A 2020 study showed that students who use laptops in the classroom engage in more multitasking, which causes negative effects on their ability to remember the course contents [27]. Furthermore, multitasking on a laptop can distract nearby students who are in direct view of the laptop [26]. It is critical to understand the student's behavior in large lectures because multitasking negatively affects GPA, efficiency, self-regulation, recall, test performance, and reading comprehension [28]. Data-mining techniques may provide useful insights for examining attention behaviors in large lecture-based courses where the students have laptops open in front of them.

Furthermore, it is important to improve the performance of the data-mining model so decisions are made by using the most accurate results possible. Khokhoni et al. used the data-mining model to predict the academic status of the learners in advance, which helped to identify the weak learners and take necessary actions [29]. The J48 algorithm obtained the highest accuracy of 99.13%, which means the model by Khokhoni et al. identifies 99.13% of weak learners [29]. Different techniques in EDM, like classifications, regressions, clustering, and feature selections, can be combined to improve the performance of the model. Such a type of combined model is called a hybrid model.

In this study, we examined methods to increase the accuracy of EDM to analyze attention in large lectures. The dataset used in this study was captured during another study that examined the relationship between attention and the active window on a student's computer [30,31]. By monitoring the active window on the student's laptop in large lecture classes (>200 seats) where students were using specific course software as part of the lecture, researchers electronically collected the students' on-task and off-task activities. The original research team completed observations they compared to the electronic active window data to validate active window as a proxy for attention. Their analysis produced error rates of 4.28% and 6.89%, which was dependent on the instructor's policies for using course software. During the validation analysis, the authors noted that discernably different attention patterns existed. Additional details about data collection, validation, and analysis for the original study are available in [31]. In our prior work, we investigated the original study's conclusion that students had discernably different attention patterns by applying a Haar wavelet classification with a support vector machine (SVM) [32]. In this current work, we explored methods to improve the model performance. To obtain better classification accuracy, we used a hybrid model that combined two different feature extraction techniques. We used principal component analysis (PCA) and linear discriminant analysis (LDA) for the feature extraction. We then used an SVM to classify the students' attention behaviors into different categories based on their attention patterns in the classroom. Specifically, we answer the following questions.

Q1. Does feature extraction technique improve the classification accuracy for active window/attention behavior of students?

Q2. Does the hybrid model improve the classification accuracy for active window/attention behavior of students?

Q3. How strongly are the classified students' attention patterns (based on active window data) related to their progress exam scores?

The objectives or motivation of this paper are first to characterize students' attention behavior in binary form (0 and 1). Second, to classify the student's attention behavior and improve the performance of the classification model by using feature extraction techniques. Third, to build a hybrid model that further improves the classification accuracy for classifying the binary attention behavior of the students. Fourth, to find the correlation between the attention patterns and the learning based on progress exam scores.

The main aim of this paper is to characterize the student's attention pattern in binary form and to use the combination of two feature extraction techniques, PCA and LDA, in the EDM field to improve the performance of the classification model in terms of accuracy. To fulfill this, a novel hybrid feature extraction model in the EDM field is proposed in this paper.

The main contributions to this paper are

- to characterize the students' attention behavior in the classroom in binary form (0 and 1) and
- to combine two different feature extraction techniques to produce a single hybrid feature extraction technique model in the EDM field.

The remaining section of our paper follows this format: Section 2 is a literature review where we have discussed the ongoing work on attention and educational data mining, its limitations, and introduced our work. Section 3 is the method section, where we have explained our dataset, performance metrics, and use of machine-learning techniques in our work. Section 4 includes the result of our study. The fifth and sixth sections are the discussion and conclusion, respectively. Section 5 answers our research questions, and Section 6 concludes our study with future work suggestions.

## 2. Literature Review

Machine-learning techniques are frequently used for classification and prediction in a wide range of fields. The use of machine learning to assess students' behavior in the classroom is a contentious issue in educational data mining. Different assessments have been carried out to assess the students' behavior in the classroom. In psychology, cognitive processes are identified by psychological signals such as eye-tracking signals, heart rate signals, or signals from electro-dermal activities [33]. These are used in educational data-mining research to measure the attention of the students in a classroom. Deng et al. developed a machine-learning model that is used for eye state classification for the students' visual attention assessment [34]. The model, which was based on the Gabor feature, obtained an accuracy of 93.1%. Adem et al. used 21 freshmen in the classroom at Usak University to detect the students' attention levels and used NeuroSky's Mindset EEG devices to detect the attention levels of the students in the classroom [35]. By using the Pearson's correlation coefficient, the results showed a positive, moderate correlation between the students' attention during class and the rate of participation in the classroom. However, using these types of high-tech devices is not feasible in a standard classroom setting because teachers should have the knowledge to use such devices, and students may not be comfortable using such devices. So, we have used software in our work that uses the active window to record the attention pattern of the students in binary form, i.e., zero for off-task and one for on-task.

Research has been done to correlate students' attention with their learning. Reference [36] used an internet proxy server to monitor student internet use during the lecture class and estimates that time off-task negatively correlates to learning. Spyware was used by [37] to monitor all computer use and, in this case, the time off-task negatively correlated with the final course grades. For the correlation between learning and attention, we see that most of the work used a single lecture performance metric to measure the learning, like a post-lecture quiz, rather than using the entire course performance metric to measure the learning, like a final exam grade. In our work, we study the effects of the students'

attention patterns on their subsequent test grades and final grades by using the Pearson's correlation coefficient ($r$) and the $p$-value.

Different feature selection and extraction techniques have been used to select the important features from the dataset. Punlumjeak et al. applied the feature selection techniques in the preprocessing stage to find out the relevant features by using four different techniques: genetic algorithms, support vector machine, information gain, and minimum redundancy and maximum relevance [38]. The comparison of these techniques was performed, and it was discovered that the minimum redundancy and maximum relevance feature selection techniques achieved the best result, with an accuracy of 91.2% [38].

Different researchers have used the hybrid model in EDM to improve the classification models. Francis et al. combined clustering and classification algorithms to evaluate students' performance in academia and found that such a hybrid model yields the best results in terms of accuracy [39]. Another study by Rawat et al. also shows that a hybrid model of classification has better performance for predicting student-related data [40]. A 2019 study used a combination of wrapper feature selection techniques and different machine-learning algorithms such as K-Nearest Neighbor, Convolutional Neural Network, Naïve Bayes, and Decision Tree. The result showed that the hybrid method improved the performance of the classifiers by 2–3% [41]. Amrieh et al. used Artificial Neural Networks, Naïve Bayes, and Decision Tree approaches to classify the students' academic performance by using behavioral features and tried to improve the performance by using the ensemble method, which achieved up to a 25.8% improvement [42]. Xiao et al. used the hybrid feature selection method RnkHEU that integrates ranking-based forward and heuristic search for predicting the academic performance of students [43]. Different classifiers such as NB, C4.5, MLP, and KNN were used as classifiers and the RnkHEU method improved the classification accuracy by 10% with the highest accuracy being 71.19%.

In EDM, most research is done to predict students' academic performance [7–9] and dropout prediction [10–12]. Very few works have been done to predict the attention level of students in the classroom [44–47]. Thomas et al. used student's facial expression, head pose, and eye gaze to distinguish the attention level of students [44]. Reference [44] used models such as SVM and LR where SVM with radial basis function performed the best with an accuracy of 90%. Zhang et al. proposed an attention inference engine by using a rule-based approach or data-driven approach and used machine-learning algorithms such as J48 DT, RF, and SVM for the classification purpose [45]. J48 DT obtained the highest accuracy of around 82%. Zaletji et al. used 2D and 3D data obtained by the Kinect One sensor that includes facial and body properties of students [16]. Seven different traditional machine-learning algorithms were used to predict the time-varying attention levels of students, and obtained the moderate accuracy of 75.3%. Poudyal et al. used Haar wavelets, PCA, and LDA separately as the feature selection techniques with SVM, Decision Tree, and KNN classifiers to classify the students' attention patterns [46].

Different learning methods have been proposed for the learning of nonnegative data [47,48]. These methods are suitable for large datasets. In our study, we used the hybrid combination of two feature extraction techniques. Because our dataset is small, we used simple PCA and LDA feature extraction techniques. A combination of PCA and LDA has been used before outside of the EDM field. Yang et al. used two feature selection techniques, kernel PCA and LDA, in a combination, and used the handwritten numerical database called CENPARMI to verify the effectiveness of such a combination [49]. Zuo et al. used bidirectional PCA plus LDA in which LDA is performed in the bidirectional PCA subspace [50]. Reference [50] used facial recognition technology and the ORL database to verify such combinations. In their research, Deng et al. used PCA plus LDA techniques to select the Gabor feature for facial expression recognition [51]. From different literature, we came to know that combinations of LDA and PCA techniques are lacking in the EDM field.

We used classifiers to classify the extracted features from our hybrid model. Different classifiers have been used in machine learning that yields high classification accuracies. Peng et al. introduced a discriminative ridge regression approach to classification which is

applicable for high-dimensional data [52]. But our dataset is small which will be explained in the Method section. Consequently, we need to choose the classifier that will perform best for the small dataset. SVM has shown that it works with high classification accuracies for a small dataset [53,54]. As a result, in our study we are using SVM as a classifier.

In our study, we have used two feature extraction techniques in a hybrid form to improve the classification performance of SVM, which is measured in terms of accuracy, precision, recall, and Kappa value. We used only the combination of two filters. We captured the students' data electronically. The electronic monitoring was supplemented by in-person observations of student behavior that quantified the method's mean percent error at 4.28% and estimated a standard error of 0.82.

## 3. Method

In this study, we used the students' attention dataset to see the effect of the hybrid feature extraction methods on the classification model. Figure 1 shows the proposed hybrid feature extraction model architecture.

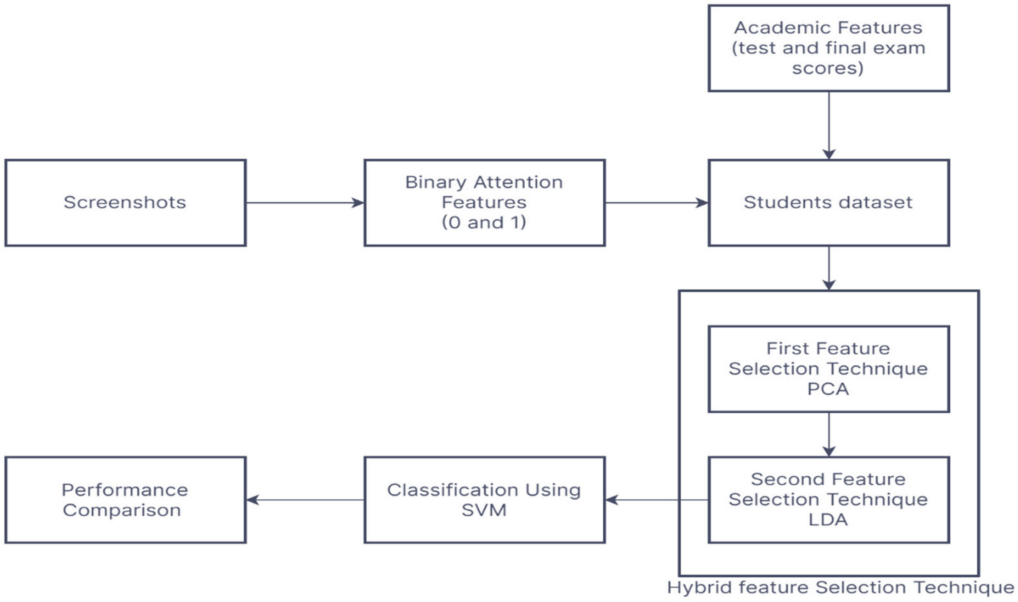


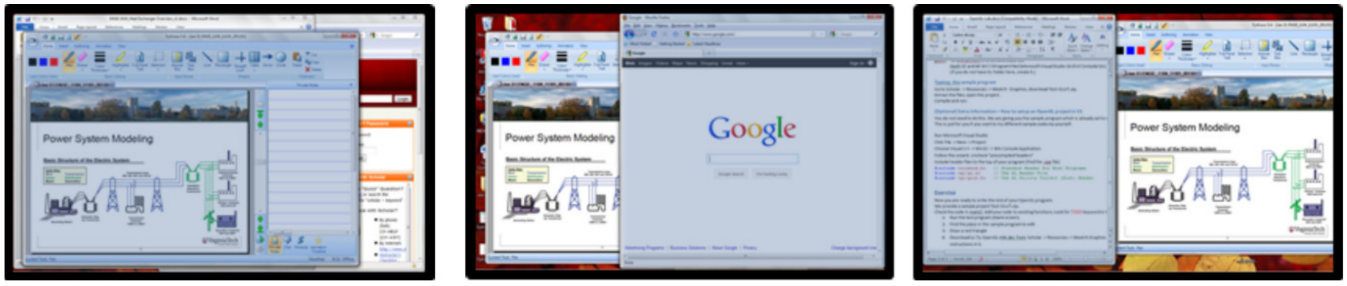

**Figure 1.** Proposed hybrid feature extraction model.

Figure 1 shows our proposed model. It shows that the dataset used in our study has two types of features, which are reduced by using our hybrid feature extraction technique in EDM. The hybrid technique is the PCA feature extraction technique followed by the LDA technique. The reduced features are then classified by using the SVM classifiers. The last module of our model is the performance comparison module wherein we compared the performance of our hybrid feature extraction technique with single feature extraction technique and the classifiers without using any feature extraction technique. The further detailed explanation of our hybrid model is given in below Section 3.1.

### *3.1. Experimental Setup*

The data was captured electronically from a large lecture classroom in a first-year engineering lecture class. The research was done with the first-semester engineering students. There were a total of 256 enrolled students, among whom 203 students consented to participate. The course was divided into a lecture session and a laboratory lesson. The lecture was fifty minutes long, and they met once a week in the early morning at eight o′clock. The laboratory session was 110-min long, and the students met once a week in a group of thirty. Our dataset includes only the data from a lecture session. The total lecture class was eight weeks throughout the semester. The demographics of students were typical for first-year engineering courses in the United States (i.e., predominately

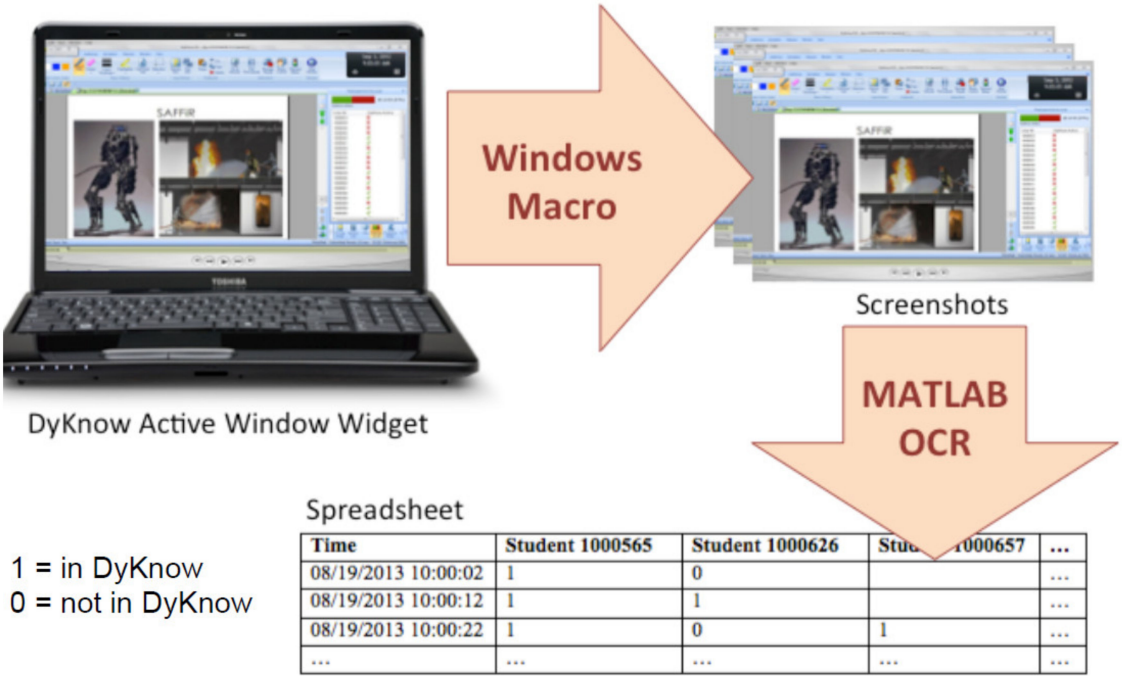
male, eighteen years of age). Data was collected with the Institutional Review Board (IRB) approval. Course software known as DyKnow was provided to the students, and they were informed to bring their laptops to the classroom and log into the software. There was a network connection between the students' and the instructor's computers to share the lecture content with the students. The network link was used to determine if the students were active on the required software on their laptops (on-task) or were using their laptops for non-academic activities (off-task). The students were recorded as on-task or off-task based on their active windows, as shown in Figure 2.




(**a**) (**b**) (**c**)

**Figure 2.** Examples of an active window. (**a**) Focus on DyKnow (on-task). (**b**) Focus on Firefox (off-task).(**c**) Focus on Word (off-task).

Figure 2 shows an example when the active window is considered on-task and off-task. The focus is on the course software in Figure 2a, so it is considered on-task, but the focus is away from the course content in Figure 2b,c, so they are considered off-task. The data was collected in the form of screenshots at an interval of every 20 s at each lecture class throughout the semester. The screenshots consist of each student's information whether they are performing on-task or off-task. The obtained screenshots were processed by using a custom MATLAB script to obtain the students' attention patterns in the form of one (on-task) and zero (off-task) as shown in Figure 3.




| Time | Student 1000565 | Student 1000626 | Stu… 1000657 | … |
|---|---|---|---|---|
| 08/19/2013 10:00:02 | 1 | 0 | | … |
| 08/19/2013 10:00:12 | 1 | 1 | | … |
| 08/19/2013 10:00:22 | 1 | 0 | 1 | … |
| … | … | … | … | … |

**Figure 3.** Processing of the screenshots.

The detailed description of the data is explained in [31] and provides a discussion of method validation, limitations, and error rates. We note that the methods involved capturing student data via a backend monitoring system to reduce student awareness of monitoring and reduce false behavior changes during data collection periods (i.e., students using a smartphone and leaving their computer set to "on-task" software).

After the students' attention pattern, other attention features were also calculated. In our study, we had two types of features: attention pattern features and academic features that included test and final exam scores. During the process of electronically capturing data, a MATLAB script was used to encode every student's data in the form of zeros (off-task) and ones (on-task). On-task means that the students are paying attention in the classroom, whereas off-task means that the students are not paying attention to the lecture content in the classroom. Once these patterns were obtained, we then calculated other characterization features by using these patterns, which are explained in Table 1. Academic features represent the scores that the students obtained on the tests and the final exam throughout the semester. The detailed description of the feature is shown in Table 1.

**Table 1.** Features description of the student's dataset.

| Features | Description | Value Range |
|---|---|---|
| On-task/Off-task | Students' attention pattern | 1 and 0 |
| Count of on-task entries | Total instances of one in a particular student's dataset | 0–154 |
| Count of off-task entries | Total instances of zeros on a particular student's dataset | 0–150 |
| On-task attention percentage | The percentage of instances that are on-task for a given total sum of on-task and off-task entries | 0–1 |
| Class period attention | Average class attention time series indicating the percentage of instances that are on-task for a given time record | 0–1 |
| Total number of switching between the on-task and off-task | Total count of number of times the student's attention pattern changes from zero to one or vice-versa | 0–62 |
| Average of the consecutive on-task entries | Average of the consecutive ones entries to the total entries on the particular student's data | 0–154 |
| Minimum consecutive on-task entries | Total count of minimum number of consecutive ones entries on the particular student's data | 1–154 |
| Maximum consecutive on-task entries | Total count of maximum number of consecutive ones entries on the particular student's data | 1–154 |
| Average of the consecutive off-task entries | Average of the consecutive zeros entries to the total entries on the particular student's data | 0–148 |
| Minimum consecutive off-task entries | Total count of minimum number of consecutive zeros entries on the student's data | 0–148 |
| Maximum consecutive off-task entries | Total count of maximum number of consecutive zeros entries on the student's data | 0–148 |
| Time duration at which the first off-task occurred | Time at which the first zero entry is seen on the student's data | 20–3060 |
| Total time duration for the first on-task period | Total time duration for which the first ones entries are seen on the particular student's data | 0–3080 |
| Test and Final Exams | Total two tests and one final exam | 0–100 |
| Course grade | Final grade of the students on that specific course | A, B, C, D, F |

In Table 1, in the value range, the minimum value represents the smallest value, and the maximum value represents the largest value among all the students' data for that particular feature. For example, the feature "maximum consecutive on-task entries" has a value range of 1–154. This means the total count of the maximum number of consecutive

one-entry entries on the particular student's data has the smallest value of one count and the largest value of 154 counts among all the students.

A MATLAB script was used to plot the attention patterns of the students, and the expert classified each attention pattern into one of the four classes, which are:

- Class 1: Attentive Students
- Class 2: Students with more attentive periods and fewer inattentive periods
- Class 3: Students with less attentive periods and more inattentive periods
- Class 4: Inattentive students

Students who did not fit into any of these categories were removed from the study. This includes the students who had almost all lower peaks on the waveform because of missing the lecture classes. Also, students whose waveforms do not fall into any of the four categories, like having equal attentive and inattentive periods, were removed from the studies. The attention pattern was plotted by using the MATLAB script to obtain the samples that represent the four different classes. Figure 4 shows the samples that represent all four classes. In Figure 4, high values (1) represent the attentive period, and low values (0) represent the inattentive period.

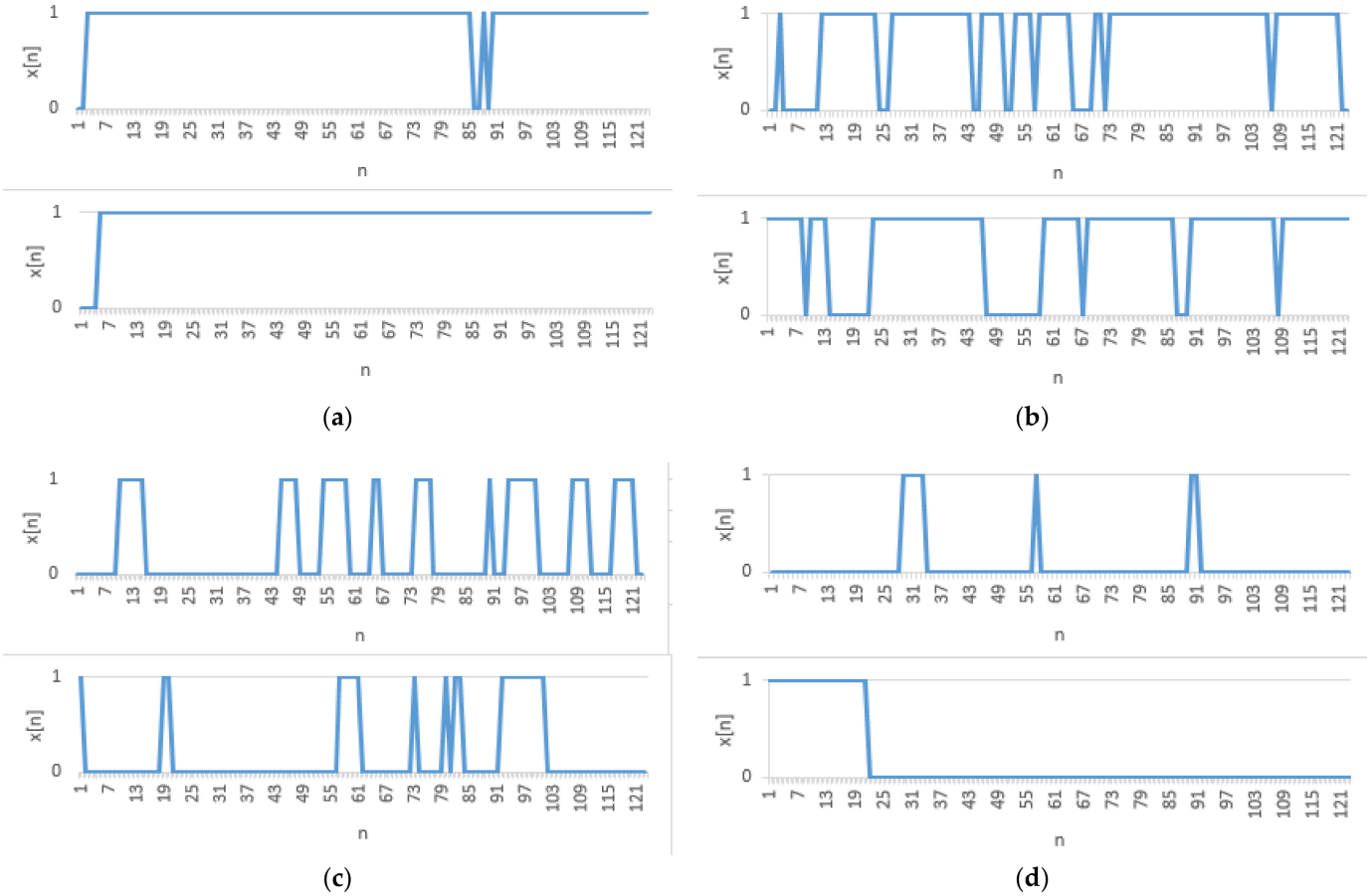



**Figure 4.** A sample to represent students in: (**a**) Class 1; (**b**) Class 2; (**c**) Class 3; (**d**) Class 4.

Class 1 represents the students where they are most attentive, as shown by the high peak of the waveform in Figure 4a. Figure 4b is the sample of class 2 where students are mostly attentive, which is represented by the high peak of the waveform, while the inattentive periods are few, which is represented by the low peak. Figure 4c is the sample of class 3 where students are mostly inattentive, which is represented by the low peak, and have few attentive periods, which is represented by the high peak. Class 4 represents students where the students are mostly inattentive, as shown by the low peak of the waveform in Figure 4d.

These attention patterns are then classified by the expert so that each student falls into one of the four classes. If the students' pattern follows the sample like in Figure 4a, then the students fall into class 1. If the pattern follows the sample like in Figure 4b, then the students fall into class 2, and so on. Because the lecture class was a total of eight weeks in the semester, we obtained one dataset for each week, making a total of eight datasets. Data in Table 2 shows the total number of students in each class as classified by the experts for eight different weeks.

**Table 2.** Expert classification of students for different weeks.

| Classification Group | Students per Week | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **First** | **Second** | **Third** | **Fourth** | **Fifth** | **Sixth** | **Seventh** | **Eighth** |
| Class 1 | 114 | 140 | 92 | 71 | 52 | 30 | 31 | 44 |
| Class 2 | 47 | 31 | 35 | 34 | 35 | 72 | 30 | 12 |
| Class 3 | 0 | 5 | 7 | 8 | 15 | 3 | 11 | 5 |
| Class 4 | 0 | 1 | 3 | 3 | 2 | 3 | 8 | 10 |
| Unclassified | 2 | 22 | 15 | 20 | 21 | 10 | 37 | 15 |
| Total | 163 | 199 | 152 | 136 | 125 | 118 | 117 | 86 |

Table 2 only represents those students who attended the entire lecture and excludes absentees, tardiness, or students who forgot to bring laptops to the classroom. All the students in Table 2 are subsets of the study participants. Some of the students' attention patterns did not fall into any of the classification groups, so they were kept as unclassified. These include students who had more lower peaks on the waveform because of missing the lecture classes. Also, students whose waveforms do not fall into any of the four categories, like having equally attentive and inattentive periods, were removed from the studies, which is one of the shortcomings of this study.

We used feature extraction techniques before performing the classification. We used a combination of two different feature extraction techniques, namely PCA and LDA. We applied the PCA to our dataset first to reduce the dimensions of our dataset. The newly obtained features were then applied to the LDA model, which produced the LDA reduced features. The new dataset with the reduced features was then used by our classifiers to classify each student into one of four different classes. We used SVM as our classifier.

SVM is a supervised machine learning algorithm. The SVM solves the unconstrained optimization problem as shown in Equation (1),

$$\frac{1}{2}w^T w + C\sum_{j=1}^{L}\xi\left(w_j; x_j; y_j\right), \tag{1}$$

where, $w = [w_1, w_2, \ldots, w_N]^T$ is the $[N \times 1]$ optimal weight vector, $x_j$ is a $[N \times 1]$ feature vector, $y_j \in \{-1, 1\}$ is the class associated with $x_j$ (1 = target, $-1$ = no target), $C$ is a penalty parameter for misclassifications, $\xi$ is known as convex optimization strategies that are used to solve for the optimal weight vector, and $L$ is the number of training samples. For the optimization, the loss function $\xi$, must be minimized. The value of loss function is based on the distance to the classification boundary. If the loss function is near zero, the training sample is correctly classified and if it is increased, the training sample is incorrectly classified.

During the training phase, we provide the target and non-target training data to the SVM, and the weight is optimized to best discriminate target cases from non-target cases. During the testing phase, we present the new feature vectors and if the SVM weighted feature is greater than zero, they are considered the target. In our work, we have used linear SVM and kernel SVM for the classification of students' attention patterns. We chose SVM because SVM is relatively more efficient. The SVM algorithm is not suitable for large datasets, and this benefited us as our dataset is relatively small.

After the classification was done, we observed whether the attention pattern had a statistically significant relationship with the students' grade. The grades of the students were measured in five categories, as explained in Table 3.

**Table 3.** Explanation of students' grade at the end of the course.

| Grade | Explanation |
|---|---|
| A | Students whose total final score lies between 87–100 |
| B | Students whose total final score lies between 80–86 |
| C | Students whose total final score lies between 70–79 |
| D | Students whose total final score lies between 60–69 |
| F | Students whose score lies below 60 |

Only the students who completed the whole course in the semester were taken for our study, and the students who dropped out in the middle were discarded. So, the total number of students in our study was $N$ = 171 students. The plot of the grades and the total number of students obtaining each grade is shown in Figure 5. We note that this grade distribution is non-normal. The highest number of students obtained the grade B, with a total count of sixty-six students. Only two students obtained Grade F, which is the least count among all the grades. The skew in the grade distribution related to the type of course studied; the course was an introductory engineering course for first year engineering students. Some topics such as professionalism, graphing, sketching, ethics, software, and globalization were likely less challenging for students than concepts taught in other engineering courses. This non-normal grade distribution impacted our ability to relate attention patterns to course grades because grades were compressed into the end of the grade distribution.

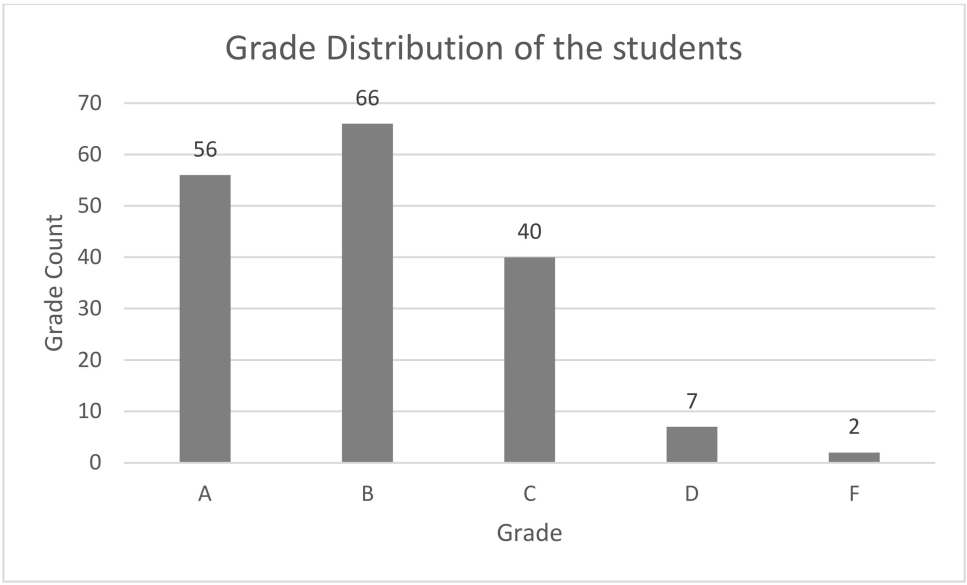



**Figure 5.** Grade distribution of students.

After the students were classified into different classes, the classes were then averaged to see the correlation between the attention pattern and the academic performance of the students. All the other features of each student, as shown in Table 1, were also averaged, and a new dataset was obtained to see the correlation between the students' attention in the classroom and their academic performance. For this purpose, we used Pearson's correlation. It gives the values between $-1$ and 1 that explain to what extent the two variables are linearly related. A $-1$ represents the negative correlation, whereas a +1 represents a positive

correlation between two variables, and 0 represents no relationship. Pearson's correlation coefficient ($r$) between two variables is calculated by using Equation (2) [55],

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}, \quad (2)$$

where, $N$ is the total number of pairs of scores, $\sum xy$ is the sum of the products of paired scores, $\sum x$ is the sum of $x$ scores, $\sum y$ is the sum of $y$ scores, $x^2$ is the sum of squared $x$ scores, and $y^2$ is the sum of squared $y$ scores. The result is explained in Section 4.

### *3.2. Performance Measures*

In our study, we used linear SVM and kernel SVM classifiers. We used the confusion metrics as a tool to explain the performance of the classification algorithm. It consists of rows and columns where the rows represent the true samples, and the columns represent the predicted samples. The values in the diagonal of the confusion matrix represent the truly classified samples. By using the confusion matrix, we calculated five performance metrics that were used as the performance metrics in our study: classification accuracy, precision, recall, F1 value, and Cohen's kappa.

Classification accuracy is calculated using Equation (3),

$$Classification\ Accuracy = \frac{\sum_{m=1}^{M} C(m,m)}{\sum_{m=1}^{M}\sum_{n=1}^{M} C(m,n)}, \quad (3)$$

where the numerator represents the total truly classified samples, and the denominator represents the total samples.

Precision tells us about the proportion of the positive identifications that were correct. It is calculated by using Equation (4),

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}}, \quad (4)$$

where the numerator denotes the truly classified samples and the denominator denotes the sum of all predicted samples.

Recall tells us about the proportion of the actual positives that were identified correctly. It is calculated using Equation (5),

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}}, \quad (5)$$

where the numerator denotes the truly classified samples and the denominator denotes the sum of all the actual samples.

The $F1$ value is the harmonic mean of the precision and the recall value, which is given by Equation (6),

$$F1 = \frac{2(Precision \times Recall)}{Precision + Recall}. \quad (6)$$

Cohen's kappa is calculated based on the confusion matrix and is used to access the performance of the classification algorithm. Cohen's kappa explains how much the classification agrees with the truth values (expert classification). The possibility of the classifier and a random guess agreeing is removed by Cohen's kappa. The number of predictions that cannot be explained by a random guess is measured by Cohen's kappa. Also, Cohen's kappa considers the correct classification by a random guess to correct the evaluation bias. The value of Cohen's kappa ranges from zero to one, where zero represents no agreement and one represents the perfect agreement. It is calculated by using Equation (7),

$$Kappa\ (K) = \frac{NC - g}{N^2 - g}, \tag{7}$$

where, $N$ is the total samples, $C$ is the total sum of truly classified samples, and $g$ is the sum of products of the total true samples and total predicted samples for each class.

## 4. Results

There were eight weeks in our study. We took the attention data of one week as the training set and the attention data of all remaining weeks as the testing set. The results are then averaged. We performed two experiments. In the first experiment, we used the data from the fourth week as a training set and the data from all other weeks as a testing set. As explained in the Method section, we used the combination of PCA and LDA to extract the features that were passed to the classifier models, linear SVM and kernel SVM. We extracted forty-four features by using PCA and sixteen features by using the LDA technique. These numbers were determined by running the experiment several times with different number of features, and the number that yielded the highest accuracy was considered. The performance of the classifiers was measured by using the confusion matrix. To compare the results of our hybrid model, we used a different combination of feature extraction techniques and classifiers. The confusion matrices when using the fifth week data as the testing set are shown in Figure 6.

Figure 6 represents the confusion matrix for the first experiment, taking the dataset from week four as a training set and the dataset from week five as a testing set. On the confusion matrix, the diagonal elements are the truly classified samples. Similarly, all other confusion matrices are also obtained with all the remaining week's data as a testing set. With the help of these confusion matrices, we now calculate different performance matrices by using Equations (3)–(7). Many results were obtained because our classification model was run with one training set and seven different testing sets. The results were then averaged and are shown in Table 4.

**Table 4.** Comparison of Different Techniques for the first experiment.

| | **Accuracy** | **Precision** | **Recall** | ***F*1** | **Kappa** |
|---|---|---|---|---|---|
| Linear SVM | 0.776 | 0.588 | 0.594 | 0.576 | 0.527 |
| Kernel SVM | 0.788 | 0.705 | 0.686 | 0.693 | 0.569 |
| PCA and Linear SVM | 0.794 | 0.639 | 0.592 | 0.681 | 0.539 |
| PCA and Kernel SVM | 0.806 | Nan | 0.558 | Nan | 0.53 |
| LDA and Linear SVM | 0.824 | 0.574 | 0.619 | 0.594 | 0.685 |
| LDA and Kernel SVM | 0.784 | 0.589 | 0.609 | 0.595 | 0.611 |
| Hybrid method and Linear SVM | **0.877** | **0.865** | **0.746** | **0.865** | **0.728** |
| Hybrid method and Kernel SVM | 0.841 | 0.852 | 0.571 | 0.76 | 0.649 |

Table 4 shows that our hybrid model with linear SVM performed the best among all in terms of all the used performance metrics. When the features from a hybrid feature extraction model were classified by using linear SVM, the classification accuracy was 0.877, the precision was 0.865, the recall was 0.746, the F1 value was 0.865, and the Kappa value was 0.728, as shown in Table 4.

In the second experiment, we used the data from the fifth week as a training set and the data from all other weeks as a testing set. As explained in the Method section, we used the combination of PCA and LDA to extract the features that were passed to the classifier models, linear SVM and kernel SVM. Also, for the second experiment, we extracted forty-four features by using PCA and sixteen features by using the LDA technique. These numbers were determined by running the experiment several times with a different number of features and the number that yielded the highest accuracy was considered. The performance of the classifiers was measured by using the confusion matrix. To compare the results of our hybrid model, we used a different combination of feature extraction

techniques and classifiers. The confusion matrices when using the fourth week data as the testing set are shown in Figure 7.

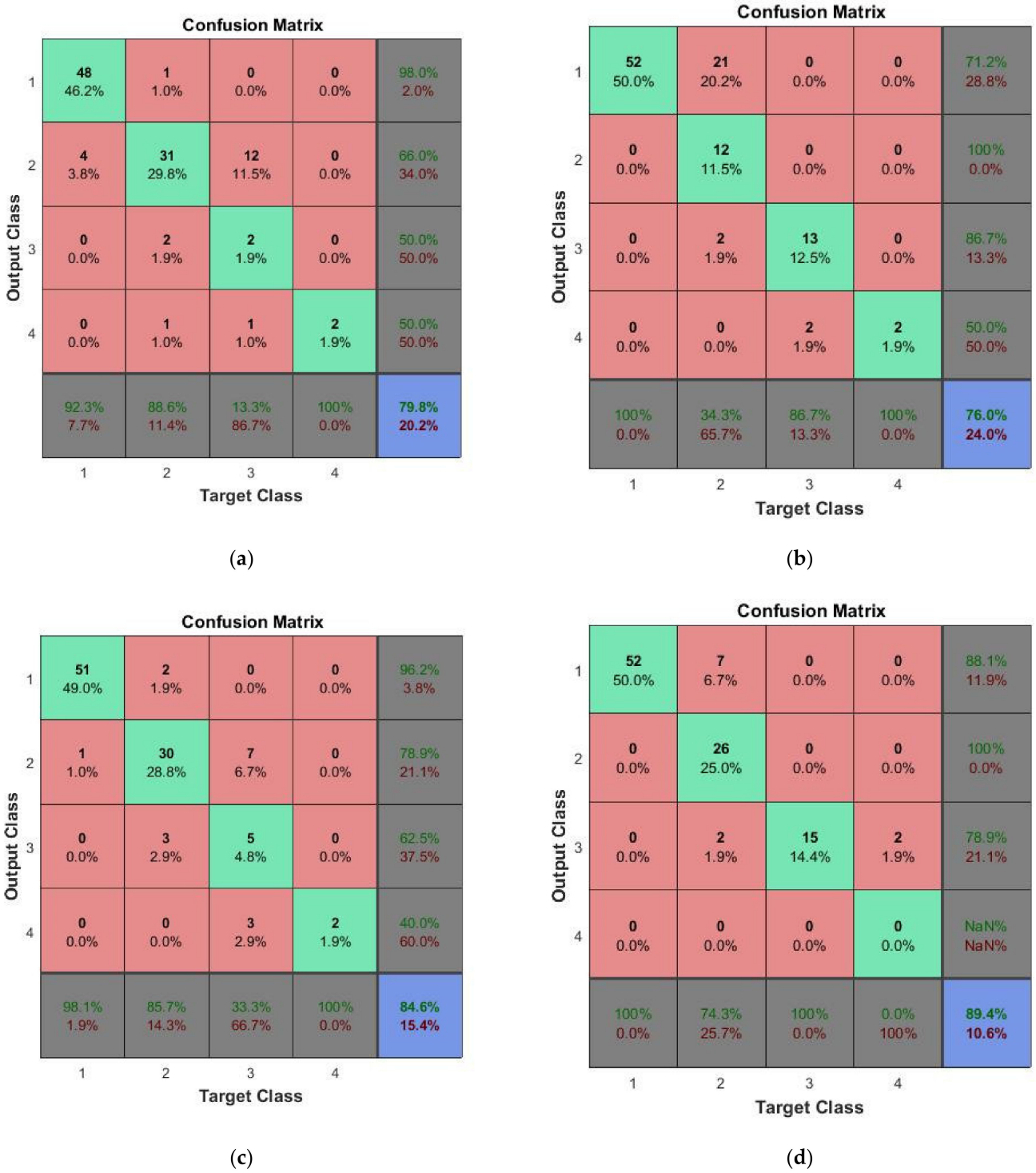



**Figure 6.** *Cont.*

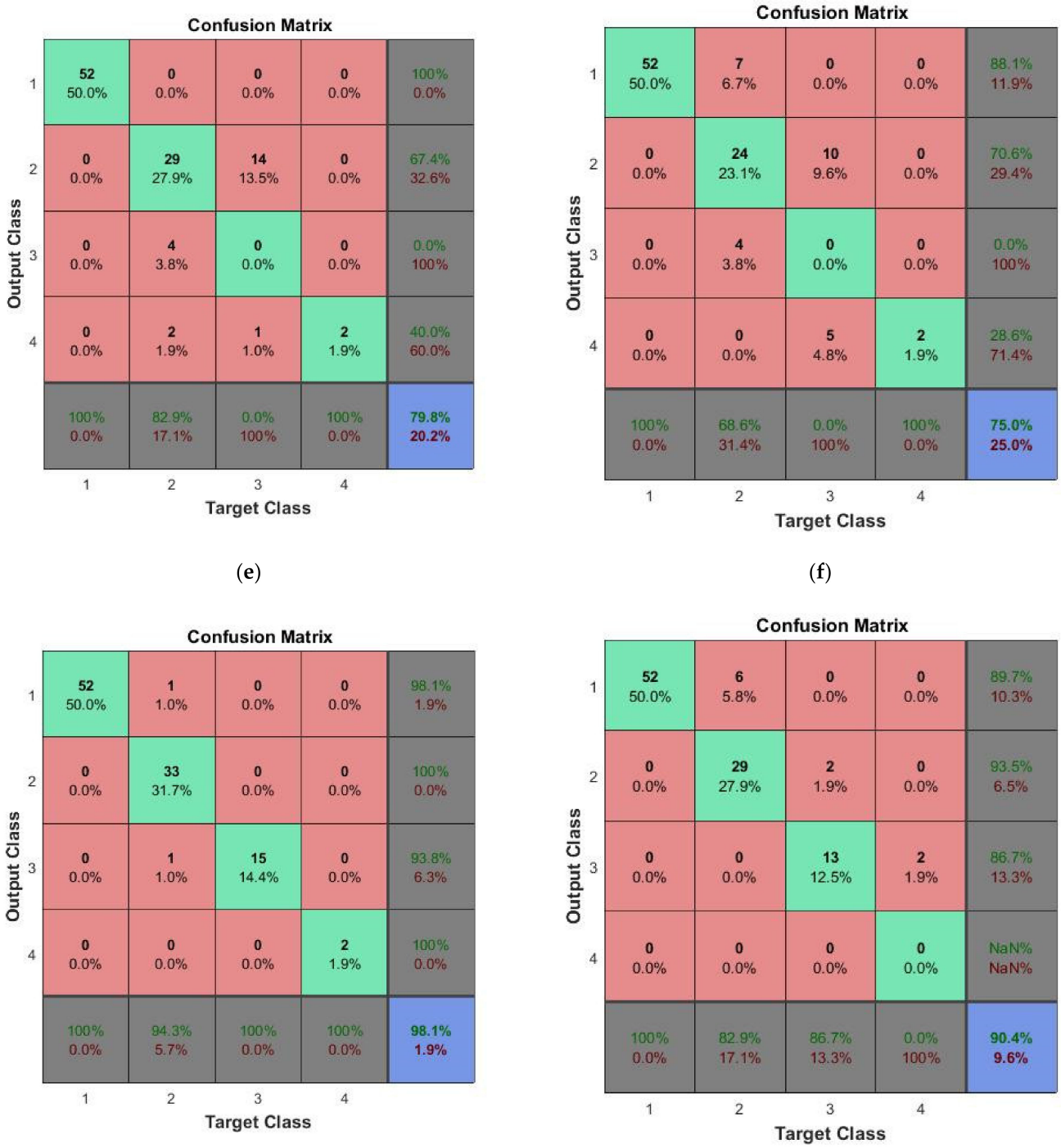



**Figure 6.** This figure represents the confusion matrix for the first experiment using (**a**) only linear SVM, (**b**) only kernel SVM, (**c**) only PCA feature extraction and linear SVM, (**d**) only PCA feature extraction and kernel SVM, (**e**) only LDA feature extraction and linear SVM, (**f**) only LDA feature extraction and kernel SVM, (**g**) hybrid feature extraction and linear SVM, and (h) hybrid feature extraction and kernel SVM.

Figure 7 represents the confusion matrix for the second experiment, taking the dataset from week five as a training set and the dataset from week four as a testing set. On the confusion matrix, the diagonal elements are the truly classified samples. Similarly, all other confusion matrices are also obtained with all the remaining week's data as a testing

set. With the help of these confusion matrixes, we now calculate different performance matrices by using Equations (3)–(7). Many results were obtained because our classification model was run with one training set and seven different testing sets. The results were then averaged and are shown in Table 5.

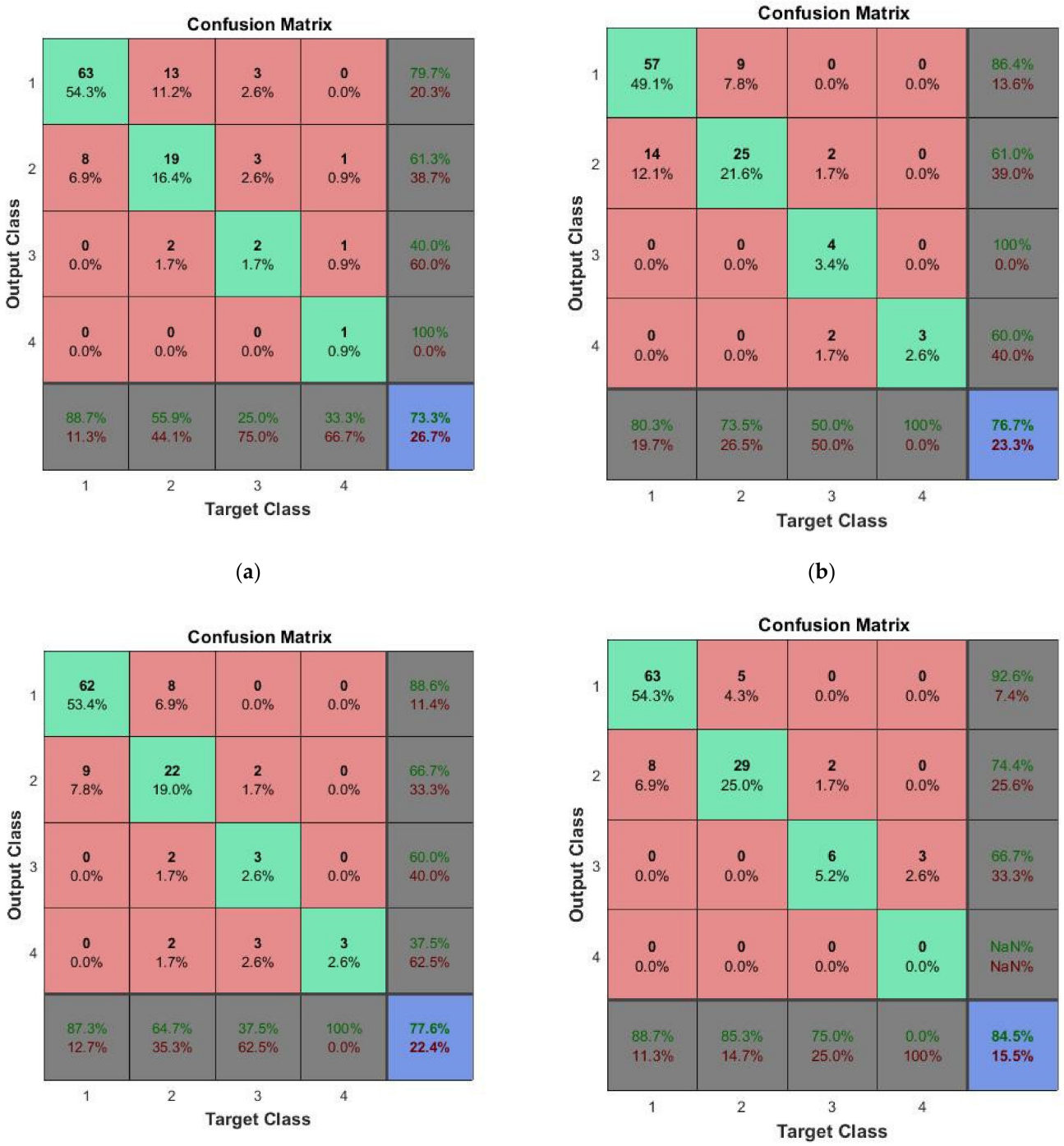



**Figure 7.** *Cont*.

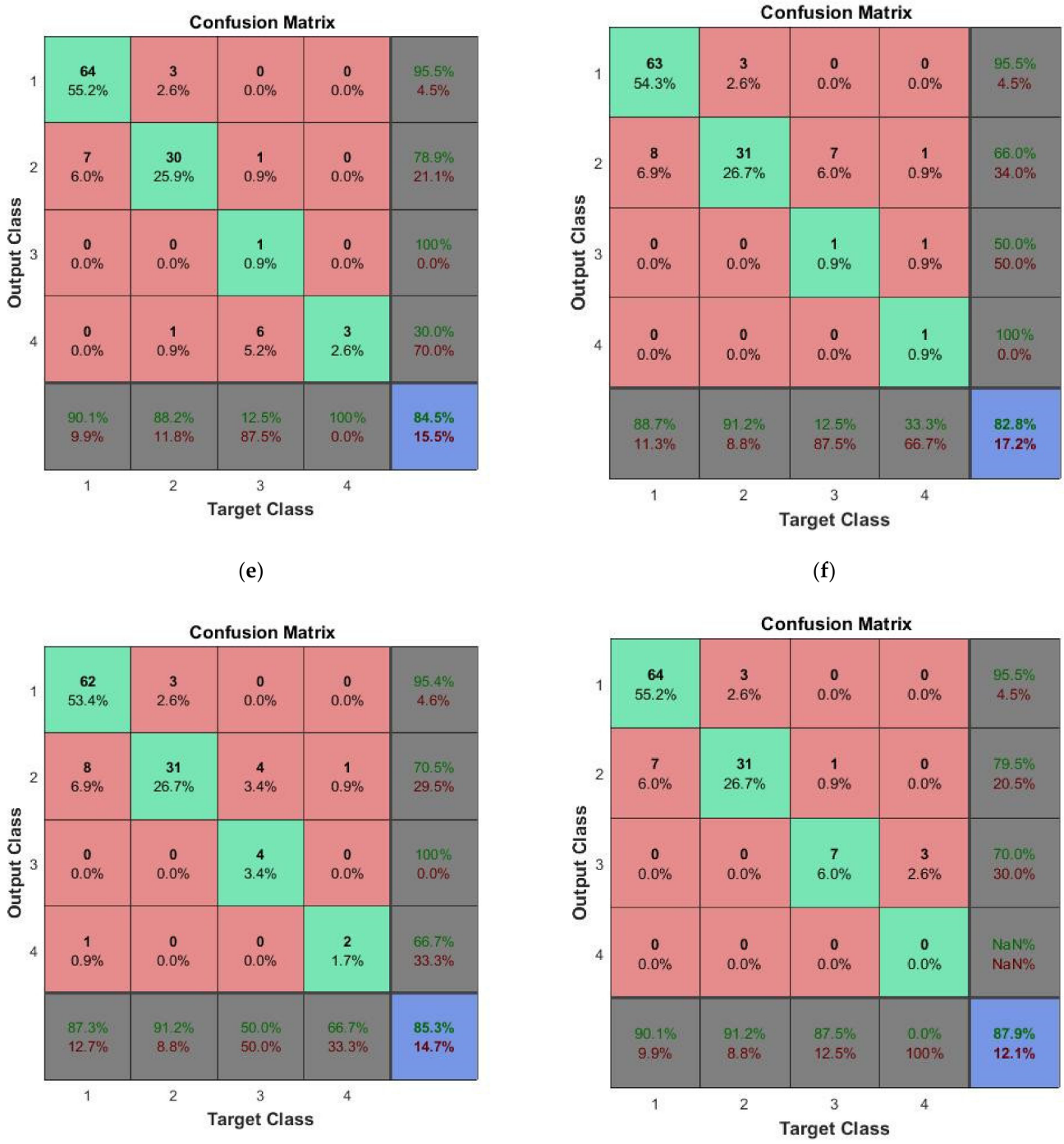



**Figure 7.** This figure represents the confusion matrix for the second experiment using (**a**) only linear SVM, (**b**) only kernel SVM, (**c**) only PCA feature extraction and linear SVM, (**d**) only PCA feature extraction and kernel SVM, (**e**) only LDA feature extraction and linear SVM, (**f**) only LDA feature extraction and kernel SVM, (**g**) hybrid feature extraction and linear SVM, and (**h**) hybrid feature extraction and kernel SVM.

Table 5 shows that our hybrid model with linear SVM performed the best among all in terms of all the used performance metrics except accuracy and kappa value, for which a hybrid model with kernel SVM performed the best. When the features from a hybrid feature extraction model were classified by using kernel SVM, the classification accuracy

was 0.915, and the kappa value was 0.833. When using the linear SVM, the precision was 0.883, the recall was 0.785, and the F1 value was 0.829, as shown in Table 5. In Tables 4 and 5 some of the values are Nan. Many reasons exist for this. First, the samples may have been wholly false negatives, and second, there may have been no samples at all.

**Table 5.** Comparison of Different Techniques for the second experiment.

| | **Accuracy** | **Precision** | **Recall** | ***F*1** | **Kappa** |
|---|---|---|---|---|---|
| Linear SVM | 0.683 | 0.58 | 0.437 | 0.508 | 0.318 |
| Kernel SVM | 0.824 | 0.775 | 0.774 | 0.774 | 0.651 |
| PCA and Linear SVM | 0.81 | 0.668 | 0.726 | 0.695 | 0.654 |
| PCA and Kernel SVM | 0.839 | Nan | 0.617 | Nan | 0.67 |
| LDA and Linear SVM | 0.87 | 0.662 | 0.694 | 0.675 | 0.754 |
| LDA and Kernel SVM | 0.87 | 0.797 | 0.636 | 0.715 | 0.753 |
| Hybrid method and Linear SVM | 0.908 | **0.883** | **0.785** | **0.829** | 0.822 |
| Hybrid method and Kernel SVM | **0.915** | Nan | 0.708 | Nan | **0.833** |

We also compared the performance of our model with previous studies as shown in Table 6.

**Table 6.** Performance comparison of our model with previous studies.

| **Authors** | **Dataset Used** | **Techniques** | **Accuracy** |
|---|---|---|---|
| Thomas et al. [44] | Videos captured dataset of students | SVM and LR | 90% |
| Zhang et al. [45] | Private dataset collected from students in classroom | J48 decision tree, Random Forest, SVM | 82% |
| Zaletelj et al. [16] | Video and 3D data recorded by Kinect One sensor | Different traditional machine learning algorithms | 75.3% |
| Xiao et al. [43] | Eight different datasets | RnkHEU hybrid method that integrates ranking-based forward and heuristic search | 71.19% |
| Our hybrid model | Student attention dataset | Hybrid feature extraction | 91.5% |

Table 6 shows the comparison of our model with other previous techniques. From Table 6, our model that used the hybrid feature extraction technique outperformed all other models.

After the classification, we produced the new dataset by averaging all the features' values from eight different weeks for each student. We did this to observe the statistical relationship between the students' attention and the final grade of the students. As explained earlier, we used Pearson's correlation for this process. We also computed the *p*-value to determine whether the correlation between variables is significant or not. We used a significance level of 0.05. Table 7 shows the Pearson's correlation coefficient ($r$) and the *p*-values between the final grade and the attention features, namely ON% and switching. The ON% is obtained by dividing the total time a student was active by the total time he/she remains in the class. Switching is the total number of times the student changes from on-task to off-task and from off-task to on-task.

**Table 7.** Pearson's Correlation Coefficient ($r$) and *p*-value between Final Grade and the Attention Features.

| **Features** | ***R*** | ***p*-Value** |
|---|---|---|
| On-task attention percentage (ON%) | 0.15 | 0.05 |
| Total number of switches between the on-task and off-task (Switching) | −0.06 | 0.44 |

We also calculated the statistical relationship between the attention pattern and the results of two tests. Test 1 was conducted after the fourth week. We averaged all the feature values up to week four. The new data was obtained by doing so. We then calculated the Pearson correlation coefficient ($r$) and $p$-value between the test 1 and attention features, which is shown in Table 8.

**Table 8.** Pearson's Correlation Coefficient ($r$) and $p$-value between Test 1 and the Attention Features.

| Features | *R* | *p*-Value |
|---|---|---|
| On-task attention percentage (ON%) | 0.025 | 0.0009 |
| Total number of switches between the on-task and off-task (Switching) | −0.05 | 0.45 |

Test 2 was conducted after the eighth week. So, we averaged all the feature values from week five to week eight. The new data was obtained by doing so. We then calculated the Pearson correlation coefficient ($r$) and $p$-value to determine the statistical relationship between the attention features and the test 2, which is shown in Table 9.

**Table 9.** Pearson's Correlation Coefficient ($r$) and $p$-value between Test 2 and the Attention Features.

| Features | *R* | *p*-Value |
|---|---|---|
| On-task attention percentage (ON%) | 0.16 | 0.02 |
| Total number of switches between the on-task and off-task (Switching) | 0.012 | 0.87 |

## 5. Discussion

We used the student's attention dataset, obtained from the attention patterns of eight different weeks, where all the participants were the subset of the total participants in the study as explained in Section 3. As explained in the Introduction section, the first objective of this study was to characterize students' attention behaviors in binary form. We did this by collecting the data in the form of screenshots at the interval of every 20 s at each lecture class throughout the semester and processing it by using a custom MATLAB 9.0 script to obtain the students' attention patterns in the form of one (on-task) and zero (off-task), which is explained in Section 3.1. The second objective was to classify the student's attention behavior and improve the performance of the classification model by using feature extraction techniques. Because we have a small dataset in our study as explained earlier, we used SVM as a classifier. We used only linear or kernel SVM to classify the attention patterns of students and again used these with feature extraction techniques such as PCA and LDA Tables 4 and 5 show the results of using the classifiers alone and the classifiers with the feature extraction techniques. We can see in Tables 4 and 5 that when we use the feature extraction technique, the classification accuracy is improved compared to without it. Thus, it answers our first research question that feature extraction techniques improve the classification accuracy for the classification of active window/attention behavior of students. The main disadvantage that we faced with SVM in our study was that it has many parameters, such as regularization, gamma, probability, etc., and we need to set these parameters correctly to achieve the best classification results.

The third objective of our study was to build a hybrid model that further improves our classification accuracy for classifying the binary attention behaviors of the students. We achieved this by using hybrid feature extraction methods prior to classification, as explained in the Method section. For the hybrid methods, we used a combination of PCA followed by LDA feature extraction techniques. The obtained features are classified by using linear and kernel SVM. As explained in the Method section, we performed two different experiments for the classification. For the first experiment, we took the dataset of a fourth week as the training set and the dataset of all the remaining seven weeks as the testing set. For the second experiment, we took the dataset of a fifth week as the training set

and the dataset of all the remaining weeks as the testing set. Bold values in Tables 4 and 5 represent the best results among all. For the first experiment, Table 4 shows that when the linear SVM is used with the hybrid feature extraction method, the result is the best in terms of all the performance metrics. Moreover, the results from Table 5 show that the classification performance is improved when using the hybrid model. The higher model performance for the hybrid method, as seen in Tables 4 and 5, answers our second research question: the hybrid method improves the performance of the model for classifying the active window/attention behavior of students. From Tables 4 and 5, it is seen that the performance of using only the linear SVM and kernel SVM without any feature extraction techniques is very poor. It is because the original dataset contains both redundant and irrelevant features that the model performance can be negatively impacted, as we can see in our case too. As shown in Tables 4 and 5, with the hybrid feature extraction technique, the accuracies are improved.

Table 6 shows the comparison of our model with other previous techniques. Thomas et al. used student's facial expression, head pose, and eye gaze to distinguish the attention level of students [44]. Reference [44] used models such as SVM and LR where, SVM with radial basis function performed the best with an accuracy of 90%. Zhang et al. proposed an attention inference engine using rule-based approach or data-driven approach and used machine-learning algorithms such as J48 DT, RF, and SVM for the classification purpose [45]. J48 DT obtained the highest accuracy of around 82%. Zaletji et al. used 2D and 3D data obtained by a Kinect One sensor that includes facial and body properties of students [16]. Seven different traditional machine-learning algorithms were used to predict the time-varying attention level of students and obtained the moderate accuracy of 75.3%. Xiao et al. used hybrid feature selection method RnkHEU that integrates ranking-based forward and heuristic search for predicting the academic performance of students [43]. Different classifiers such as NB, C4.5, MLP, and KNN were used as classifiers and the RnkHEU method improved the classification accuracy by 10% with the highest accuracy being 71.19%. On comparing previous studies, our model performed the best for classifying the students' attention level, and also our hybrid feature extraction model obtained the highest classification accuracy of 91.5%.

The fourth objective was to find a correlation between the attention patterns and the learning based on progress exam scores. To achieve this, we observed the statistical relationship between the students' attention patterns and learning after the classification. It is measured in terms of test grades and final grades. For the students' attention features, we selected two features, ON% and switching, as explained in the Results section. We performed the Pearson's correlation and calculated the $p$-value with a significance level of 0.05. The value of Pearson's correlation coefficient lies somewhere between $-1$ and +1. We can interpret the Pearson's coefficient value as: for the value of nearly $\pm 1$, it is a perfect correlation. That is, if one variable increases, the other variable also increases (for +1) or decreases (for $-1$), and for the coefficient between $\pm 0.5$ and $\pm 1$, two variables have a strong correlation, and for $\pm 0.30$ and $\pm 0.49$ it is said to have a medium correlation, and for values less than $\pm 0.29$ it is said to have a small correlation, and for the value of 0, there is no correlation. The result of Pearson's correlation coefficient between the attention features and the final grade is shown in Table 7. The results show that there is not even a medium correlation between the attention features and the final grade, which is also supported by the large $p$-values. Tables 8 and 9 show the result of Pearson's correlation coefficient between the attention features and the test 1 grades, and the test 2 grades, respectively. Again, the results show that there is not even a medium correlation between the attention features and the test grades, which is again supported by the large $p$-values. All the results showed a very limited correlation between attention and learning. These results help to answer our third research question: the classified students' attention patterns (based on active window data) are not highly correlated to their progress exam scores. There could be multiple possible reasons for this. First, because the course is a first-year engineering course, it is a relatively easy course and students may have some prior knowledge of the

course. It can be explained by the fact that nearly 40% of the students who obtained a grade A in the course fall into an attention class greater than class 2 (i.e., they have a higher off-task period than the on-task period). Second, it may be that students did not engage in cognitively demanding off-task behavior.

We have some limitations in our study that form threats to the internal validity (trustworthiness of our results) and external validity (extend results are generalizable). First, the way the attention pattern of the students is recorded for our work using the values zero and one in binary form, attention is simplified. In reality, attention is not a simplified construct. Students may be paying attention to the computer screen, but their minds may be wandering somewhere else. In such cases, our study shows that the student is on-task, but they are not. The claim is that the active window is serving as a good approximation for attention is supported by prior work, but the threat to validity related to equating an active window with attention is still present and worth acknowledging. Second, our choice to set four classes of attention patterns was based on an examination of the data by engineering education experts, who introduce their own bias. It is possible that more or less categories would be more appropriate for future examinations. The third limitation of our work is that the electronic monitoring produces known error in the dataset—a mean percent error at 4.28% and estimated a standard error of 0.82 for the specific classroom data used in our analysis. Although this error rate is relatively low for data captured within a classroom setting (versus a laboratory where variables can be controlled), the error is present and, to a degree, impacts the degree of confidence in our results. In addition to the above limitations, there is also a threat to external validity for our study. One threat is that the data were captured in a first-year engineering course with specific instructor policies and practices for using computers. This means that in different types of classes (e.g., upper-level engineering courses), attention patterns could be very different. Also, in cases where computers are not used for instructional activities, attention patterns based on active windows would be invalid. Thus, collecting data in those types of classrooms and then applying our classification techniques described herein may not be appropriate. That said, if data were collected in courses according to the methods of the original study that produce the dataset we used, the methods we described herein could be appropriate. Finally, although the majority of the above threats to validity are related directly to data collection and the dataset used in our analysis, the final threat relates to the sparseness of cases for a few of our classification categories. This limitation is directly related to the EDM algorithm classification accuracies reported in this paper. All these limitations and threats are important to note and address if possible in future studies, but they do not invalidate our findings or impact the comparisons of the various EDM methods we used because we used the same dataset for all comparisons.

## 6. Conclusions

Large amounts of data are obtained from the students in the classroom that can be used by educators to understand the behavior of the students and take positive action toward their welfare. Data mining is needed to obtain the information from such students' data. In our work, in the beginning, we begin with the research question of whether the feature extraction technique can be used to improve the classification accuracy of students' attention patterns. The answer to this question is supported by the higher classification accuracy by using feature extraction techniques as opposed to the use of only the classification algorithm as seen in Tables 4 and 5. The accuracy was increased to around 90% by using the hybrid methods, which explains our second research question. Based on the result, the instructor could determine whether the students are paying attention. We do not require 100% classification accuracy for the instructors to make the correct decision. For the hybrid method, we obtained an average classification accuracy of about 90%. That means if we say 55% are paying attention in the classroom, only about 49 to 53% are paying attention, which is an acceptable range. The classification error can be due to several reasons. First, the expert classification was done manually, so there may be some errors during such a

classification. Second, some of the students may have been missed out when collecting the data because of software errors. Third, during the plot of the attention pattern, some students did not fall into any of the categories, and they were removed from our study. We saw the poor correlation between attention and the learning as shown in the result section. This result helped to answer the third research question: the classified students' attention patterns (based on active window data) are poorly related to their progress exam scores. This result implies that attention alone does not impact learning.

In the future, we could increase the total number of classes so that all the students that were removed for not fitting into any of the four classes will be incorporated. To further improve the classification accuracy of the students' attention patterns, we could use some other advanced techniques like deep learning. In this study, we used the PCA technique, but in future we could use robust PCA which is a factorization-based approach with linear complexity. Also, by using feature extraction we could combine multiple feature selection and extraction techniques like the wrapper method, filter methods, and embedded methods. Because our dataset is small, explainable AI is an interesting topic that could be included in the future. Explainable AI is ideal for smaller datasets. We used only SVM classifiers, but in the future we could use other classifiers such as KNN, DT, RF, discriminative ridge machine, or some other deep-learning techniques. In our study, we have taken exams and test grades as learning. But learning is a broad term as it involves complex cognitive processes such as internal organization, elaboration, repetition, and review. Students may be paying attention, but they may not be good at memorizing the lecture content, so they will perform poorly on tests and the final exam. Consequently, it is not fair to relate attention with grades only. In the future, we could combine the attention pattern of the students with other aspects like intrinsic motivation levels, study habits, and correlate them with learning.

The study demonstrates how the educational institute could use hybrid feature extraction techniques to anticipate students' attention behaviors in the classroom and take appropriate action to assist the students. Our work has produced acceptable classification accuracy, so it can be used to produce a tool that would be helpful for the educational institute to understand the attention behavior of the students in the classroom where computers are used. It could also be used to provide feedback to the students about their performance in the classroom. Our model's great accuracy has prompted us to continue experimenting with students' data in the classroom by using data mining to create insights in the pedagogical sector.

**Author Contributions:** Conceptualization, S.P. and M.J.M.-A.; methodology, S.P. and J.E.B.; software, S.P. and M.J.M.-A.; validation, S.P.; formal analysis, S.P.; investigation, S.P.; resources, S.P.; data curation, S.P.; writing—original draft preparation, S.P. and M.J.M.-A.; writing—review and editing, S.P., M.J.M.-A. and J.E.B.; visualization, S.P.; supervision, M.J.M.-A.; project administration, M.J.M.-A.; funding acquisition, M.J.M.-A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in part by the National Science Foundation under grant 2047625.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Truong, D. More students are learning on laptops and tablets in class. Some parents want to hit the off switch. *Washington Post*, 2020. Available online: https://www.washingtonpost.com/local/education/more-students-are-learning-on-laptops-and-tablets-in-class-some-parents-want-to-hit-the-off-switch/2020/02/01/d53134d0-db1e-11e9-a688-303693fb4b0b_story.html (accessed on 25 September 2021).
2. Get the 411: Laptops and Tablets in the Classroom. Available online: https://www.educationworld.com/a_tech/tech/tech194.shtml (accessed on 25 September 2021).
3. Winstead, S. How to Implement 1:1 Technology using Tablets in the Classroom. *My Elearning World*. 2016. Available online: https://myelearningworld.com/10-benefits-of-tablets-in-the-classroom/ (accessed on 25 September 2021).

4. Virtual Reality in Education: Benefits, Tools, and Resources. Available online: https://soeonline.american.edu/blog/benefits-of-virtual-reality-in-education (accessed on 25 September 2021).
5. Baig, M.I.; Shuib, L.; Yadegaridehkordi, E. Big data in education: A state of the art, limitations, and future research directions. *Int. J. Educ. Technol. High Educ.* **2020**, *17*, 44. [CrossRef]
6. Ferguson, R. Learning analytics: Drivers, developments and challenges. *Int. J. Technol. Enhanc. Learn.* **2012**, *4*, 304. [CrossRef]
7. Poudyal, S.; Mohammadi-Aragh, M.J.; Ball, J.E. Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. *Electronics* **2022**, *11*, 1005. [CrossRef]
8. Okubo, F.; Yamashita, T.; Shimada, A.; Ogata, H. A neural network approach for students' performance prediction. In Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK '17), Vancouver, BC, Canada, 13–17 March 2017; Association for Computing Machinery: New York, NY, USA, 2017.
9. Kim, B.H.; Vizitei, E.; Ganapathi, V. GritNet: Student Performance Prediction with Deep Learning. *arXiv* **2018**, arXiv:1804.07405.
10. Poudyal, S.; Morteza, N.; Mohammad, N.; Ghodsieh, G. Machine Learning Techniques for Determining Students' Academic Performance: A Sustainable Development Case for Engineering Education. In Proceedings of the International Conference on Decision Aid Sciences and Applications (DASA), Online, 8–9 November 2020.
11. Nagahi, M.; Jaradat, R.; Nagahisarchoghaei, M.; Ghanbari, G.; Poudyal, S.; Goerger, S. Effect of Individual Differences in Predicting Engineering Students' Performance: A Case of Education for Sustainable Development. In Proceedings of the International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020.
12. Wei, W.; Han, Y.; Chunyan, M. Deep Model for Dropout Prediction in MOOCs. In Proceedings of the 2nd International Conference on Crowd Science and Engineering (ICCSE'17), Beijing, China, 6–9 July 2017; Association for Computing Machinery: New York, NY, USA, 2017.
13. Whitehill, J.; Mohan, K.; Seaton, D.; Rosen, Y.; Tingley, D. Delving Deeper into MOOC Student Dropout Prediction. *arXiv* **2017**, arXiv:1702.06404.
14. Xing, W.; Dongping, D. Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *J. Educ. Comput. Res.* **2019**, *57*, 547–570. [CrossRef]
15. Chinchu, T.; Dinesh Babu, J. Predicting student engagement in classrooms using facial behavioral cues. In Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education (MIE), Glasgow, UK, 13 November 2017; Association for Computing Machinery: New York, NY, USA, 2017.
16. Zaletelj, J.; Košir, A. Predicting students' attention in the classroom from Kinect facial and body features. *J. Image Video. Proc.* **2017**, *1*, 1–12. [CrossRef]
17. Bohong, Y.; Zeping, Y.; Hong, L.; Yaqian, Z.; Jinkai, X. In-classroom learning analytics based on student behavior, topic and teaching characteristic mining. *Pattern Recognit. Lett.* **2020**, *129*, 224–231.
18. Nigel, B.; Sidney, K.; Jaclyn, O.; Ryan, S.; Valerie, S. Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Trans. Interact. Intell. Syst.* **2016**, *6*, 2–26.
19. Goldberg, P.; Sümer, Ö.; Stürmer, K. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educ. Psychol. Rev.* **2021**, *33*, 27–49. [CrossRef]
20. Cetintas, S.; Si, L.; PingXin, Y.; Hord, C. Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Trans. Learn. Technol.* **2010**, *3*, 228–236. [CrossRef]
21. Peter, B.; Paul, B.; Penelope, B. Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learn. Instr.* **2011**, *21*, 715–730.
22. Driver, M. Exploring student perceptions of group interactions and class satisfaction in the web-enhanced classroom. *Internet High. Educ.* **2002**, *5*, 35–45. [CrossRef]
23. The Chronicle of Higher Education. Available online: http://www.chronicle.com (accessed on 25 September 2021).
24. Close, B.; Lipson, A.; Lerman, S. Wireless laptops as means for promoting active learning in large lecture halls. *J. Res. Technol. Educ.* **2006**, *38*, 245–263.
25. Finn, S.; Inman, J.G. Inman Digital unity and digital divide: Surveying alumni to study effects of a campus laptop initiative. *J. Res. Technol. Educ.* **2004**, *36*, 297–317. [CrossRef]
26. Faria, S.; Tina, W.; Nicholas, C. Laptop multitasking hinders classroom learning for both users and nearby peers. *Comput. Educ.* **2013**, *62*, 24–31.
27. Eric, J.; Corentin, G.; Salomé, C.; Tiphaine, C.; Séverine, E. Does multitasking in the classroom affect learning outcomes? A naturalistic study. *Comput. Hum. Behav.* **2020**, *106*, 106264.
28. May, K.E.; Elder, A.D. Efficient, helpful, or distracting? A literature review of media multitasking in relation to academic performance. *Int. J. Educ. Technol. High Educ.* **2018**, *15*, 13. [CrossRef]
29. Ramaphosa, K.; Zuva, T.; Kwuimi, R. Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools. In Proceedings of the International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 6–7 August 2018.
30. Mohammadi-Aragh, M.J.; Williams, C.B. Student attention in unstructured-use, computer-infused classrooms. In Proceedings of the ASEE Annual Conference & Exposition, Atlanta, GA, USA, 23–26 June 2013; ASEE: Atlanta, GA, USA, 2013; pp. 23–1093.
31. Mohammadi-Aragh, M.J. Characterizing Student Attention in Technology-Infused Classrooms Using Real-Time Active Window Data. Ph.D. Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2013.

32. Mohammadi-Aragh, M.; Ball, J.; Jaison, D. Using wavelets to categorize student attention patterns. In Proceedings of the IEEE Frontiers in Education Conference (FIE), Eire, PA, USA, 12–15 October 2016.
33. Gerjets, P.; Walter, C.; Rosenstiel, W.; Bogdan, M.; Zander, T.O. Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Front. Neurosci.* **2014**, *8*, 385. [CrossRef]
34. Qingshan, D.; Wu, Z. Students' Attention Assessment in eLearning based on Machine Learning. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Banda Aceh, Indonesia, 26–27 September 2018.
35. Sezer, A.; Inel, Y.; Seçkin, A.C.; Uluçınar, U. The Relationship between Attention Levels and Class Participation of First-Year Students in Classroom Teaching Departments. *Int. J. Instr.* **2017**, *10*, 55–68. [CrossRef]
36. Hembrooke, H.; Gay, G. The laptop and the lecture: The effects of multitasking in learning environments. *J. Comput. High. Educ.* **2003**, *15*, 46–64. [CrossRef]
37. Kraushaar, J.M.; Novak, D.C. Examining the affects of student multitasking with laptops during the lecture. *J. Inf. Syst. Educ.* **2010**, *21*, 241–251.
38. Punlumjeak, W.; Rachburee, N. A comparative study of feature selection techniques for classify student performance. In Proceedings of the 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, Thailand, 29–30 October 2015.
39. Francis, B.K.; Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J. Med. Syst.* **2019**, *43*, 162. [CrossRef] [PubMed]
40. Rawat, K.S.; Malhan, I.V. A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In Proceedings of the 2nd International Conference on Communication Computing and Networking, Chandigarh, India, 29–30 March 2018; Springer: Singapore, 2019.
41. Turabieh, H. Hybrid Machine Learning Classifiers to Predict Student Performance. In Proceedings of the 2nd International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan, 9–11 October 2019.
42. Amrieh, E.; Hamtini, T.; Aljarah, I. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *Int. J. Database Theory Appl.* **2016**, *9*, 119–136. [CrossRef]
43. Xiao, W. RnkHEU: A Hybrid Feature Selection Method for Predicting Students' Performance. *Sci. Program.* **2021**, *2021*, 167059. [CrossRef]
44. Thomas, C.; Jayagopi, D. Predicting student engagement in classrooms using facial behavioral cues. In Proceedings of the 1st ACM SIGCHI International Workshop, Glasgow, UK, 13 November 2017.
45. Zhang, X.; Wu, C.; Viger, P.F.; Van, L.; Tseng, Y. Analyzing students' attention in class using wearable devices. In Proceedings of the IEEE 18th International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM), Macau, China, 12–15 June 2017.
46. Poudyal, S.; Mohammadi-Aragh, M.J.; Ball, J.E. Data Mining Approach for Determining Student Attention Pattern. In Proceedings of the IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 21–24 October 2020.
47. Chong, P.; Zhilu, Z.; Zhao, K.; Chenglizhao, C.; Qiang, C. Nonnegative matrix factorization with local similarity learning. *Inf. Sci.* **2021**, *562*, 325–346. [CrossRef]
48. Chong, P.; Yongyong, C.; Zhao, K.; Chenglizhao, C.; Qiang, C. Robust principal component analysis: A factorization-based approach with linear complexity. *Inf. Sci.* **2020**, *513*, 581–599. [CrossRef]
49. Jian, Y.; Zhong, J.; Yang, J.; Zhang, D.; Frangi, A.F. Essence of kernel Fisher discriminant: KPCA plus LDA. *Pattern Recognit.* **2004**, *37*, 2097–2100.
50. Wangmeng, Z.; Zhang, D.; Jian, Y.; Kuanquan, W. BDPCA plus LDA: A novel fast feature extraction technique for face recognition. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 946–953. [CrossRef]
51. Deng, H.; Jin, L.W.; Zhen, L.; Huang, J. A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA. *Inf. Technol. IT* **2005**, *11*, 86–96.
52. Chong, P.; Qiang, C. Discriminative Ridge Machine: A Classifier for High-Dimensional Data or Imbalanced Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2595–2609. [CrossRef]
53. Abu, Z.L.M. Prediction of Student's performance by modelling small dataset size. *Int. J. Educ. Technol. High Educ.* **2019**, *16*, 27. [CrossRef]
54. Chandra, M.A.; Bedi, S.S. Survey on SVM and their application in image classification. *Int. J. Inf. Tecnol.* **2021**, *13*, 1–11. [CrossRef]
55. Pearson Correlation Coefficient: Introduction, Formula, Calculation, and Examples. Available online: https://www.questionpro.com/blog/pearson-correlation-coefficient/ (accessed on 26 September 2021).