Using Wavelets to Categorize Student Attention Patterns

M. Jean Mohammadi-Aragh, John E. Ball, and Donna Jaison
Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS, USA

***Abstract*—Wavelet analysis is a pattern recognition technique that has been used to analyze signals and images such as medical scans and hyperspectral imagery. We hypothesize that wavelets can also be used to analyze educational data. We explore the use of Haar wavelets to differentiate various patterns in a student attention data set. The data consists of a very large set of binary attention data (on-task/off- task) for individual students that was electronically captured at 20-second intervals. The data set contains information for multiple large lectures of a first year engineering course and was captured over an entire semester. For each lecture within a course, there are approximately 200 student attention charts with 150 entries per student record (3 per minute for 50 minutes). We investigate the accuracy of using a Haar wavelet classification system to distinguish student attention patterns in a given lecture. Our results and discussion demonstrate the usefulness of applying traditional engineering analysis techniques to educational data**

Keywords—learning analytics, data mining, large lectures

I. Introduction and Motivation

Initially driven by college of engineering laptop computer requirements laptops and tablets, and now mobile phones and other electronic devices are commonplace in engineering classrooms. While the debate continues as to whether these devices increase learning [1-3] or cause a distraction that decreases learning [4-7], the fact remains, they are present. We view their presence as an opportunity to use the students' devices as a data collection tool to examine how students learn in order to improve educational practice and to design interventions that promote student learning.

We ultimately envision a type of "educational alert system." For example, if students' devices report students' attention or engagement levels throughout a lecture, an instructor could initiate well-time active exercises to reengage the class when average attention drops below a certain threshold. Alternatively, a student's device could alert a student when they engage in prolong off-task behavior and encourage the student to reengage with lecture. An educational alert system could be beneficial to both instructors and students.

In addition to providing real-time behavior alerts, such a system could be used for real-time data collection. Current computer-classroom research relies heavily on surveys and observations (e.g., [1-3,5,7]), which captures an incomplete, and arguably inaccurate, depiction of students' in-class computer use. Electronic data collection has the potential to capture more accurate data (e.g., [4]) and more complete data (e.g., [6]). Data collection via existing student devices is an especially exciting opportunity in large lecture halls. In large lectures, students gain a measure of anonymity and are more likely to engage in off-task behavior [5]. In addition, off-task behavior spreads as a distracted student distracts their neighbors. By collecting data through existing student devices, researchers have the potential to significantly impact educational practice.

As others work towards increasing electronic data collection capabilities, one required component towards realizing an educational alert system is real-time, automated processing algorithms. When considering binary on-task / off-task data, wavelets, a pattern recognition and classification technique, have the potential to isolate at-risk attention patterns. Through this exploratory quantitative investigation detailed in this paper, we investigate the research question: *How accurately can a Haar wavelet classification system distinguish on-task and off-task attention patterns?*

This initial section of the paper has introduced our larger research goal and motivated the research problem discussed in this paper. In subsequent sections of this paper, we justify our focus on student attention by providing an overview of relevant learning theories (Section II), we provide a brief introduction to wavelet analysis (Section III), we detail our study methods (Section IV), we present the results of our classification experiments (Section V), we discuss our classification results in the context of our research question (Section VI), and we provide closing remarks and future research directions (Section VII).

II. Importance of Attention for Learning

Attention is an important construct to investigate in the classroom since it is a precursor to learning. Situated within the cognitive perspective on learning are theories of information processing that focus on how people process, remember, and use information (e.g., [8-9]). Though there are variations within this group of theories, the general consensus is that attention is required to move information from one's sensory register to long-term memory (permanent storage). Multiple educational researchers have established the direct link between learning and attention (e.g., [10-11]). Based on their work, it is evident that attention is a critical element of the learning process.

Robert Gagne is credited with shifting the information processing discussion from the research lab to the practical realm of instructional design with his introduction of the Conditions of Learning [12]. Gagne's original theory stipulates that there are nine instructional events that must occur for learning to take place, the first of which involves obtaining the learner's attention [13]. The nine instructional events are: 1) gain attention, 2) inform learners of the objective, 3) stimulate recall of prior knowledge, 4) present information, 5) provide guidance, 6) elicit performance, 7) provide feedback, 8) assess performance, and 9) enhance retention and transfer. While these nine events do not guarantee learning will take place, each event is a necessary condition for learning to take place. Gagne has revised his original theory since its inception [14]. However, throughout all revisions, attention has remained the critical first step for learning. Gagne's work emphasizes the importance of attention for learning in the classroom context. This requirement of attention for learning has driven the design of our data collection method, and focused our analysis on identifying patterns of attention that may be detrimental to learning.

III. Brief Introduction to Wavelets

Wavelets are used extensively in signal and image processing, due to their ability to simultaneously analyze signals in time and frequency, and efficient implementations [15]. Wavelets have been used in many data processing applications, including image and signal processing, time-series analysis, and image processing, [16] and image compression and denoising [17]. Wavelet processing allows analysis of phenomena occurring at multiple scales to be analyzed and processed.

The simplest wavelet is the Haar wavelet, which allows a function to be represented by an orthonormal basis of Haar wavelets. The discrete Haar decomposition analyzes a signal by repeatedly applying two filters, then subsampling the signal. A two-level Haar decomposition is shown in Fig. 1.

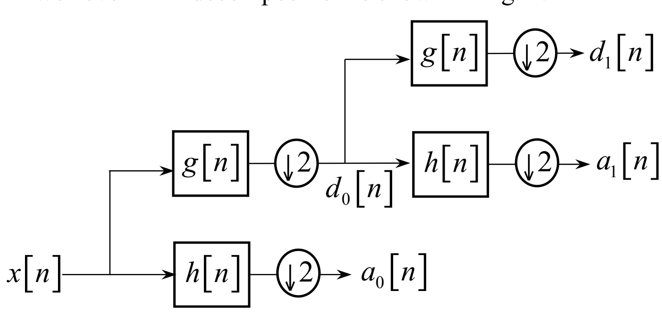

Fig. 1. Two-level Discrete Wavelet Transform (DWT) with Haar wavelet.

The two filters are defined by (1), which yields the detail coefficients, and (2). The approximation and detail coefficients at level m are $a_m[n]$ and $d_m[n]$, repsectively. The circles in Fig. 1 represent downsamplers: the output of the downsampler is a new sequence with every odd component removed.

$$h[n] = \frac{1}{\sqrt{2}}[-1,1] \tag{1}$$

$$g[n] = \frac{1}{\sqrt{2}}[1,1] \tag{2}$$

A. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) have been used extensively in classification and pattern recognition applications [18]. Properly trained SVMs can provide high classification accuracies [18]. SVMs have been used in a variety of applications where data requires classification. In this work, the linear SVM (L-SVM) is utilized. The L-SVM implementation is the LIBLINEAR library [19], which is an extension to the LibSVM library. The L-SVM solves the unconstrained optimization problem depicted in

$$\frac{1}{2}w^T w + C\sum_{j=1}^{L}\xi(w_j; x_j, y_j) \tag{3}$$

where $\boldsymbol{w} = [w_1, w_2, \dots, w_N]^T$ is the $[N \times 1]$ optimal weight vector, $\boldsymbol{x}_j$ is a $[N \times 1]$ feature vector, $y_j \in \{-1,1\}$ is the class associated with $\boldsymbol{x}_j$ (1=target, −1=no target), C is a penalty parameter for misclassifications, and ξ is a convex loss function, which allows for well-known convex optimization strategies to be employed in solving for the optimal weight vector, and L is the number of training samples. The loss function is used in classifier training; it will be near zero when a training sample is correctly classified, and will increase in value based on the distance to the classification boundary if the training sample is incorrectly classified.

In the training phase, target and non-target training data is presented to the L-SVM and the weight is optimized to best discriminate target cases from non-target cases. The target data in this case is simulated, and can also be simulated in a radar since targets may not be present. In the testing phase, new feature vectors are presented and they are considered targets if the L-SVM weighted feature z is greater than zero. The L-SVM implementation is very fast and computationally efficient. LIBSVM provides SVM implementations that support multi-class problems

B. Features

For each student, tasking was sampled every 20 seconds. A one indicated on-task, while a zero indicated off-task. The testing and training data had different class lengths, so the data was cropped to the shorter of the two. The Haar coefficients were created, and the following features were also used: 1) total of on task entries, 2) total of off-task entries, 3) total entries that were on-task followed by on-task, 4) total entries that were off-task followed by off-task, 5) total of two consecutive readings that were the same, 6) total of two consecutive readings with on followed by off, 7) total of two consecutive readings with off followed by on, 8) the time index of first on task, 9) the standard deviation of the students data.

The SVM parameters were chosen as the L2-regularized L2-loss support vector classification solver (-s 2 option) with cost set at 50 (-c 50) option. The cost was varied from 1 to 200 and 50 provided the best results. The other solvers provided worse results

IV. Methods

This study utilized electronically captured binary attention data (on-task/off-task) captured during two large lecture sections with the same instructor during two weeks in the midpoint of the semester. Data was classified by an expert and those classifications were compared to the output of a Haar wavelet based classification system.

A. *Setting*

Data were captured during a first-year engineering lecture at a large land-grant university in the southeastern United States. The participants were enrolled in the first semester of their engineering degree program and represented students from all engineering disciplines at the university. The three credit, semester-long course required students to attend one 50-minute lecture and one three-hour laboratory per week. The lecture section was taught in the sage-on-stage format with short active-learning activities interspersed throughout the lecture. The lecture was intended to introduce key topics that students would be gaining hands-on experience with during the laboratories later that week. Data were recorded in the lecture section only.

The data consisted of a very large set of binary attention data (on-task/off-task) for individual students that was electronically captured at 20 second intervals. The data was captured in two lectures of the same course. To capture the data, all 200 students brought their college of engineering required laptop computer to class and logged into the course software. A network connection facilitated communication between students' and the instructor's computers. Lecture slides were pushed to students' computers, and students' in-class exercise work was pulled from students' computers via the instructor interface. Our data collection method utilized this network connection and the course software to identify whether students were in the course software (on-task) or not in the course software (off-task). Data were captured with an Institutional Review Board (IRB) approval and with a waiver of informed consent. The waiver allowed data to be captured without altering student behavior by informing them of the study. Full details regarding our data collection procedures including IRB approval and validation considerations are available in [20].

B. *Expert Classification*

Attention data were recorded in two lectures during the mid-point of the course. Individual student records were examined and incomplete records were removed from analysis. Examples of incomplete records include those in which students 1) logged into course software late, 2) left the lecture early, and 3) had network issues that resulted in dropped data during the middle of the course. Removing incomplete data records resulted in 136 records for lecture 1 and 86 records for lecture 2. Note that the primary difference in participant numbers is due to lower attendance during lecture 2.

The data was graphed and examined by an expert for patterns. The expert was able to differentiate students who were primarily on-task or off-task but it appeared there were varying degrees of attention. Based on their recommendation, the data were classified into one of four classes as follows: Class 1: Primarily attentive students, Class 2: Students who were primarily attentive but had short inattentive periods (i.e., checked out of lecture), Class 3: Students who were primarily inattentive but had short attentive periods (i.e., checked-in to lecture), and Class 4: Primarily inattentive students. Examples of the various classes are shown in Figs. 2 – 5. Some student records could not be classified as one of these four classes, and those records were omitted from the study. In most cases, the omitted records were students who were approximately 50% attentive and 50% inattentive for the lecture. The distribution of expert classification type by lecture is given in Table 1.

TABLE I. DISTRIBUTION OF CLASSES BY LECTURE

Classification	Number of students per lecture	
	Lecture 1	***Lecture 2***
Class 1	73	44
Class 2	32	12
Class 3	8	5
Class 4	3	10
Not classified	20	15

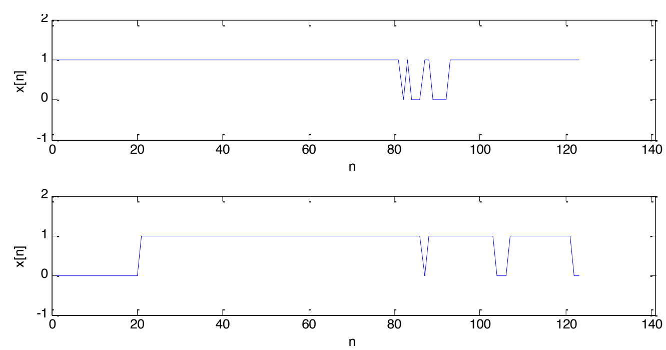

Fig. 2. Class 1 representative samples

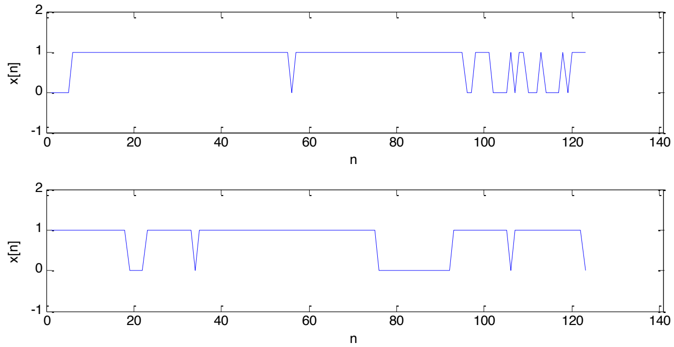

Fig. 3. Class 2 representative samples

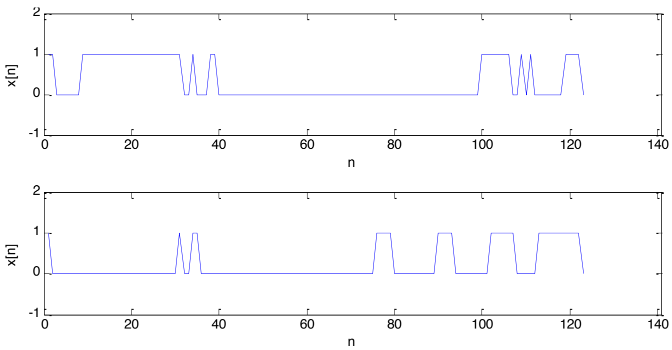

Fig. 4. Class 3 representative samples

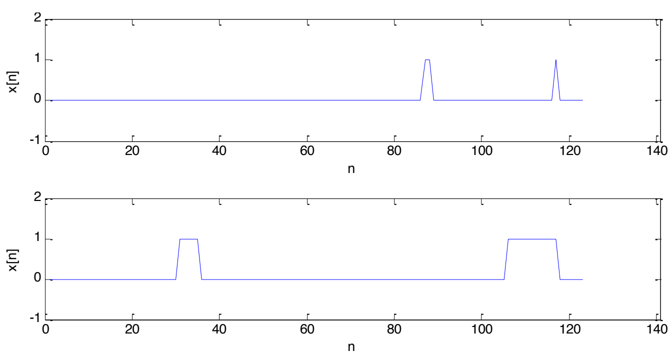

Fig. 5. Class 4 representative samples

C. *Haar Classificaiton and Confusion Matrices*

A standard Haar wavelet classifier was developed in MATLAB using the principles discussed in Section 3. For classification error rates, expert classification was considered "truth" data. The Haar wavelet classifier's classification was compared to the truth data using confusion matrices. A confusion matrix provides insight for a *M*-class problem using a *M* X *M* matrix, C, where C(*m,n*) is the number of classification results that are truly class *m* but the classifier result was class *n*. Correct entries are on the diagonal and incorrect ones are off the diagonal. The overall accuracy is the diagonal sum over the sum of all matrix entries. The user's accuracy for a class *m* is given in Eq. 4 and gives an estimate of how accurately the individual class results will be. In our case with four classes, there are four user accuracies.

$$UA(m) = \frac{C(m,n)}{\sum_{n=1}^{m} C(m,n)} \qquad (4)$$

V. Results

We conducted two experiments. In experiment 1, lecture 2 served as training data and lecture 1 served as testing data. In experiment 2, we swapped the training and testing data. The results of each experiment are detailed below.

A. *Experiment 1*

The classifier was trained on the 71 classified student records from lecture 2 (44 class 1, 12 class 2, 5 class 3, and 10 class 4). The Haar coefficients were extracted from the training data for each class. Running the linear classifier on the training data resulted in an accuracy of 100% (71/71). Running the linear classifier on the testing data resulted in an accuracy of 82.8% (96/116). The confusion matrix shown in Table 2 has training data in columns and linear classifier data in rows. For experiment 1, 72 entries were correctly classified as class 1, 15 as class 2, 6 as class 3, and 3 as class 4. The majority of the error was with 17 class 2 entries incorrectly classified as class 1. The user accuracies were 98.63%, 46.88%, 75.00%, and 100% for classes 1-4, respectively.

TABLE II. CONFUSION MATRIX FOR EXPERIMENT 1

	1	**2**	**3**	**4**
1	72	1	0	0
2	17	15	0	0
3	0	2	6	0
4	0	0	0	3

B. *Experiment 2*

The classifier was trained on the 116 classified student records from lecture 1 (73 class 1, 32 class 2, 8 class 3, and 3 class 4). The Haar coefficients were extracted from the training data for each class. Running the linear classifier on the training data resulted in an accuracy of 100% (116/116). Running the linear classifier on the testing data resulted in an accuracy of 90.1% (64/71). The confusion matrix shown in Table 3 has training data in columns and linear classifier data in rows. For experiment 2, 42 entries were correctly classified as class 1, 9 as class 2, 3 as class 3, and 10 as class 4. Unlike experiment 1, no single misclassification was responsible for the majority of the error. The user accuracies were 95.45%, 75.00%, 60.00%, and 100% for classes 1-4, respectively.

TABLE III. CONFUSION MATRIX FOR EXPERIMENT 2

	1	2	3	4
1	42	2	0	0
2	2	9	1	0
3	0	2	3	0
4	0	0	0	10

VI. DISCUSSION

To determine the sources of misclassification, we examined the classifier's ability to separate the classes at the feature level. We also examined the discriminate function separation. Finally, we examined the individual misclassified student records and compared them to student records in their "truth" class and records in their misclassified class.

A. Examining individual feature separation

We examined three individual features and their ability to separate the classes. First, we examined feature 120, the sum of all the zeros in the data. This feature corresponds to the number of instances a student was off-task. Fig. 6 is a plot of feature 120. Classes 3 and 4 are well separated by this feature. However, classes 1 and 2, and 3 and 4 have a large amount of overlap. This overlap means that the classifier is unable to correctly classify the classes based on this feature alone.

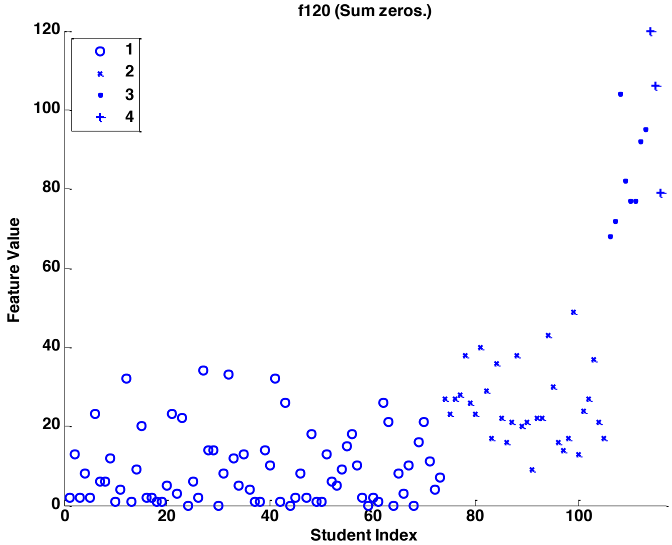

Fig. 6. Feature plot – sum of all the zeros (number of off-task instances).

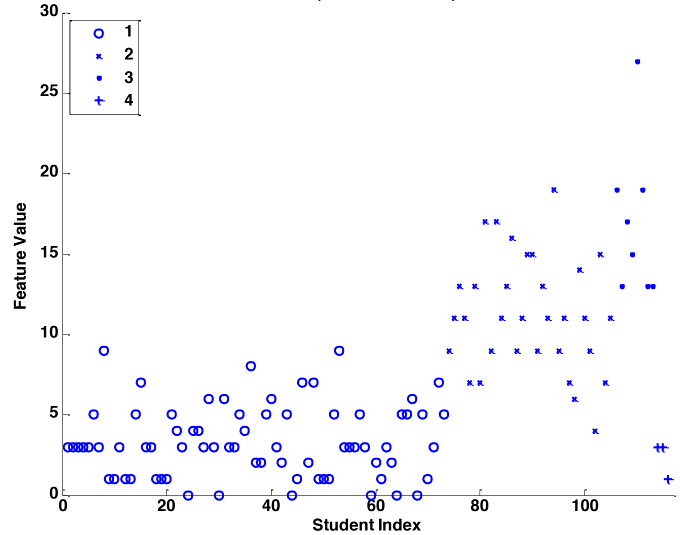

Fig. 7. Feature plot – sum of dx/dn not 0 (swapping on-task and off-task)

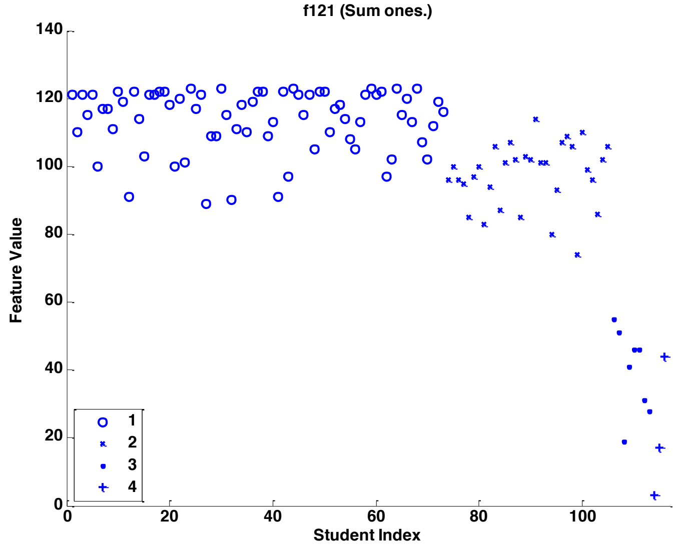

Fig. 8. Feature plot – sum of all the ones (number of on-task instances)

Second, we examined feature 125, the sum of derivative not zero. This feature corresponds to the number of times a student switch from on-task to off-task and is plotted in Fig. 7. Classes 1 and 4 have lower values, while classes 2 and 3 are higher due to the increase in students' swapping between on-task and off-task activities in these classes. Finally, we examined feature 121, the sum of the ones in the data. This feature plot in Fig. 8 shows that Classes 1 and 2 have a high amount of overlap, as well as classes 3 and 4. However, classes 1 and 2 are very well separated from classes 3 and 4. The Haar features all showed significant overlap between the classes, which was unexpected. However, the combinations of features selected do provide significant class discrimination capability (overall accuracy of 82.8% and 90.1%).

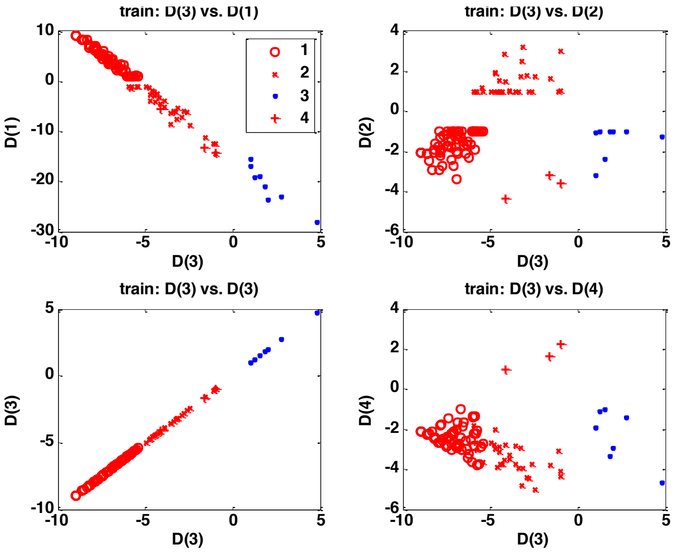

Fig. 9. Discriminate functions for training data for experiment 2

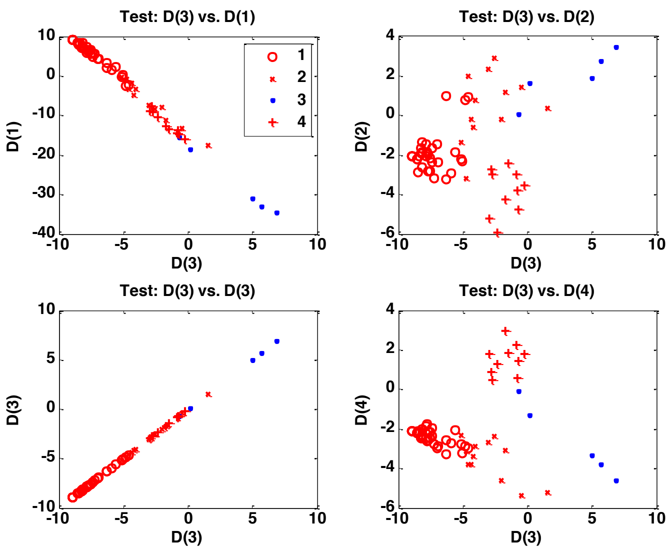

Fig. 10. Discriminate functions for testing data for experiment 2

B. *Examing discriminant functions*

With four classes, we have four discriminant functions. The SVM creates the optimal set of weights, w_n, for each class such that the classes are separated as much as possible. For a given input vector, x, the discriminate functions are $D_n(x) = w_n * x$ for each class n. The classifier chooses a class based on all the discriminant functions. For the interested reader, the details are in [21]. Fig. 9 shows the four discriminant functions for the training data for experiment 2 and highlights the classifier discriminant functions ability to separate the classes. In each case, the decision boundary is 0 and we would like the discriminant function to separate each class into a tight cluster, and have each cluster overlap with the other clusters as little as possible. The testing discriminant functions are shown in Fig. 10. Comparing the two, one can see more overlap between classes for the testing data. With few training samples and highly-overlapping features, SVMs can over-train and not generalize as well with new data. We believe this is a cause of error in our experiments.

C. *Examining misclassified data*

In Figs. 11-14, we plot the test data vectors for experiment 2. The classes start at time index 1 at the top and time progresses down (i.e., a student record is in a single column). The various class instances are plotted from left to right. In the figure, black indicates not paying attention, while green and red indicate paying attention. Green indicates the instance was classified correctly, and red indicates an incorrect classifier decision. As expected, for the students who are mostly attentive (Class 1 and 2), the plots in Figs. 11 and 12 are mostly green or red (on-task coloring). For the students who were primarily inattentive (Class 3 and 4) the plots in Figs. 13 and 14 are mostly black (off-task coloring). This visually confirms expert classification.

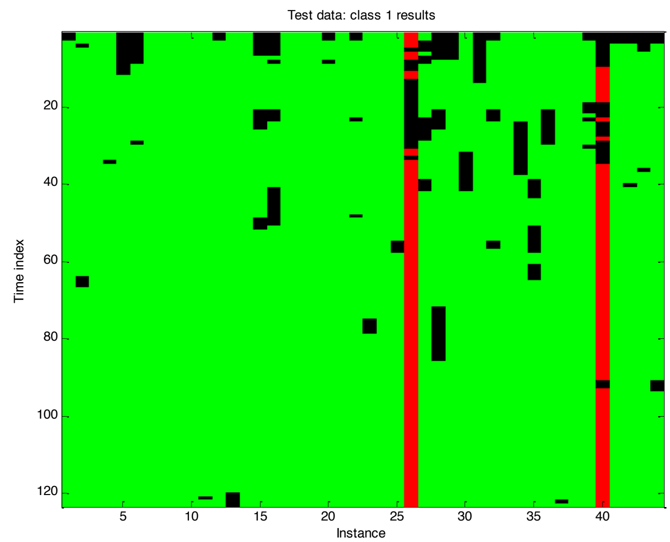

Fig. 11. Class 1 classification for testing data colored for correct classification (green) and misclassification (red)

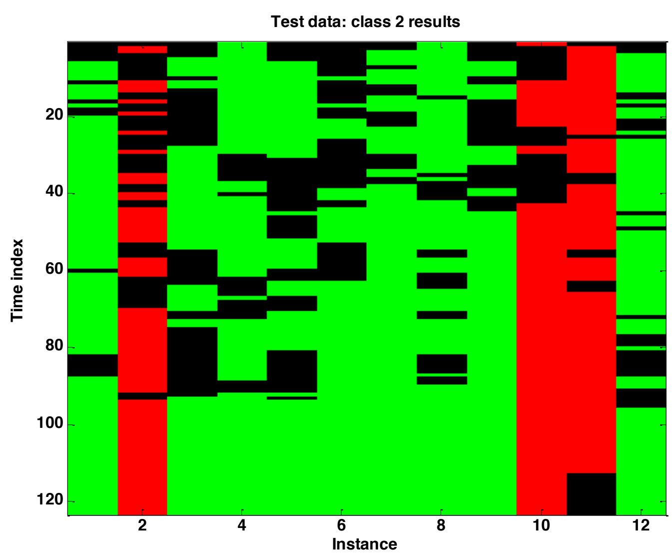

Fig. 12. Class 2 classification for testing data colored for correct classification (green) and misclassification (red)

Fig. 11 and 12 highlight how overlapping features cause confusion and misclassification. Fig. 11 shows misclassified test data vectors 26 and 40 (red). From the confusion matrix for experiment 2, we know these two instances were misclassified as class 2. Comparing instance 40 to instance 9 in Fig. 12, we see similarities in the patterns. In fact, one might argue that vector 40 is in actuality a class 2 instance, and the error is in the "truth" data that was classified by a human rather than in the automated classification system. Similarly, instances 10 and 11 in Fig. 12 were misclassified as class 1 by the classifier, and upon inspection, one could argue that they do indeed belong to class 1. Finally, the misclassifications between Class 2 and 3 (instances 2 in Fig. 12 and 1 and 2 in Fig. 13) highlight the similarity in the definitions of "mostly on task but checking out" and "mostly off task but checking in" as all those instances are nearly 50% black and red.

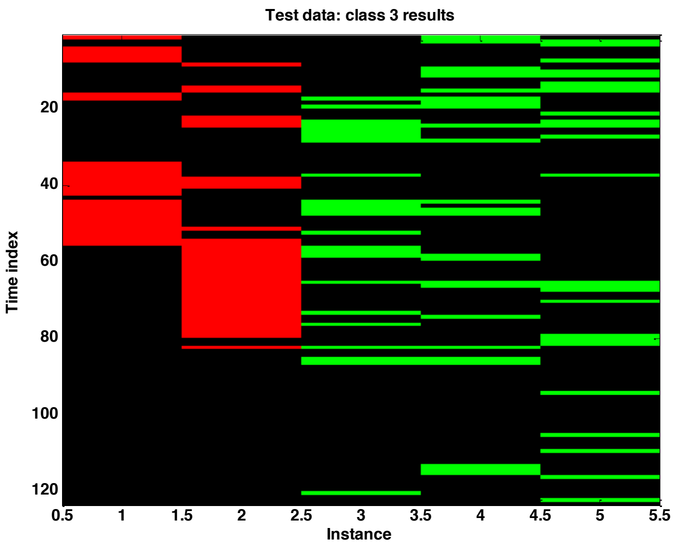

Fig. 13. Class 3 classification for testing data colored for correct classification (green) and misclassification (red)

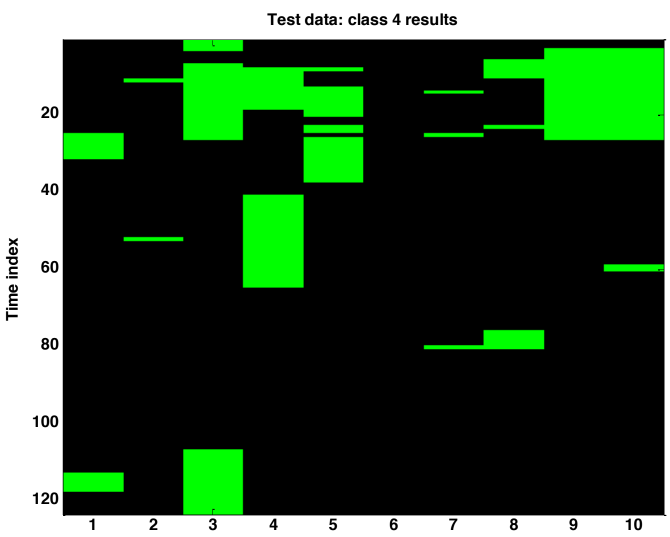

Fig. 14. Class 4 classification for testing data colored for correct classification (green) and misclassification (red)

VII. Conclusions and Future Work

While exploring the research question, "*How accurately can a Haar wavelet classification system distinguish on-task and off-task attention patterns?*", we found initial classification accuracy rates of 82.8% and 90.1%. There are two main factors that contribute to these error rates and are worth additional exploration.

First, the expert initial recognized and separated the data into four classes. However, during the course of the expert classification, numerous student attention records were not classified because they did not fit clearly into one of the four classes. While the patterns were easily described qualitatively and many student attention records did clearly fit the description, the definitions were difficult to quantify and resulted in many fuzzy cases. In the future, classification based on attention combined with a learning metric may be more useful. It is also important to examine the four classes to see if the number of classes should be increased or decreased. This recommendation is supported by the examination of how well the class features separated the testing data (Section VI.A) and the examination of the individual misclassified student attention records (Section VI.C).

Second, this investigation utilized a simple Haar-based linear classifier. In the future more advanced classification schemes can be examined. For example, windowing could be used to classify small sections of a student's attention record, which could then be combined into an overall attention classification. Windowing would be especially helpful in the event that a specific pattern of attention was shown to be detrimental to learning. As an example, consider the top representative sample in Fig.3. This student attention record has extensive task switching towards the end of the lecture. Windowing could be used to look for this specific pattern, and warn the student that they are engaging in activity that is detrimental to their learning. In addition to windowing, a redundant DWT could be utilized to remove the sampling associated with the two-level DWT used in this analysis (see Fig. 1). By employing more complex wavelet analyses, we may be able to further separate the student attention classes.

The positive results from this initial exploration of using wavelets to classify student attention records motivates us to continue our pursuit of developing classifiers for educational contexts. As stated in the introduction, we are motivated to develop tools that could provide feedback to educators regarding their instructional design. We are equally excited about the possibility of developing a system for alerting students to behavior that is detrimental to their learning so that they can better self-regulate their own learning. We further believe that it is possible to develop an understanding of attention patterns in large lectures, especially in regards to identifying those patterns that have a strong negative or strong positive impact on learning. Finally, with a robust classification system, we also believe it will be possible to examine attention patterns over time to understand whether students engage in certain behaviors throughout the semester or if inattentiveness is more likely due to, for example, an upcoming test in another class, rather than directly related to a specific lecture or course.

While our data and the focus on our study is large in-person lecture halls where electronic devices are ubiquitous, we are also excited about the possibility of using wavelets to analyze and classify other forms of educational data. Specifically, with an increase in online classes, distance education, and other electronically-delivered courses comes an opportunity to collect data through electronic devices. As engineering educators, we are excited about the prospect of mining this data for patterns and information that can help us understand more about learning in a technology-rich environment.

REFERENCES

[1] T. L. Doolen, J. D. Porter, and J. Hoag, "The relationship between PDA usage and student performance in an introductory engineering course." Journal of Engineering Education, 92, 263-268, 2003.

[2] V. Lohani, R. Castles, J. Lo, and O. Griffin, "Tablet PC applications in a large engineering program." Proc. 114th ASEE Annual Conference and Exposition, Honolulu, HI, 2007.

[3] P.J. Samson, "Deliberate engagement of laptops in large lecture classes to improve attentiveness and engagement." *Computers in Education* 20(2), 2010.

[4] H. Hembroke and G. Gay, "The laptop and the lecture: The effects of multitasking in learning environments," *Journal of Computing in Higher Education*, 15, 46-64, 2003.

[5] C. Fried, "In-class laptop use and its effects on student learning," *Computers & Education*, 50, 906-914, 2008.

[6] J.M. Kraushaar and D. C. Novak, "Examining the affects of student multitasking with laptops during the lecture," *Journal of Information Systems Education*, 21(2), 241-251, 2010.

[7] J. Connor, "Behind the laptops in a large lecture," in proceedings of the 3rd Annual Conference on Higher Education Pedagogy. Blacksburg, VA, February 3-4, p. 9-10, 2011.

[8] N. Cowan, *Attention and memory: An integrated framework*, Oxford Psychology Series vol 26. New York, NY: Oxford University Press, Inc., 1995.

[9] J.E. Ormrod (2008). *Human learning* (5th ed). Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[10] N. J. Mackintosh "A theory of attention: variations in the associability of stimuli with reinforcement." Psychological review 82, no. 4 (1975): 276.

[11] C.J. Mitchell, C. and M. E. Le Pelley. *Attention and associative learning: From brain to behaviour*. Oxford University Press, USA, 2010.

[12] M.E. Gredler, *Learning and Instruction: Theory into practice.* Upper Saddle River, NJ: Prentice-Hall, Inc.

[13] R.M. Gagne *The Conditions of* Learning. New York: Holt, Rinehart, and Winston, 1965

[14] R. M. Gagne, *Instruction and the conditions of learning. Psychology of school learning: Views of the learner*, 1, 153-175, 1974.

[15] Ingrid Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics (1992), SIAM, ISBN 0898712742.

[16] T. Li, Tao, et al. "A survey on wavelet applications in data mining." *ACM SIGKDD Explorations Newsletter* 4.2 (2002): 49-68.

[17] A.N. Akansu, W.A. Serdijn, and I.W. Selesnick. "Emerging applications of wavelets: A review." *Physical communication* 3.1 (2010): 1-18.

[18] D. He and H. Leung, "CFAR Intrusion Detection method Based on Support Vector Machine Prediction," IEEE Intl. Conf. on Comp. Intell. for Meas. Systems and Intell., 14 Jul 2004, pp. 10-15.

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Jrnl. Machine Learn. Rsch.*, Vol. 9, August 2008, pp. 1871 – 1874.

[20] M.J. Mohammadi-Aragh, "Characterizing student attention in technology-infused classrooms using real-time active window data", 2013.

[21] S. S. Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "A sequential dual method for large scale multi-class linear SVMs." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 408-416). ACM, 2008.