EXPLORING SEMI-SUPERVISED LEARNING FOR AUDIO-BASED AUTOMATED CLASSROOM OBSERVATIONS

Akchunya Chanchal and Imran Zualkernan
Computer Science and Engineering
American University of Sharjah, UAE

ABSTRACT

Systematic classroom observation is often used in evaluating and enhancing the quality of classroom instruction. However, classroom observation can potentially suffer from human bias. In addition, the traditional classroom observation is too expensive for resource-constrained environments (e.g., Sub-Saharan Africa, South and Central Asia). A cost-effective automation of classroom observation could potentially enhance both quality and resolution of feedback to the teacher, and hence potentially result in enhancing quality of instruction. Audio-based automatic classroom observation using supervised deep learning techniques has yielded good results in limited contexts. However, one challenge when using supervised techniques is the high cost of collecting and labelling the classroom audio data. One solution for such data-starved scenarios is to use semi-supervised learning (SSL) which requires significantly lesser data and labels. This paper explores an audio-adaptation of the state-of-the-art SSL FixMatch algorithm to automate classroom observation. An adaptation of the FixMatch algorithm was proposed to automate the coding for the Stallings class observation system. The proposed system was trained on classroom audio data collected in the wild. The supervised approach had an F1-score of 0.83 on 100% labeled data. The proposed FixMatch adaptation achieved an impressive F1-score of 0.81 on 20% labeled data, 0.79 on 15% labeled data, 0.76 on 10% labeled data, and 0.72 using only 5% of labeled data. This suggests that algorithms like FixMatch that use consistency regularization and pseudo-labeling have a great potential for being used to automate classroom observation using a small labelled set of audio snippets.

KEYWORDS

Classroom Observation, Audio Classification, Mel Spectrograms, Stallings, Semi Supervised Learning, Developing Country, FixMatch

1. INTRODUCTION

Quality and inclusive education is one of the United Nation's sustainability development goals. Education has long been considered as one of the most important factors in improving the quality of life of individuals. Increased quality of education has been shown to increase Gross Domestic Product (GDP) per capita by about 0.5% and consequently results in a decrease in Infant Mortality Rate (IMR) by about 0.6% (Jamison et. al, 2007). It is important to highlight that merely increasing the quantity of educational institutions does not result in reduction in poverty or IMR as seen in Tanzania (Wedgwood, 2007). Classroom teaching is backbone of most educational system and teachers represents a key leverage point for improving student learning outcomes, both in the long and the short term (Chetty et al, 2014). Therefore, improving classroom teaching practice and quality is paramount. Classroom observations provides effective feedback to teachers and can potentially result in better classroom practice and improved learning outcomes. For example, classroom observation has been shown to be an effective tool in assessing teacher quality in low-income countries like the Ugandan secondary schools where classroom observation was successful in quality of instructional practices with sufficient variance and high inter-rate reliability (Seidman et al., 2018).

Formally, classroom observation is defined as a tool that offers an unobstructed view and an understanding of how the teacher teaches in a realistic classroom setting (Martinez et al., 2016). There are many classroom observation systems (e.g., Stallings, CLASS, etc.) (Bell et al., 2018). One advantage of using the much simpler Stallings's observation system is that it was modestly correlated with the more expensive and complex CLASS

system. Benefits of Stallings's system have also been validated in developing countries in addressing student learning outcomes (Bruns et al., 2016).

Regardless of the observation systems being used, class observation as performed today is not without drawbacks. One obvious issue is that the observers may not hold a completely objective viewpoint, and may make a biased judgement based on their own preferences (Werner, 2018). Furthermore, it is entirely conceivable that the teacher may change their behavior when being observed. This may be more likely if the observer was an administrator. In addition, the students' behavior also may change as well. Another issue is that many classroom observation systems have been known to give inflated teacher evaluations (Werner, 2018). Classroom observation can also cause teachers further stress and anxiety as they are generally not accustomed to being observed and this causes them to believe that their professional competence is going to be questioned or judged (Borich, 2016). Finally, the low frequency with which classroom observation are conducted (Lam, 2001) means the teachers do not have an opportunity to correct their classroom practice midstream. Rather, they must typically wait for a year before receiving any feedback.

This paper is based on the Stallings class observation because this simpler approach lends itself easily to automation via deep learning. In specific, the use of a state-of-the-art semi-supervised learning algorithms that requires a fraction of labeled data was used to train a neural network model that can use classroom audio data to automatically extract Stallings's classroom observation categories. The remainder of the paper is organized as follows. Related work is discussed next, followed by a brief description of the Stallings classroom observation system. The semi-supervised learning (SSL) approach used is then briefly explained. This is followed by the results of the SSL-based system with different percentage of labelled data used along with the comparison against the fully supervised approach. The paper ends with a conclusion and future work.

2. RELATED WORK

Machine learning has been used previously to perform classroom observations. Typically, such systems use audio or multimodal data. Some examples of each are provided below.

2.1 Systems using Audio Data Only

Earlier, many approaches to automate classroom observation used traditional machine learning methods. For example., classroom audio was collected in a Randomized Control Trial (RCT) from 4th grade classrooms of low performing schools in Chile using a microphone connected to the teachers' mobile phones, length of such recording was between 18 minutes to 77 minutes (Schlotterbeck et al., 2021). The data was then classified using Random Forest into a modified version of the COPUS (Classroom Observation Protocol for Undergraduate STEM) where the different codes of the system are categorized into three categories; presenting, guiding and administration (Owens et al., 2017). Different representations of the data were used (e.g., Decibel scaled Mel Spectrograms, Mel Frequency Cepstral Coefficients (MFCC), and Amplitude) and fairly high Accuracy of 86%, 83%, and 96% respectively were reported for the three categories. However, there were large disparities in the recall scores across all categories (e.g., recall of 1 for negative samples and 0 for positive samples in the case of administration). Similarly, another system (Wang et al., 2014) used recordings of mathematics lessons from 13 different elementary schools in Michigan, ranging from the 1st grade to the 4th grade. The system was based on the Language Environment Analysis System (LENA) (Ford et al., 2008) and provided information about the quantity and distribution of classroom lectures. The teachers were asked to turn on the LENA recorder and put it in a pouch worn around their necks, and teach the lesson normally. At the end of the lesson, they were asked to turn off the recorder and connect it to a laptop. The data was then extracted from the recorder and sent to a remote server, where the data was analyzed to yield the four metrics of teacher talk, student talk, overlapping speech, and non-speech (silence or noise). The recorded audio was divided into 30 second segments and coded by 2 independent coders into 3 different activity types of teacher lecturing, the whole class discussion, and class group work. In each of the audio segments, each of the 4 metrics were calculated and were used as the independent variables to classify the audio into one of the three activities. Random Forest achieved classification accuracies of 88.1%, 79.7% and 83.3% for lecture, discussion and group work respectively leading to an overall classification accuracy of 84.37% (Wang et al., 2014). Another approach with audio only data was presented in (James et al., 2019) where data from 92 classrooms from

multiple preschools in Singapore was used. Each datapoint was about 20 minutes long in classrooms of 10-15 students and consisted of different types of activities such as small group discussions between students and teacher-student interactions. The audio was collected using a microphone worn by the teacher. Since this was a single microphone setup the audio was not of high fidelity and consisted of a large amount of background noise in addition to the audio of the teacher and students, even the audio between the teacher and student were not always intelligible. This was a largely accurate representation of a classroom environment where the activities are largely dynamic and uncontrolled. Preliminary speech detection was used to recognize and remove silences and any non-speech acoustic features from the audio recordings. Diarization was then performed using the LIUM toolkit (Meignier, Merlin, 2010). The conversational features were extracted from both the LIUM clusters and the low-level audio features like MFCC. Kruskal-Wallis test and other correlation algorithms were used for feature selection. The features selected were then passed through nine different conventional machine learning algorithms using 10-fold cross-validation to classify the lessons as having a negative or positive climate based on the CLASS observation system. The traditional machine learning methods achieved accuracy scores of between 70% and 80%.

2.2 Systems using Multimodal Data Only

Recently, there has been a surge in methods that utilize multimodal data (i.e., data from different modalities such as audio, video, text). Multimodal methods typically result in better performance since the different modalities are used in conjunction with each other resulting in a richer representation of the data giving the model more information (Summaira et al, 2021). One such system that utilized multimodal data was CLEVER (Classroom Evaluation and Video Retrieval) (Qiao & Beling, 2011) that used audio and video data from classrooms and classified them based on the CLASS protocol. CLEVER used video and audio metrics to bridge the gap between semantic assessment concepts to make them quantifiable and measurable. The system did this by relating the audio/video metrics with feature variables that could be using video and audio processing techniques such as topical detection, synchronization, silence detection, etc. These feature variables were then used to classify the classroom videos according to CLASS categories. The system used a variation of supervised learning known as Multiple Instance Learning (MIL) where instead of a label, there was a set of training bags. MIL is used in cases where there is ambiguity associated with the labels of the training examples. The dataset used for training and testing the system consisted of 40 video clips of 3 minutes each of upper-level systems engineering courses at the University of Virginia which were labeled by 10 expert CLASS coders. The model performed differently for labels set by each of the different coders, ranging from a minimum accuracy of 58.5% to maximum accuracy of 94.7%. The variance in the results of the model was attributed to the different interpretations of each video clip by the different coders.

While the aforementioned research shows the potential of using machine learning for automating aspects of classroom observation, however, none of these approaches addressed the issue of the cost of labelling the data. To the best of our knowledge, this paper is the first to present using an SSL based strategy to address the problem of classroom observation.

3. STALLINGS CLASSROOM OBSERVATION SYSTEM

This paper used the Stallings classroom observation system. The Stallings classroom observation system was developed by Jane Stallings in the late 1970s to evaluate the efficiency and quality of basic education teachers in the United States (Stallings, 1977). In this method, the human observer takes a 15 second 360-view of the entire classroom every 5 minutes in a 50-minute lesson. The observer notes down their observations on a structured coding sheet. This classroom observation systems evaluates the frequency of activities that the teacher conducts (e.g., question-answering, lecturing, classwork etc.), the materials that teacher utilizes (e.g., whiteboard, textbook, computer etc.) and the student group sizes that teacher is working with. Table 1 shows a summary of some of the activities that are coded under the Stallings system (World Bank Group, 2017).

Table 1. Classroom activities and their descriptions in the Stallings system

Activity	Description
Classwork	One or more students engaged in solving problems on the board, writing papers or engaged in any other conventional classwork activities.
Classroom Management	Teachers/Students engaged in classroom administrative tasks such as handing our graded exams, taking attendance, switching activities etc.
Lecture/Demonstration	Teacher or another medium explaining the academic content to the students.
Practice & Drill	Activities/tasks undertaken by the student in order to improve retention and reinforce the material such as multiplication tables, vocabulary, spelling etc.
Discussion/ Q & A	Students and/or teachers engage in a discussion/conversation in regards to some academic material e.g., clarification of queries, exchange of ideas.
Reading Aloud	The teacher and/or student reading aloud the contents from a medium e.g., presentation slides, textbook, paper etc.

Many of the Stallings system activities are not amenable to be detected using audio only. This paper considered only four Stallings categories of classroom management, lecture, practice and drill, and Q&A.

4. DATASET

The audio data used to train and test the models were obtained from Stallings-type classroom observation videos from semi-rural schools in Pakistan (Zualkernan et al., 2014). The classroom observation videos were collected by enumerators who used their mobile phone cameras and/or low-end video cameras to record the classroom videos. In total, 646 audio sessions were coded according to the Stallings coding manual (The World Bank, 2015). The 646 audio files were split into 3-second interval audio clips without overlap, resulting in 5,392 total audio clips. The audio data collected consisted of 263 female teachers and 383 male teachers across a variety of subjects such as English, Arabic, Mathematics and Science. Table 2 below shows the distribution of the four Stalling categories of classroom management, lecture, practice and discussion/Q&A across the different subjects.

Table 2. Distribution of audio clips according to subject

Subject	Classroom Management	Lecture	Practice	Q & A
Arabic	0	35	0	0
English	766	630	424	544
Math	580	1221	200	492
Science	147	353	424	0

5. METHODOLOGY

5.1 Proposed FixMatch Adaptation

FixMatch is a state-of-the-art semi-supervised learning (SSL) training algorithm for deep neural networks (Sohn et al., 2020). FixMatch uses consistency regularization and pseudo-labeling. Consistency regularization utilizes unlabeled data by in essence, assuming that the model will output similar prediction to inputs where perturbations have been applied and the same input when fed to the model as is (Bachman et al., 2014). This technique is used by many of the recent state-of-the-art SSL algorithms such as the Π-model and temporal ensembling (Laine, Aila, 2017). When using consistency regularization, the model is trained using both labeled data via a standard supervised learning classification loss function and on unlabeled data using something similar to (1). Consider $\gamma(.)$ to be a perturbation function which is applied to some input μ_b, where μ_b is b^{th} example in the batch. $(y|\gamma(\mu_b))$ represents the class distribution produced by the model for the perturbed example. B is the number of unlabeled examples in a batch.

$$\sum^{B}_{b=1} \|(y|\gamma(\mu_b) - p(y|\gamma(\mu_b)\|^2_2 \quad (1)$$

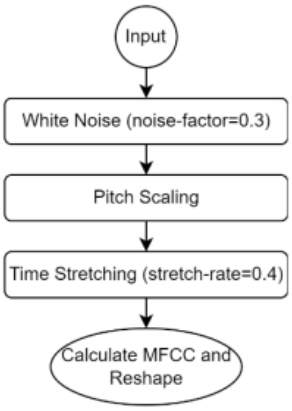
Since both p and γ are stochastic functions, the values of the two terms will be distinct. One is essentially calculating squared l^2 loss between the values, however cross-entropy loss between the two values can be used as well. Pseudo-labeling is the idea of using the model itself to obtain artificial or pseudo-labels for the unlabeled examples in the dataset. Examples, whose "hard labels" (i.e., $argmax$ of the predicted class distribution for the input) are greater than a predefined threshold τ are classified as such. (2) is used as the loss function for pseudo-labeling. Let $q_b = (y| \mu_b)$ and $\hat{q_b} = \arg\max (q_b)$. $\mathbb{1}(predicate)$ is a boolean function which is 1 when the predicate is true and 0 when the predicate is false. H is the cross-entropy loss function.

$$\frac{1}{B}\sum^{B}{}_{b}(\max (q_b \geq \tau)\ H(\hat{q_b}\ ,\ q_b) \qquad (2)$$

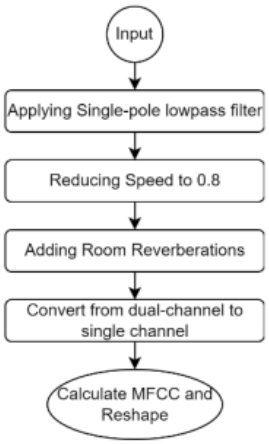
Using hard labels in pseudo-labelling makes it closely related to entropy minimization (Grandvalet & Bengio, 2004) where the model's prediction is encouraged to have low-entropy on unlabeled data.

FixMatch uses weak and strong transformations on the inputs of the model. Weak transformations slightly distort the input example, are initially applied to the input example, and used to produce the artificial label of the example. This label is then used as the target for the example when it is fed to the model after applying the strong transformations, while heavily perturbing the example.

This paper proposes the weak and strong augmentations for audio data as shown in Figure 1 (a) and Figure 1 (b).



(a) Weak Transformation



(a) Strong Transformation

Figure 1. Sequence of Weak Transformations Applied

The loss function for FixMatch uses two cross-entropy loss terms; a supervised loss ℓs applied to the labeled data and an unsupervised loss ℓ_u for the unlabeled data. ℓ_s (3) is the standard-cross entropy loss function that is used when using the weakly-perturbed examples as the input to the model. Consider β as the number of labeled examples in a batch. γ(.) represents a function which applies the sequence of weak transformations. x_b represents the b^{th} example in the batch.

$$l_s = \frac{1}{\beta}\sum\beta_{b=1}(p_b\ ,\ p(y|\ \gamma(x_b))) \qquad (3)$$

The unlabeled loss, ℓ_u computes cross-entropy loss between the artificial label $\hat{q_b}$ ($arg\ max$ of the model's predicted class distribution on the weakly transformed version of the same unlabeled example) and the model's output for the strongly transformed version of the unlabeled example. (4) represents the equation for the same. Consider Γ(.) to represent a function which applies the sequence of strong transformations, τ is a scalar hyperparameter which denotes a threshold above which the artificial label is retained. B is the number of unlabeled examples in a batch.

$$l_u = \frac{1}{B}\sum^{B}{}_{b=1}(max(q_b) \geq \tau)\ H(\hat{q_b}, p(y|\Gamma(x_b)) \qquad (4)$$

Both labeled and unlabeled loss are combined as in (5), which is the loss function ℓ minimized by FixMatch. λu is a fixed scalar hyperparameter which represents the relative weight of the unlabeled loss.

$$l = l_s + \lambda_u \, l_u \tag{5}$$

This paper proposes the adapted FixMatch algorithm shown in Fig. 2 to classify classroom observation audio. The classification Model used in the FixMatch algorithm is shown in Figure 3. The classification Model was inspired by the architecture described in (Salamon & Bello, 2017) which was successful in classifying audio data from the UrbanSound8K (Salamon et al., 2014) with 98.60% accuracy, and the ESC-10 and ESC-50 with 97.25% and 95.5% accuracy respectively (Piczak, 2015). The original CNN proposed in (Salamon & Bello, 2017) had three convolutional layers with 24, 48 and 48 layers respectively and used an input size of 128x128. However, (Zualkernan & Khan, 2020) found that using 5 convolutional layers with 128,96,96,32 and 64 filters respectively and an input shape of 8x16x1 gave better results on the classroom audio data. Batch normalization was applied after each convolutional layer and dropout is applied after the 1st, 3rd and 5th convolutional layer. This followed by three fully-connected layers. SGD optimizer with Nesterov momentum were used to train the model. The model was implemented using Pytorch.

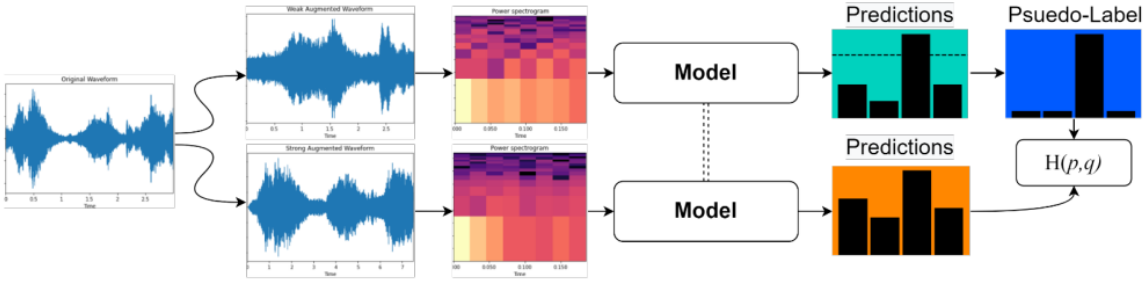


Figure 2. Adapted FixMatch Algorithm for classroom observation audio

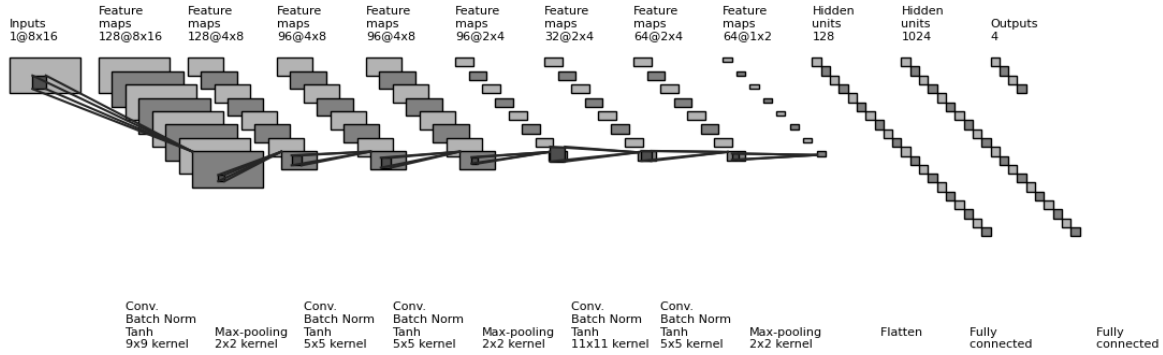


Figure 3. The classification Model used in the FixMatch Algorithm

5.2 Training and Testing

The original classroom audio data was resampled for each 3-second interval at 48 KHz and scaled between -1 and 1. The data was then split according into an 80/20 training and testing set such that there was equal number of examples of each of the four classroom activities in the test set. Accordingly, then, three separate training sets were generated, a labeled set where no transformations were applied, a weak set where the weak transformations were applied, and a strong set where the strong transformations were applied. For each of the training sets, the MFCCs were then calculated with n_mfcc value of 128 using the Librosa library. The mean across the first axis of the calculated MFCCs was then reshaped it into (8,16,1) to be fed to the CNN. The same training/testing scheme as proposed in (Sohn et al., 2020) was followed.

The Initial Learning rates for the model was set to 1.6e-5 for each of the different experiments, informed by the experiments performed in (Zualkernan & Khan, 2020). For a learning rate schedule, a cosine learning rate decay (Loshchilov & Hutter, 2016) was used which sets the learning rate to (6). η is the initial learning rate, k is the current training step and K is the total number of training step. For all models, the total number of training steps was 5000.

$$\eta \cos(\frac{7\Pi k}{16K}) \quad (6)$$

Finally, a batch size of 64 was used for training and threshold τ is set to 0.95. To observe the viability of the proposed adapted FixMatch, the proposed model was trained and tested under four experimental conditions and then compared with a fully supervised baseline. The four conditions used 20% labeled data, 15% labeled data,10% labeled data and 5% labeled data. Table 3 shows the number of training cases used for each condition.

Table 3. Number of training cases for various conditions

Activity	% Labeled data used for training				
	100%	20%	15%	10%	5%
Class Mgmt.	1223	225	163	108	54
Lecture	1969	225	163	108	54
Practice	354	225	163	108	54
Q&A	766	225	163	108	54
Total	4312	900	652	432	216

6. RESULTS

Table 4 shows the F1-scores of the trained systems under different experimental conditions. Table 4 clearly shows that even with a fraction of the amount of labeled data, the system performed well. For example, there was only a 0.08 drop in F1 score of the system when only 10% of labelled data used for training. There is, however, an expected larger drop of in performance when one moved from 10% to 5% of data. The results generally show that that the proposed architecture worked well and was able to use only 5% of the data to achieve very reasonable results.

Table 4. F1-Score of the proposed model (N = 1080)

Activity	% Labeled data used for training				
	100%	20%	15%	10%	5%
Class Mgmt.	0.83	0.81	0.79	0.76	0.72
Lecture	0.75	0.72	0.70	0.66	0.60
Practice	0.72	0.68	0.72	0.67	0.59
Q&A	0.61	0.62	0.54	0.50	0.51
Macro Avg.	0.73	0.71	0.69	0.65	0.61
Weight Avg.	0.73	0.71	0.69	0.65	0.61

As Figure 4 shows the confusion matrices under various conditions. As the Figure shows, there does not seem to be a systematic difference between the types of errors being made as the percentage of labels changes. In addition, Q&A seems to be confused with Practice. This could largely be attributed to the fact that both sets of activities follow similar patterns that consist of prolonged quiet spells with sporadic spoken responses. The ROC curves shown in Figure 5 also confirm that qualitatively the models seem to training in a similar manner. Figure 5 also shows that at lower percentages of labelling (5% and 10%) Lecture and Q&A are confused, most likely due to the sound predominantly being the voice of the teacher, however, this issue is largely handled better at larger label percentages.

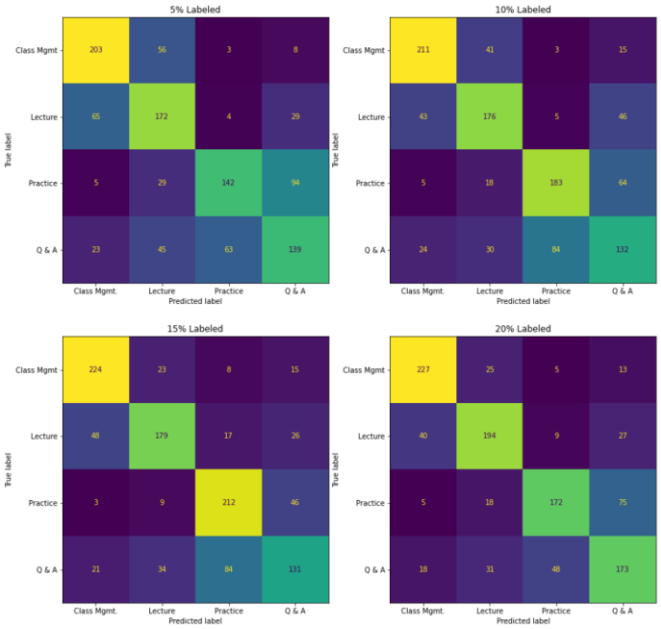


Figure 4. Confusion Matrix at 5%, 10%, 15% and 20% respectively

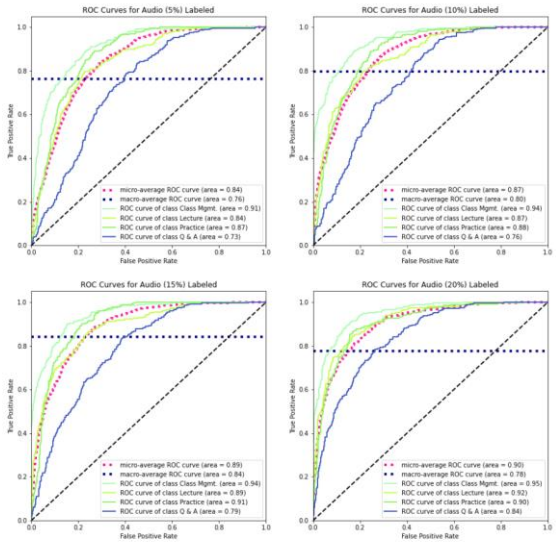


Figure 5. One-vs-Others ROC Curves at 5%, 10%, 15% and 20% labeled data respectively

7. CONCLUSION

The process of automating any aspect of the classroom observation task in resource-constrained environment remains a challenge. Unavailability of labelled data is a key detriment. This paper proposed and evaluated an audio version of the FixMatch algorithm to automate classroom observation using on audio classroom data only. Preliminary results are promising. If successful, this stream of work will allow for automating classroom observation in the resource-constrained regions in a more economical way. The primary limitations of this study are the relatively small size of the dataset and the requirement of manual pre-cleaning of the data. These can however be addressed in the future. Furthermore, to improve the performance of the system, other architectures like transformers can be leveraged.

REFERENCES

Bachman, P., Alsharif, O. and Precup, D. (2014) 'Learning with Pseudo-Ensembles', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2014/hash/66be31e4c40d676991f2405aaecc6934-Abstract.html (Accessed: 24 October 2022).

Bell, C.A. *et al.* (2019) 'Qualities of classroom observation systems', *School Effectiveness and School Improvement*, 30(1), pp. 3–29. Available at: https://doi.org/10.1080/09243453.2018.1539014.

Borich, G.D. (2016) *Observation Skills for Effective Teaching: Research-Based Practice*. 7th edn. New York: Routledge. Available at: https://doi.org/10.4324/9781315633206.

Bruns, B., De Gregorio, S. and Taut, S. (2016) *Measures of Effective Teaching in Developing Countries*. Research on Improving Systems of Education (RISE). Available at: https://doi.org/10.35489/BSG-RISE-WP_2016/009.

Caldwell, R. (2018) *The Problem with Teacher Observation*, *Chemonics International*. Available at: https://chemonics.com/blog/problem-teacher-observation/ (Accessed: 24 October 2022).

Chetty, R., Friedman, J.N. and Rockoff, J.E. (2014) 'Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood', *American Economic Review*, 104(9), pp. 2633–2679. Available at: https://doi.org/10.1257/aer.104.9.2633.

Grandvalet, Y. and Bengio, Y. (2004) 'Semi-supervised Learning by Entropy Minimization', in *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://proceedings.neurips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html (Accessed: 24 October 2022).

James, A. *et al.* (2019) 'Automated Classification of Classroom Climate by Audio Analysis', in L.F. D'Haro, R.E. Banchs, and H. Li (eds) *9th International Workshop on Spoken Dialogue System Technology*. Singapore: Springer (Lecture Notes in Electrical Engineering), pp. 41–49. Available at: https://doi.org/10.1007/978-981-13-9443-0_4.

Jamison, E.A., Jamison, D.T. and Hanushek, E.A. (2007) 'The effects of education quality on income growth and mortality decline', *Economics of Education Review*, 26(6), pp. 771–788. Available at: https://doi.org/10.1016/j.econedurev.2007.07.001.

Laine, S. and Aila, T. (2017) 'Temporal Ensembling for Semi-Supervised Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.1610.02242.

Lam, S. (2001) 'Educators' opinions on classroom observation as a practice of staff development and appraisal', *Teaching and Teacher Education*, 17(2), pp. 161–173. Available at: https://doi.org/10.1016/S0742-051X(00)00049-4.

Loshchilov, I. and Hutter, F. (2017) 'SGDR: Stochastic Gradient Descent with Warm Restarts'. arXiv. Available at: https://doi.org/10.48550/arXiv.1608.03983.

Martinez, F., Taut, S. and Schaaf, K. (2016) 'Classroom observation for evaluating and improving teaching: An international perspective', *Studies in Educational Evaluation*, 49, pp. 15–29. Available at: https://doi.org/10.1016/j.stueduc.2016.03.002.

Meignier, S. and Merlin, T. (2010) 'LIUM SPKDIARIZATION: An Open Source Toolkit for Diarization', in *CMU SPUD Workshop*. Dallas, United States. Available at: https://hal.archives-ouvertes.fr/hal-01433518 (Accessed: 24 October 2022).

Mundial, B. (2015) 'Conducting classroom observations: analyzing classrooms dynamics and instructional time, using the Stallings' classroom snapshot'observation system. User guide'.

Owens, M.T. *et al.* (2017) 'Classroom sound can be used to classify teaching practices in college science courses', *Proceedings of the National Academy of Sciences*.

Piczak, K.J. (2015) 'ESC: Dataset for Environmental Sound Classification', in *Proceedings of the 23rd ACM international conference on Multimedia*. New York, NY, USA: Association for Computing Machinery (MM '15), pp. 1015–1018. Available at: https://doi.org/10.1145/2733373.2806390.

Qiao, Q. and Beling, P.A. (2011) 'Classroom Video Assessment and Retrieval via Multiple Instance Learning', in G. Biswas et al. (eds) *Artificial Intelligence in Education*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 272–279. Available at: https://doi.org/10.1007/978-3-642-21869-9_36.

Richards, J.A. *et al.* (2008) *The LENATM automatic vocalization assessment*. Technical Report LTR-08-1). LENA Foundation.

Salamon, J. and Bello, J.P. (2017) 'Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification', *IEEE Signal Processing Letters*, 24(3), pp. 279–283. Available at: https://doi.org/10.1109/LSP.2017.2657381.

Salamon, J., Jacoby, C. and Bello, J.P. (2014) 'A Dataset and Taxonomy for Urban Sound Research', in *Proceedings of the 22nd ACM international conference on Multimedia*. New York, NY, USA: Association for Computing Machinery (MM '14), pp. 1041–1044. Available at: https://doi.org/10.1145/2647868.2655045.

Schlotterbeck, D. *et al.* (2021) 'What Classroom Audio Tells About Teaching: A Cost-effective Approach for Detection of Teaching Practices Using Spectral Audio Features', in *LAK21: 11th International Learning Analytics and Knowledge Conference*. New York, NY, USA: Association for Computing Machinery (LAK21), pp. 132–140. Available at: https://doi.org/10.1145/3448139.3448152.

Seidman, E. *et al.* (2018) 'Assessment of pedagogical practices and processes in low and middle income countries: Findings from secondary school classrooms in Uganda', *Teaching and Teacher Education*, 71, pp. 283–296. Available at: https://doi.org/10.1016/j.tate.2017.12.017.

Sohn, K. *et al.* (2020) 'FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 596–608. Available at: https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html (Accessed: 24 October 2022).

Summaira, J. *et al.* (2021) 'Recent Advances and Trends in Multimodal Deep Learning: A Review'. arXiv. Available at: https://doi.org/10.48550/arXiv.2105.11087.

Wang, Z. *et al.* (2014) 'Automatic classification of activities in classroom discourse', *Computers & Education*, 78, pp. 115–123. Available at: https://doi.org/10.1016/j.compedu.2014.05.010.

Wedgwood, R. (2007) 'Education and poverty reduction in Tanzania', *International Journal of Educational Development*, 27(4), pp. 383–396. Available at: https://doi.org/10.1016/j.ijedudev.2006.10.005.

World Bank Group (no date) *The Stallings classroom observation system*, *World Bank*. Available at: https://www.worldbank.org/en/programs/sief-trust-fund/brief/the-stallings-classroom-snapshot (Accessed: 24 October 2022).

Zualkernan, I. and Khan, M.S. (2020) 'Towards an Audio-based CNN for Classroom Observation on a Smartwatch', in *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G). 2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*, pp. 224–229. Available at: https://doi.org/10.1109/AI4G50087.2020.9311083.

Zualkernan, I.A., Lutfeali, S. and Karim, A. (2014) 'Using tablets and satellite-based internet to deliver numeracy education to marginalized children in a developing country', in *IEEE Global Humanitarian Technology Conference (GHTC 2014). IEEE Global Humanitarian Technology Conference (GHTC 2014)*, pp. 294–301. Available at: https://doi.org/10.1109/GHTC.2014.6970295.