Eye Gaze Feature Classification for Predicting Levels of Learning

Saurin Parikh

Florida Atlantic University, Boca Raton, Florida, USA, sparikh2014@fau.edu,
Institute of Technology Nirma University, Ahmedabad, Gujarat, India, saurin.parikh@nirmauni.ac.in

Hari Kalva

Florida Atlantic University, Boca Raton, Florida, USA, hari.kalva@fau.edu

Abstract

E-Learning courses reach online to millions worldwide. Amidst the geo-flexibility of registering students, the main challenges are instructor feedback and student retention. Ability to predict difficult content in real time enables eLearning systems to adapt content to students' needs dynamically. Recently, we examined eye responses as an indicator of levels of learning and introduced a non-parametric, non-probabilistic and statistical feature weighted linguistics classifier (FWLC) capable of predicting difficult words (terms) and concepts. FWLC achieved 85% accuracy for predicting levels of learning of big words using eye responses. In this paper, we analyze the performance of FWLC with five machine learning classifiers. FWLC has a higher true positive rate (TPR) and a lower ratio of FNR/FPR (the novel is a positive class). FWLC achieves a TPR gain of 43% over the best performing machine learning classifier. Prediction accuracy of FWLC for big words is lower by 6.6% than the best performing machine learning classifier. However, this accuracy tradeoff is worth the higher TPR of FWLC as the objective is to predict novel words (positive class) more accurately so that content can adapt to student's need.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI) → HCI theory, concepts and models; Applied computing → Education → E-learning**

KEYWORDS

Predicting levels of learning; eye and pupil response analysis; e-learning, machine learning classification.

1 Introduction

E-Learning has transformed the way we deliver education to students across the globe. E-learning classrooms have a diverse group of students from various demographics. Self-paced learning is a popular eLearning model, and its primary challenge is to address to students' concern in real time. Lack of immediate teacher feedback to a student's learning concern causes delay and interruptions in learning. E-Learning needs to be adaptive to the student's learning level. Content adaptation requires prediction of difficult content (words, visual elements or concepts causing learning difficulty). Cognition is an individual characteristic, dependent on readers' skill such as logical reasoning, quantitative analysis, and verbal skills. These skills vary because of demographics, culture, experience, education and biological factors such as cognition, working memory capacity, psychomotor skills, ocular deficiency, oculomotor dysfunctions, and reading disorders [1-5]. Ability to predict students' learning difficulty (difficult term/concept) in real time enables e-learning systems to dynamically adapt content, provide online supplementary material, and classify learners into various learning groups.

Recent developments have made it possible to use eye trackers to track eye movements to assess emotional and cognitive state during learning, scene perception, program debugging, and building dynamic online teams for projects. [4], [6-12]. Eye trackers show the exact visual element (text, a graphic on display) that is causing the response. Eyes receive visual inputs and send it to the visual cortex for creating a perception. To predict learning difficulty, we do a spatio-temporal analysis of eye responses and stimulus (visual element: term/concept) that is causing the response. We hypothesize that variations in eye responses to the same concept over time are indicative of learning level.

Prior research had analyzed various eye response signals such as fixations and saccades for learning assessment [6], [13-17]. Fixation is a virtual point in the visual field, where we fixate to read or look at a scene and saccade is the movement of eyes from current fixation to the next one in the visual field. These prior research studies analyzed eye response signals to generalize the emotional, cognitive or affective state of learners. However, eye response to a stimulus is an individual characteristic, driven by individuals' cognition and perception. There is no known work, where eye responses of individuals were extracted as features and used by machine learning classifiers for predicting levels of learning

In our previous work, we applied the theories of psycholinguistics, information context processing, and human visual system to develop 12 eye movement features [1]. All features may not contribute equally to indicate learning level because eye response to reading is an individual characteristic.

Therefore, we had proposed a feature weighted linguistics classifier (FWLC) that uses a feature selection method (either ReliefF, Information Gain or Ensemble) to assign weight to each eye response feature based on its relevance [1]. Three FWLC variants: FWLC-ReliefF, FWLC-InfoGain, and FWLC-Ensemble used their corresponding set of weighted features for predicting learning level for a word/concept. FWLC-ReliefF achieved 85% prediction accuracy for big words, which was 24% greater than baseline, a majority voting classifier, which assigns equal weight to features for making a prediction decision [1]. In this paper, we analyze the prediction performance of three FWLC variants with five popular classifiers Naïve Bayes, SVM, KNN, decision trees (C4.5), and random forest. FWLC does show 6.6% lower accuracy than the best performing machine learning classifier however the tradeoff is worth the high true positive rate and low false negative rate shown by all three FWLC variants. It means that FWLC predicts novel words (novel is a positive class) more accurately, which is more desirable for adapting content to a students' level. We have organized the paper into three sections. Section two briefly provides the list of eye response features used for this study and the overview of our FWLC classifier. A detailed description of FWLC can be found in [1]. We describe the experimental setup in section three, discuss results in section four and conclusion in section five.

2 Eye Movement Features and Overview of FWLC Classifier

2.1 Eye Response Features

Eye response to visual content is driven by cognition and perception of individuals'. Therefore, eye response measures derived from a large population are not suitable for predicting learning difficulty for individuals [1]. Eye response to reading depends on individual characteristics. For example, dyslexic or non-native readers read a word at a syllable level, and this reading behavior causes more fixations on a word and shorter saccades. Whereas, skilled readers learn a word at morpheme/word level and as a result, have fewer fixations and longer saccades [3]. Dyslexia readers show more top-down processing delay of visual signals and are more prone to delayed sentence comprehension (slow learners) [2].

Hence reading behavior will directly influence the fixations and saccades on a word. Fixation is a point in the visual field, where the reader fixates for a short duration to read the word/visual content. The word read during a fixation triggers bottom-up processing of the visual signal. The signal received by fovea/parafovea is sent to visual cortex for creating the perception. On receiving visual signals, the upcoming input is anticipated even before receiving the next bottom-up visual information. The result of anticipation controls the saccadic movement of eyes to the next point (saccade) in the visual field (This process is often called as the top-down processing of visual signal) [18], [1-2]. Readers make longer saccades when anticipation matches the upcoming visual information and gaze moves to the next word following the parafoveal word otherwise

gaze moves to the parafoveal word or remaining part of the foveal word (shorter saccade). Hence, the number of fixations, fixation duration, and saccade length on a term/concept is reader specific. Familiar term/concept attracts no fixations or fewer fixations of shorter duration in comparison to novel terms [6] [19]. Word familiarity also varies individually. More fixations of longer duration on a term/concept may represent learning difficulty [20], [4]. Even pupil response varies as per content familiarity and complexity. On reading a novel term/concept, perceiving a complex scene, or processing complex tasks, the pupil may dilate due to higher cognitive load [15], [21-22], [11]. However, a term/concept which a reader perceives as novel during first pass (first time reading) of the term may not always remain novel after reading its information context; as readers may derive its meaning by looking at the context [23], [1]. The reading of information context causes regressions (relook or rereading of the content). Regression demonstrates various reading patterns. In our previous work, we had derived five reading patterns, and three eye movement features related to reanalysis (regressions) [1]. Hence, we modeled fixation, fixation duration, saccade and regression as eye response features.

By theories of psycholinguistics, contextual information processing, anticipatory behavior analysis, recurrence fixation analysis and pupillary response, we had derived 12 eye response features that indicate learning level [1]. We divided the 12 features into two groups: (i) six first pass features and (ii) six reanalysis features. We computed first pass features from eye responses collected during first pass reading of a word and the reanalysis features from eye responses recorded during rereading or revisiting a word during subsequent passes. We list the first pass eye response features in Table 1 and reanalysis features in Table 2. Nine features are quantitative, and three are categorical with two categories: observed/not observed. The eye response to a term is captured in a term-response map using its' associated 12 eye response features.

Table 1: First pass eye response features [1]

Feature Name	Feature Type	Description
Single fixation duration (SFD)	Quantitative	It represents the total amount of time spent during a single fixation on a target word.
First fixation duration (FFD)	Quantitative	Amount of time spent by a reader on the first fixation; In this case, total no of fixations on a word is higher than one.
Gaze duration (GD)	Quantitative	Sum of the duration of all consecutive fixations on a target word from the first fixation until the first time that reader moves the gaze away from the word.
Mean fixation duration (AFD)	Quantitative	Mean of the total duration of all first pass fixations on a target word.

| Fixation count (FC) | Quantitative | Total no of fixations on a word during first pass reading of the word. |
| Mean pupil diameter of left eye/right eye (FPPupilDiaLeft/ FPPupilDiaRight) | Quantitative | Mean of all pupil diameters of the left/right eye, which is measured during the entire duration of all fixations on a specific term recorded during first pass reading of a term. |

Source: Parikh et al. [1]

Table 2: Reanalysis eye response features [1]

Feature Name	Feature Type	Description
Regressions (RC)	Quantitative	Measured as a number of look backs to the target word after the reader's initial encounter with the word or context has ended.
Second pass time (SPT)	Quantitative	Amount of total fixation time spent on the target word after leaving the word and later revisiting it second time before navigating to next term/concept
Look back Fine Detail (LFD)	Categorical. Two categories: Observed / Not Observed.	The reader makes brief fixation during the first pass and revisits term in more detail (more fixations) during the second pass.
Look back re-glance (LRG):	Categorical. Two categories: Observed / Not Observed.	Reader visits term/context in more detail during the first pass and takes a re-glance at it during the second pass.
Determinism (Dm)	Categorical. Two categories: Observed / Not Observed.	Eye movements show recurrence of fixation sequences both during first and second pass reading for establishing an association between lexical co-occurrence of novel words.
Mean pupil diameter of left eye/right eye (RAPupilDiaL eft/RAPupilDi aRight)	Quantitative	Mean of all pupil diameters of the left/right eye, which is measured during the entire duration of all fixations on a specific term recorded during reanalysis.

Source: Parikh et al. [1]

2.2 FWLC Classifier

Figure 1 depicts the conceptual model of the levels of learning prediction system. A detailed description of this system is at [1]. The levels of learning prediction system have two phases: training phase (reading surveys) and prediction phase. During the training phase, subjects read slides with easy passages containing terms of high frequency and common words (two term classes). We had derived the word frequencies from [24]. Tobii pro x2-60 eye tracker was used to capture eye responses at a rate of 60 samples per second. We extracted words from slides using our "slide words extraction engine" and mapped eye responses on the words using the word-to-eye-response-mapping engine. The words with its eye responses (term-response map) are given to the training engine. During the training phase (please refer left section of Figure 2), we compute 12 eye response features for each term (term-response map). We group term-response maps by term class and compute the feature thresholds across the related term-response maps of the same class. We call this map as term-class response threshold map in the paper. We input term-response maps to three feature selection methods: ReliefF, Information Gain, and Ensemble. These methods assign feature weight to features on the basis of its relevance. We call these three sets of feature weights as ReliefF-FW, InfoGain-FW, and Ensemble-FW and the corresponding FWLC variants as FWLC-ReliefF (FRF), FWLC-InfoGain (FIG) and FWLC-Ensemble (FEN). Training phase gives three set of feature thresholds, one set for each term length class (big/mid/small) and three set of feature weights: ReliefF-FW, InfoGain-FW, and Ensemble-FW for each subject.
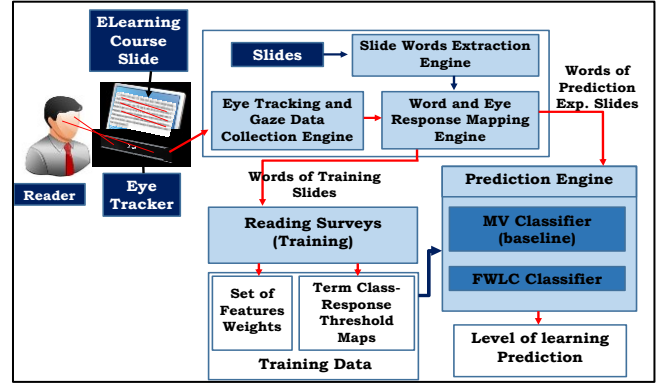
Figure 1. Conceptual model of our levels of learning prediction system

During prediction phase (please refer right section of Figure 2), subjects read separate passage slides containing words of three categories: low-frequency, high-frequency and common. Term-response map is created for each word. We clean the maps having no fixation count or pupil responses, The FWLC classifier with linguistics knowledge uses the set of feature thresholds and any one set of feature weights (training data) to classify a term/concept of prediction phase as novel or familiar. FWLC assigns different weights to features using one of the Feature selection methods. We also input the same term-response maps of prediction phase to majority voting classifier (baseline). The baseline classifier uses the threshold maps from training phase and assigns equal weights to all features. Detailed classification process of both baseline and FWLC classifiers can be found at [1].
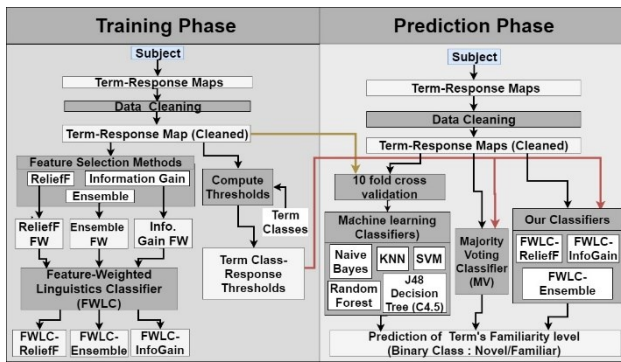
Figure 2. Block diagram depicting levels of learning prediction process (training and prediction phases)

3 Experiment

3.1 Subjects and Dataset

Ten subjects (undergraduate/graduate students) participated in our experiment. Their age is 21 to 32. We removed data for two human subjects' due to data collection error. During the training phase, subjects read six passage slides and five slides during prediction phase. Subjects answered questions related to word comprehension and pronunciation difficulty, and we used this subjective assessment data as the ground truth for measuring the performance of the classifier. During the prediction experiment, subjects looked at a total of 56 big size words (word of length > eight letters). 28 each were low frequency and high-frequency big words. Subjects fixated on an average of 60% words. Subject's dataset is the term-response maps for big words.

3.2 Selection of Machine Learning Classifiers

FWLC, a non-parametric classifier does not use a fixed set of features for classification. Relevant features are selected using feature selection methods on training data. Hence, subject wise, the set of relevant features used may increase or decrease depending upon its relevance for each subject because reading is an individual characteristic and all features may not be relevant for all subjects. As observed in [1], FWLC-InfoGain considers only five out of 12 features as relevant for making a prediction, and it works best for a specific group of readers whereas ReliefF-FWLC used 12 relevant features for prediction [1]. We compared the classification performance of the nonparametric FWLC with a popular simple non-parametric classifier, KNN. It uses K nearest neighbors for classification typically with majority voting method for assigning the most common class from its K nearest neighbors. Moreover, in this paper, we have used three nearest neighbors (3NN) and have assigned a weight of 1/d (d is the distance from the object) to the neighbors so that the nearest neighbors will have more contribution in the classification decision. FWLC is not a probabilistic classifier as it does not assign the probability to classes but outputs the most likely class from the observation. Hence, we have compared the performance of FWLC variants

with a non-probabilistic classifier, SVM, and also with a probabilistic classifier, naïve Bayes.

Statistical classifiers such as decision trees (C4.5) are good performers when a labeled training set of observations is available for learning. As our training set has labeled observations, we have selected C4.5 (j48) for comparing the prediction results with FWLC. However, C4.5 is known for overfitting its decision to the training set because prediction from a single tree is sensitive to noise in the training set. Random forest, an ensemble learning classifier corrects this characteristic of C4.5 and can overcome the low bias and high variance (overfitting) scenario of decision trees by using bagging procedure. It trains many trees with randomly selected parts of the training sets, and the average prediction of many trees will be less sensitive to noise as trees are de-correlated by showing different training sets. Hence, we have also selected random forest classifier with bagging and 100 iterations.

The focus of this paper is to compare the classification results of FWLC variants with five machine learning classifiers: Naïve Bayes (NB), KNN, SVM, C4.5 and Random forest (RF) for predicting levels of learning using eye responses.

3.3 Cost-Sensitive Classification

Weka 3.7.11 tool [25] was used to evaluate the performance of machine learning classifiers. The test mode used was 10-fold cross-validation. We divided our labeled dataset of each subject (term-response maps) into ten subsets. Out of 10 subsets, we randomly selected a testing subset during ten iterations (folds). Each classifier predicts a level of learning for a term and classifies it into two levels of learning: a novel (positive class) or familiar (negative class). The overall purpose of the experiment is to detect learning difficulty during a learning exercise and adapt content. Hence, we consider predicting novel terms as familiar (false negative) more expensive than predicting familiar terms as novel (false positive) and therefore, we applied a double classification cost to false negatives than false positives. We used the knowledge flow tool of Weka to run the selected cost-sensitive classifiers and the cost-sensitive classification process is described below.

- The eye response dataset (term-response maps) is converted to 'arff' format using Weka.
- Term-response map ('arff' record) has nine quantitative and three categorical features and one class attribute. We have listed six features each in Table 1 and Table 2.
- Subject-word-assessment' (subject's self-learning assessment for the word), a categorical response variable was selected as a class attribute (class assigner).
- We selected the novel class as a positive class (class value picker).
- We added 10-fold cross validation evaluator method to the flow.
- Five classifiers: SVM, 3NN, J48, Random Forest and, Naïve Bayes were selected with their cost matrix. False negatives (FN) were assigned double cost than the false positives (FP).
- We ran the five cost-sensitive classifiers and noted the prediction accuracy for predicting levels of learning for big

words and the corresponding TPR, FNR, TNR, and FPR for each subject separately and finally computed the mean prediction accuracy, mean TPR, FNR, TNR, and FPR across all subjects.

4 Results

The primary objective of our study is to predict levels of learning in real time and adapt content as per the learning level. Hence we consider predicting novel word as familiar more expensive then predicting familiar word as novel (false negatives are more expensive than false positive). Table 3 shows subject wise True-positive rate (TPR) for all classifiers. TPR is the proportion of total novel words, which are correctly classified as novel. All FWLC variants are performing better for predicting novel words than the five machine learning classifiers. FWLC-InfoGain shows the highest mean TPR of 0.91 followed by FWLC-ReliefF and Ensemble with 0.76. Random forest performed the worst for predicting novel words. For four subjects (S3–S5 and S7) the three FWLC variants show similar best performance.

Table 3 shows the subject-wise best-performing classifiers in bold-italics and the highest performer from machine learning group in bold. FWLC-InfoGain, the best performing classifier achieves a mean TPR gain of 43% in comparison to SVM, the highest scoring machine learning classifier from the machine learning group.

Table 3: True Positive Rate (TPR) [Actual Novel word predicted as Novel]

Subject ID	Baseline	Proposed FWLC Classifier Variants			Machine Learning Classifiers				
	MV	FRF	FIG	FEN	NB	KNN	RF	SVM	J48
S1	0.33	0.50	*0.67*	0.50	0.29	0.14	0.14	0.14	0.29
S2	0.25	0.50	*0.75*	0.50	0.25	0.50	0.00	0.50	0.50
S3	0.67	*1.00*	*1.00*	*1.00*	0.00	0.00	0.00	0.33	0.33
S4	0.60	*1.00*	*1.00*	*1.00*	0.83	0.67	0.33	0.83	0.50
S5	0.00	*1.00*	*1.00*	*1.00*	0.33	0.07	0.00	0.33	0.00
S6	0.83	0.67	*1.00*	0.67	0.50	0.17	0.17	0.33	0.33
S7	0.00	*1.00*	*1.00*	*1.00*	0.17	0.00	0.00	0.67	0.67
S8	0.63	0.38	*0.88*	0.38	0.38	0.69	0.50	0.69	0.69
Mean TPR	0.41	0.76	*0.91*	0.76	0.34	0.28	0.14	**0.48**	0.41

Table 4 shows the subject wise and mean prediction accuracy of all classifiers. FWLC-Relief performs slightly better than other FWLC variants. However, SVM shows the best overall prediction accuracy of 91.29%, an improvement of 6.6% over the best performing FWLC variant (FWLC-ReliefF). Subject wise performance of classifiers shows that J48 performs better for subjects S1 and S2. FWLC variants perform better for S3. SVM works better for S4, S6 and, S7. KNN works better for S5 and S8, and even J48 shows similar performance for S8. No one classifier shows the best performance for the majority of the subjects. Naïve Bayes performs worst among the machine learning group. The

slightly higher prediction accuracy of machine learning classifiers in comparison to FWLC comes with a cost because they increase the number of false negatives. Machine learning classifiers predict more novel terms as familiar (false negatives), and hence this conditions is not desirable for adapting eLearning content as providing supplementary learning content may be skipped due to incorrect prediction of novel content as familiar (prediction of false negatives).

Table 4: Prediction Accuracy

Sub. ID	Baseline	Proposed FWLC Classifier variants			Machine Learning Classifiers				
	MV	FRF	FIG	FEN	NB	KNN	RF	SVM	J48
S1	57.89	86.84	86.84	86.84	84.09	85.23	88.64	88.63	*92.04*
S2	66.67	83.33	88.89	83.33	80.00	93.33	*91.66*	90.00	*91.66*
S3	76.92	*92.31*	*92.31*	*92.31*	88.57	82.85	91.42	85.71	88.57
S4	44.00	84.00	80.00	80.00	88.31	92.20	92.20	*96.10*	88.31
S5	55.17	79.31	72.41	75.86	94.73	*97.36*	96.05	97.36	85.52
S6	63.89	86.11	88.89	86.11	89.13	89.13	91.30	*93.47*	86.95
S7	72.41	79.31	75.86	79.31	86.07	86.07	89.87	*92.40*	88.60
S8	75.56	86.67	84.44	86.67	89.62	*89.62*	88.88	86.66	*89.62*
Mean Pred. Acc.	64.06	*84.74*	83.71	83.80	87.57	89.47	91.25	**91.29**	88.91

Table 5 shows the subject wise and mean false negative rate of all classifiers. All FWLC variants show lowest mean FNR in comparison to all the machine learning classifiers. FWLC-InfoGain with lowest mean FNR performs the best and random forest is the worst performer. The graph of Figure 3 shows that all FWLC variants are good at predicting novel terms (minority class) more accurately (Achieving highest TPR and lowest FNR). Figure 4 shows the ratio of FNR and FPR plot where FWLC variants achieve lowest FNR at negligible cost of FPR and perform best in comparison to all machine learning classifiers. SVM achieves highest prediction accuracy in comparison to FWLC variants and other four machine learning classifiers. However, it performs worst for predicting novel terms.

Table 5: False Negative Rate (FNR) [Actual Novel word predicted as Familiar]

| Sub. ID	Baseline	Proposed FWLC Classifier variants			Machine Learning Classifiers				
	MV	FRF	FIG	FEN	NB	KNN	RF	SVM	J48
S1	0.67	0.50	0.33	0.50	0.71	0.86	0.86	0.86	0.71
S2	0.75	0.50	0.25	0.50	0.75	0.50	1.00	0.50	0.50
S3	0.33	0.00	0.00	0.00	1.00	1.00	1.00	0.67	0.67
S4	0.40	0.00	0.00	0.00	0.17	0.33	0.67	0.17	0.50
S5	1.00	0.00	0.00	0.00	0.67	0.33	1.00	0.67	1.00
S6	0.17	0.33	0.00	0.33	0.50	0.83	0.83	0.67	0.67
S7	0.00	0.00	0.00	0.00	0.83	1.00	1.00	0.33	0.33
S8	0.38	0.62	0.13	0.63	0.63	0.31	0.50	0.31	0.31
Mean FNR	0.46	0.24	*0.09*	0.25	0.66	0.65	**0.86**	0.52	0.59
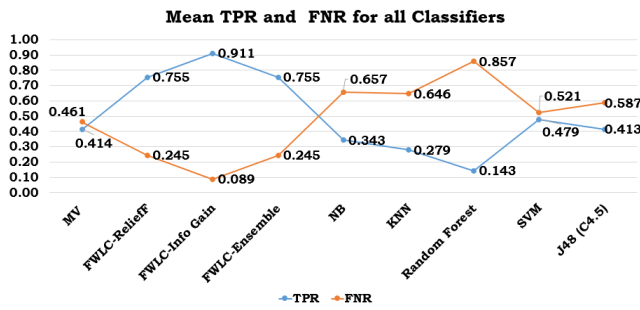
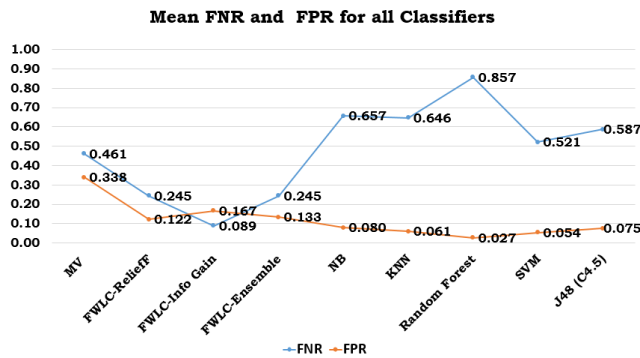**Figure 3. Mean TPR / Mean FNR plot for all classifiers**

Figure 4. Mean FNR / Mean FPR plot for all classifiers

5 Conclusion

FWLC variants perform better for predicting novel terms at a low cost of predicting false positives. Learning level prediction in real time will help eLearning systems to adapt the content as per student's learning level and hence predicting novel terms correctly in this scenario is highly desirable. This model may work best for assessing learning level of students from diverse demographics and adapt learning content for an individual on a need basis using eye responses. FWLC performance looks promising for making content adaptation decisions in real time.

References

[1] S. Parikh and H. Kalva, "Feature Weighted Linguistics Classifier for Predicting Learning Difficulty Using Eye tracking," ACM Trans on. Applied Perception [under review].

[2] F. Huettig and S. Brouwer, "Delayed Anticipatory Spoken Language Processing in Adults with Dyslexia—Evidence from Eye-tracking," Dyslexia, vol. 21, no. 2, pp. 97–122, May 2015.

[3] A. Vagge, M. Cavanna, C. E. Traverso, and M. Iester, "Evaluation of ocular movements in patients with dyslexia," Ann. Dyslexia, vol. 65, no. 1, pp. 24–32, Apr. 2015.

[4] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," Q. J. Exp. Psychol., vol. 62, no. 8, pp. 1457–1506, Aug. 2009.

[5] K. Rayner and G. W. McConkie, "What guides a reader's eye movements?," Vision Res., vol. 16, no. 8, pp. 829–837, 1976.

[6] R Williams and R. Morris, "Eye movements, word familiarity, and vocabulary acquisition," Eur. J. Cogn. Psychol., vol. 16, no. 1–2, pp. 312–339, Jan. 2004.

[7] J. L. Sibert, M. Gokturk, and R. A. Lavine, "The Reading Assistant: Eye Gaze Triggered Auditory Prompting for Reading Remediation," in Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, 2000, pp. 101–107.

[8] F. C. Pop, M. F. Vaida, and M. Cremene, "An alternative strategy for grouping students in eLearning using an Enneagram methodology and eye tracking," in 2010 9th International Symposium on Electronics and Telecommunications, 2010, pp. 395–398.

[9] M. Malčík, E. Mechlová, Z. Sikorová, and A. Mentel, "Analysis of visual perceptual patterns on screen using eye tracker technology," in 2014 IEEE 12th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2014, pp. 311–317.

[10] S. Husain, S. Vasishth, and N. Srinivasan, Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. Journal of Eye movement Research, 2015.

[11] Z. Dobesova and M. Malcik, "Workflow diagrams and pupil dilatation in eye-tracking testing," in 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2015, pp. 1–6.

[12] Y. T. Lin, C. C. Wu, T. Y. Hou, Y. C. Lin, F. Y. Yang, and C. H. Chang, "Tracking Students #x2019; Cognitive Processes During Program Debugging #x2014;An Eye-Movement Approach," IEEE Trans. Educ., vol. 59, no. 3, pp. 175–186, Aug. 2016.

[13] H. Wang, M. Chignell, and M. Ishizuka, "Empathic Tutoring Software Agents Using Real-time Eye Tracking," in Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, New York, NY, USA, 2006, pp. 73–78.

[14] M. Porta, S. Ricotti, and C. J. Perez, "Emotional e-learning through eye tracking," in Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON), 2012, pp. 1–6.

[15] Saurin Parikh and Hari Kalva. 2018. Predicting Learning Difficulty Based On Gaze And Pupil Response. In Proceedings of 26th Conference on User Modeling, Adaptation and Personalization Adjunct, Singapore, July 2018 (UMAP'18 Adjunct), 5 pages. DOI: 10.1145/3213586.3226224.

[16] V. Manuel et al., "AdeLE: A Framework for Adaptive E-Learning through Eye Tracking," in Proceedings of IKNOW 2004, 2004, pp. 609–616.

[17] V. Cantoni, C. J. Perez, M. Porta, and S. Ricotti, "Exploiting Eye Tracking in Advanced e-Learning Systems," in Proceedings of the 13th International Conference on Computer Systems and Technologies, New York, NY, USA, 2012, pp. 376–383.

[18] M. J. Traxler, "Trends in syntactic parsing: anticipation, Bayesian estimation, and good-enough parsing," Trends Cogn. Sci., vol. 18, no. 11, pp. 605–611, Nov. 2014.

[19] L. Copeland and T. Gedeon, "The Effect of Subject Familiarity on Comprehension and Eye Movements During Reading," in Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, New York, NY, USA, 2013, pp. 285–288.

[20] K. Rayner and S. A. Duffy, "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity," Mem. Cognit., vol. 14, no. 3, pp. 191–201, May 1986.

[21] C. Sibley, J. Coyne, and C. Baldwin, "Pupil Dilation as an Index of Learning," Proc. Hum. Factors Ergon. Soc. Annu. Meet. vol. 55, no. 1, pp. 237–241, Sep. 2011.

[22]] G. J. Siegle, S. R. Steinhauer, V. A. Stenger, R. Konecky, and C. S. Carter, "Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data," NeuroImage, vol. 20, no. 1, pp. 114–124, Sep. 2003.

[23] R. Chaffin, R. K. Morris, and R. E. Seely, "Learning new word meanings from context: a study of eye movements," J. Exp. Psychol. Learn. Mem. Cogn., vol. 27, no. 1, pp. 225–235, Jan. 2001.

[24] Davies and Mark, "The Corpus of Contemporary American English (COCA): 520 million words, 1990-present." 2008.

[25] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, Fourth Edition, 2016.