IHJPAS. 36 (3) 2023

Ibn Al-Haitham Journal for Pure and Applied Sciences
Journal homepage: jih.uobaghdad.edu.iq

Applying Ensemble Classifier, K-Nearest Neighbor and Decision Tree for Predicting Oral Reading Rate Levels

Jwan Abdulkhaliq Mohammed
Department of Computer Science, College of Science
University of Duhok, Iraq.

Article history: Received 8 November 2022, Accepted 27 December 2022, Published in July 2023.

doi.org/10.30526/36.3.3102

Abstract

For many years, reading rate as word correct Per Minute (WCPM) has been investigated by many researchers as an indicator of learners' level of oral reading speed, accuracy, and comprehension. The aim of the study is to predict the levels of WCPM using three machine learning algorithms: which are Ensemble Classifier (EC), Decision Tree (DT), and K- Nearest Neighbor (KNN). The data for this study were collected from 100 Kurdish EFL students in the 2nd-year English language department, at the University of Duhok in 2021. The outcomes showed that the ensemble classifier (EC) obtained the highest accuracy of testing results, with a value of 94%. Also, EC recorded the highest precision, recall, and F1 scores with values of 0.92 for the three performance measures. The Receiver Operating Character curve (ROC curve) also got the highest results compared to other classification algorithms. Accordingly, it can be concluded that the ensemble classifier is the best and most accurate model for predicting the reading rate (accuracy) of WCPM.

Keywords: Word Correct Per Minute, Prediction, Machine learning, Reading rate, Accuracy

1. Introduction

Assessing EFL and ESL learners' Oral Reading Fluency (ORF) is a necessity that researchers should attempt to investigate more as the number of such studies is still insufficient [1,2]. According to Grabe [1], ORF is "the ability to read rapidly with ease and accuracy, and to read with appropriate expression and phrasing" (p.72). Based on the given definition, it is evident that rate and accuracy are significant dimensions of ORF assessment. Reading rate (WCPM) is the number of words a reader maintains to read correctly in a minute [3].

Researchers have used many scales and models to evaluate learners' ORF. Two scales of ORF are quite well-known worldwide as they include measuring rate and accuracy: the National Assessment of Educational Progress (NAEP) and the Comprehensive Oral Reading Fluency Scale (CORFS) [4].

Reading accuracy and especially reading rate (speed) as word correct per minute (WCPM) has been and still is the main focus for many researchers in assessing learners' ORF as it is highly important for teachers, educators, and educational institutions. The significance of assessing readers' (WCPM) can be summarized in the following: (1) it is a fast and reliable indicator that shows readers' level of reading speed and accuracy; (2) it can be used in the process of evaluating and developing reading curriculum (textbooks); (3) it enables researchers to easily do and assess the effectiveness of their interventions while doing experimental studies; and (4) it saves readers, educators and researchers' time and effort in detecting the ineffective reading material [5].

Accordingly, as an outcome, WCMP has been a significant indicator of learners' reading rate (speed), reading accuracy, and comprehension. Therefore, for several decades, researchers have concentrated on evaluating readers' speed and accuracy levels using WCPM [4].

Reading rate WCPM depends mainly on the number of errors readers make, the number of correct words they maintain to read, and the total time (total reading time) they take to read the selected passage [3]. In general, many factors may affect EFL and ESL readers' WCPM scores, such as: (1) the influence of their mother tongue; (2) their phonological, syntactical, and semantical awareness of the target language; (3) their memory span quality in the foreign language; (4) their knowledge and experience of making correct associations of sounds; and (5) the difference between their mother tongue and foreign language script direction, such as from left to right or right to left [6]. Other factors were also pointed out by Rahmawati et al. [7] regarding the quality of the selected text, such as the volume of the text, the level of difficulty of the sentence structure of the text, and the quality and number of difficult terms in the text.

Previously, many researchers have used software programs focusing on using WCPM as a component of ORF. However, a few studies have been performed on predicting reading rates using machine learning algorithms. Reading rate WCPM reflects students' reading accuracy and speed levels. For this reason, it is very important to assess and predict the WCPM to measure the levels of reading rate or accuracy for students using machine learning algorithms.

The algorithms of machine learning consider the best solutions for many issues that are hard to fix. Classification is a data extraction approach (machine learning) that is utilized to predict and classify a wide range of domains [8]. Classification is utilized to categorize data into multiple classes based on specified constraints. The classification is an example of supervised learning, since it uses training data linked with class labels as input. Classification algorithms are used for several applications, such as medical disease diagnosis, social network analysis, artificial intelligence, document categorization, etc. There are several types of classification techniques, like K-Nearest Neighbor classifier, Ensemble Classifier, Decision Trees, Support Vector Machine etc. [9]. Data classification is the process of classifying data into groups so that data items within the same group are more similar and data items from various groups are not similar. The classification procedure is separated into two stages. The first stage is the training phase, during which the classification model is constructed. The second stage is classification, in which the trained model is utilized to allocate an unknown data item to one of a set of class labels [10].

Machine learning algorithms have been used in several studies to assess ORF using WCPM as a component. For example, Bolaños et al. [11] have used word accuracy, reading rate (WCPM), and expressiveness as components of oral reading fluency in order to measure the reading ability of

children using speech recognition and machine learning algorithms. Kim et al. [12] used acoustic features to estimate oral reading fluency measured by WCMP using machine learning algorithms.

Therefore, the aim of the study is to show and examine findings obtained using EC, KNN, and DT algorithms to predict the WCPM reading rate. In other words, the goals of this research are to predict the levels of reading rate WCPM using the three machine learning methods and to identify which algorithm produces the best classification results in terms of the confusion matrix and ROC curve.

2. Materials and Methods

2.1 Data Collection

The source of data was the mid-term oral reading exam of 100 Kurdish EFL 2nd-year university students of the English language department, University of Duhok, in 2021. Students were individually tested as they were given a reading passage of 514 words from the book Cover-to-Cover2 by Day and Harsch [13]. The book is designed to include intensive and extensive reading passages that are joyful, interesting, and relatively brief. The selected passage consisted of five paragraphs, including a variety of simple, compound, complex, and compound-complex sentence structures.

2.2 Human Scoring:

Printed copies of the selected reading passage were prepared for each participant in advance, and participants were tested individually by the subject teacher, asking them to read the whole passage out loud. The subject teacher timed the oral reading of each participant at a time, and their errors were highlighted on the printed text sheets. The work of Rasinski [3] illustrates the procedure for calculating reading accuracy and reading rate (speed) as WCPM. To measure the WCPM, several calculations such as the number of errors, the number of words read correctly, and the total reading time (converting the readers' time of reading the passage from minutes into seconds and then into decimals) are also required to be done to assess the accuracy percentage and the WCPM of a particular reader. The table below shows the procedure for calculating each of the above-mentioned aspects:

Table1. The Process of Calculating WCPM

#	Aspects	Formulas
1	Words read correctly	$\text{Total number of words} - \text{The number of errors made}$
2	Total reading time	$\frac{The\ number\ of\ seconds}{60}$
3	Reading rate WCPM	$\frac{\text{The number of words read correctly}}{\text{The total reading time}}$

Accordingly, their calculated scores of WCPM were classified according to the four classifications of the CORFS model [4], because it is the only model that classifies readers' WCPM into four different levels with a specific number of words for each category. The table below shows the CORFS scale.

Table 2. The CORFS classification of WCPM

CORFS classifications of WCPM	Levels
137+	Skilled reader
107 - 136	Moderate reader
78 - 106	Non-fluent reader2
1 - 78	Non-fluent reader1

2.3 Machine Learning Algorithms:

2.3.1 K-Nearest Neighbor Classifier

The K-Nearest Neighbor (KNN) is considered one of the simplest algorithms in supervised machine learning. It is based on the idea that similar samples are often found in close proximity. The KNN algorithm is an instance-based machine learning model. Instance-based learning is also known as a "lazy learner" algorithm because it keeps the whole training sample and does not create a classifier until a newer and unlabeled sample wants to be categorized. During the training stage, the algorithms of lazy learners request less computing time than the algorithms of eager learners, like decision trees, Bayes networks, and neural networks, but extra calculation time is required for the classification stage [14].

KNN classifiers rely on similarity learning. Similarity learning involves comparing a given test sample to the available training samples that are similar to it. To classify a data sample X, its K-nearest neighbors are found, and then X is allocated to the class label to which the majority of its neighbors belong. The performance of the KNN classifier is also affected by the value of k. A very small value of k may lead to the KNN classifier being susceptible to over fitting due to noise in the training dataset.

However, a very high value of k may lead the KNN classifier to misclassify the test sample because its list of nearest neighbors includes data that are set far away from its neighbors. K-NN is based on the assumption that data is related in a feature space. As a result, all points are examined in order to calculate the distance between the data points. The similarity between the datapoints is calculated using a distance metric, such as Euclidean distance. Hamming distance or Euclidian distance is utilized based on the data type of the data classes used [15].

The below equation presents the general form of the Euclidean distance. The Euclidean distance is a distance measurement between two points. The similarity between d1 and d2 is between 0 and 1; smaller values indicate lower similarity, while larger numbers indicate greater similarity.

$$E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

Where n is the number of attributes and x_i and y_i are the k-th attributes.

As seen in **Figure 1**, the following example illustrates three classes: X, Y, and Z. It is now necessary to determine the class label for data sample P. The Euclidean distance and the value of K=5 are computed for each sample pair, and it is discovered that the four closest neighbor samples belong to class label X, whereas a single tuple belongs to class label Z. As a result, sample P is allocated to class X because it is the primary class for that sample [16].

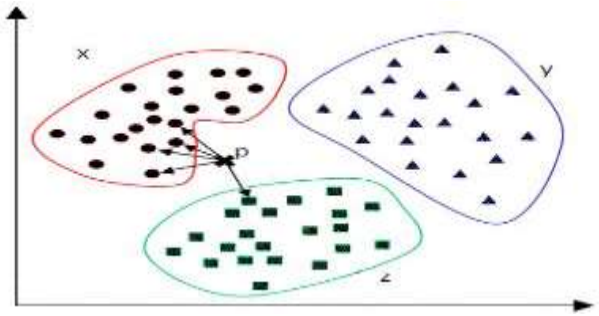


Figure1. An example of KNN classifier

For using KNN for reading rate classification, it considers the feature vectors (number of errors, the number of words read correctly, and total reading time), then computes the similarity of each neighbor to the test data using Euclidean distance. The Euclidean distance calculates the distance between the test data and all WCPM training data (i.e., each row of training data). After calculating the distance between them, the distance is sorted and the nearest neighbors are determined to test the data based on the Kth minimum distance.

2.3.2 Decision Tree

A decision tree is used as a predictive model in decision tree learning, which translates observations about an object into inferences about the item's goal value. The algorithm of a decision tree iteratively splits a dataset of records utilizing either a breadth-first strategy or a depth-first greedy approach until all data items belong to a certain class. The structure of a decision tree consists of the root, internal, and leaf nodes. It is a tree structure that looks like a flow chart, with each internal node representing a test condition on an attribute, each branch representing the outcome of the test condition, and each terminal node (or leaf node) specified as a class label. The root node is the upper node. A decision tree is built using a divide-and-conquer strategy [17]. Each branch of the decision tree represents a decision rule. In general, it takes a greedy approach from the top down; see the below **Figure**.

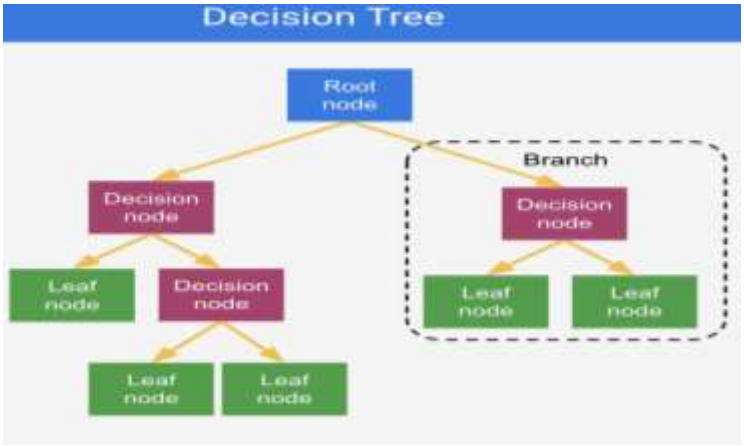


Figure 2. The structure of Decision tree algorithm

The approach to decision trees is carried out in two stages: tree construction and tree pruning [18].

Tree construction is done from the top to the down strategy. The tree is iteratively partitioned at this phase until all data items belong to the same class label. Because the training dataset is scanned frequently, it is quite computationally expensive. Tree pruning is done from the ground up. It is used to increase the algorithm's prediction and classification performance by reducing the over fitting issue of trees. The decision tree's over fitting issue leads to a misclassification mistake.

Many decision tree-based algorithms exist, such as CART, C4.5, ID3, C5.0, etc. These algorithms have the advantages of fast classification, good learning capacity, and simple design [19].

For our study predicting reading rate levels, the first split, or the root, of the three features (number of errors, the number of words read correctly, and total reading time) is considered, and the WCPM training data is divided into groups based on this split using the greedy algorithm. The three features will have three candidate splits, and then we will calculate how much accuracy each split will cost us. The split that costs least is chosen.

2.3.3 Ensemble Classifier

An Ensemble Classifier (EC) is defined as a collection of individual classifiers that are jointly trained on a dataset in a supervised classification issue. The goal of the ensemble approach is to create a prediction model by combining numerous models. Ensemble approaches are well known for their ability to improve prediction performance. Classification methods for ensemble learning typically combine multiple base classifiers in some way to improve accuracy [20]. These classifiers are capable of solving the same issue and collectively attaining a predicted result with greater accuracy and stability by developing many independent models and joining them. The traditional reasons for using ensemble classifiers to improve effectiveness are representational reasons, statistical concerns, and computational issues. Firstly, a single classifier is not always enough to achieve the best representation in the hypothesis space; thus, independent classifiers must be combined to improve predictive performance. Secondly, if the input dataset is insufficient to train the learning algorithm, the resulting hypothesis may be weak or false. In the latter case, an individual classifier may spend a considerable duration of computing time generating a suitable hypothesis, in which case the procedure is more likely to cause problems. There are five well-known methods of ensemble classifiers: Bagging, Boosting, Voting, Bayesian parameter averaging and Stacking [21].

The bagging method was used for this study. 'Bagging' refers to bootstrap aggregation, in which random samples are drawn by replacing the training datasets. It achieves good classification results and is widely used to build ensemble models. Bagging is a simple method for reducing variance in machine learning algorithms that have a high variance. Learners are divided into two categories: 'stable' and 'unstable.' The variance of stable learners is low, whereas the variance of unstable learners is high. Decision Tree (DT) is an unstable learner who responds to particular training patterns. If the training pattern is altered, then the resulting DT may differ significantly from the original. The prediction accuracy of the resulting DT will change in such a case. Over fitting of training patterns may occur as a result of an individual DT. To overcome the DT's challenge, a bagged DT can be a good option. Individual DT overfits are less of a concern for bagged DT. It gathers the results of multiple DTs by picking a majority vote on their decisions, which helps minimize the over fitting problem and improve the generalization of the DT. As a result, the individual DTs are developed deeply and the trees are not pruned, yielding higher variance and lower bias. These significant characteristics are utilized to combine DT predictions during the bagging stage [22]. See the below **Figure**:

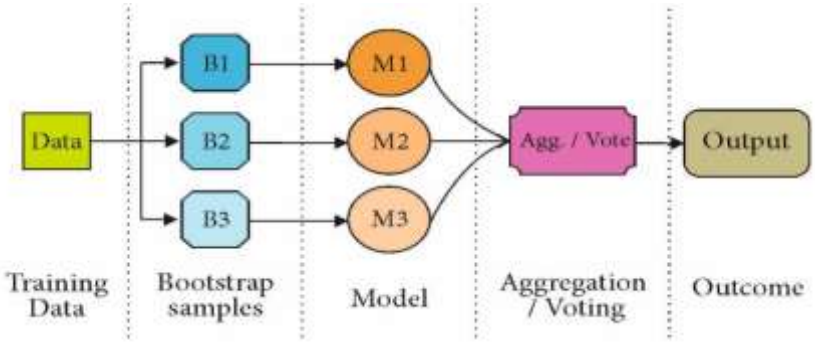


Figure 3. Bagging Ensemble Classifier

In our study, the WCPM training data was divided into several subsets of data chosen randomly with replacement. Each group of subset data is utilized to train their decision trees. After that, it ends up with an ensemble of various models. An average of all the reading rate level predictions from various trees is utilized, which is more robust than a single decision tree.

3. Results and Discussion

This section assesses the efficiency of several methods used to categorize readers based on their levels of reading rate (WCPM). In this study, three machine learning algorithms were used: K-nearest neighbor, Decision Tree and Ensemble Classifiers using Matlab 2021 to predict WCPM for the 2nd year university students of the English language department, University of Duhok, Duhok governorate, Iraq. The participants' assessment of ORF was done in 2021, when a total of 100 participants were tested. Participants' reading rate WCPM was measured using Rasinski's (2004) formulas, and then the calculated WCPM was classified based on the four categories given in the CORFS scale. The input parameters for calculating the WCPM were the number of errors, the number of words read correctly, and the total reading time, and the output was the levels of reading rate.

In order to evaluate the three machine learning algorithms, the dataset was divided into two groups (70% for training and 30% for testing). The three input variables were used to evaluate the levels of WCPM. Using K-fold (k=5), the model was trained using the three machine learning classification techniques and assessed for each classifier's performance using test data. After the testing stage, the data was evaluated using a confusion matrix. The confusion matrix is a successful method for displaying the outcomes of two or more class classification tasks. It simplifies the performance of classifiers on test data and compares the classified data based on its real class label. It indicates that, for each machine learning classifier, the higher the accuracy, the better the model will estimate the real class. To evaluate the performance, the five most prevalent assessment indices were used: Accuracy, Precision, Recall, F1-score, and Receiver Operating Character curve (ROC curve).

Accuracy is the most important performance metric. The accuracy of a classifier reveals its capacity. The higher the accuracy, the better the classifier. So, the main work of our paper is to determine the accuracy of all three classification systems. One will be more accurate than the others, and this will be the optimal algorithm. Accuracy is the computation of all the actual groups that were successfully predicted per the total testing; see the below equation:

$$Accuracy = \frac{corectly\ predicted\ class}{total\ testing\ class} \times 100\% \qquad (2)$$

Precision is the capacity of a classification algorithm to detect only relevant items. Precision is calculated by dividing the number of true positives by the number of true positives plus the number

of false positives. Recall is the capacity of a model to discover all relevant situations within a data collection. Recall is calculated by dividing the number of true positives by the number of true positives plus the number of false negatives, as shown in equations (3) and (4):

$$Precision = \frac{True\ positives}{True\ positives + flase\ positives} \quad (3)$$

$$Recall = \frac{True\ positives}{True\ positives + flase\ negatives} \quad (4)$$

True positives are data values that the model classifies as correct, false positives are data values that the model wrongly identifies as positive but are actually negative, and false negatives are data items identified as negative by the model but are really positive. The higher the accuracy and recall, the better the model's output. However, in other cases, the two contradict each other.

F1-score is also called F measure. The test accuracy in supervised learning issues is measured by F1-score. F1-score is the measure of mean precision and recall

$$F1 - Score = \frac{2\times Precision \times Recall}{Precision + Recall} \quad (5)$$

Figure 4 depicts the accuracy, precision, recall, and F1-score results for the KNN, DT, and EC algorithms. From the figure, it can be seen that the three algorithms have good performance for all classification results, but the EC model outperformed them in terms of accuracy with a value of 0.94 and precision, recall, and F1-score with a value of 0.92 for all of them.

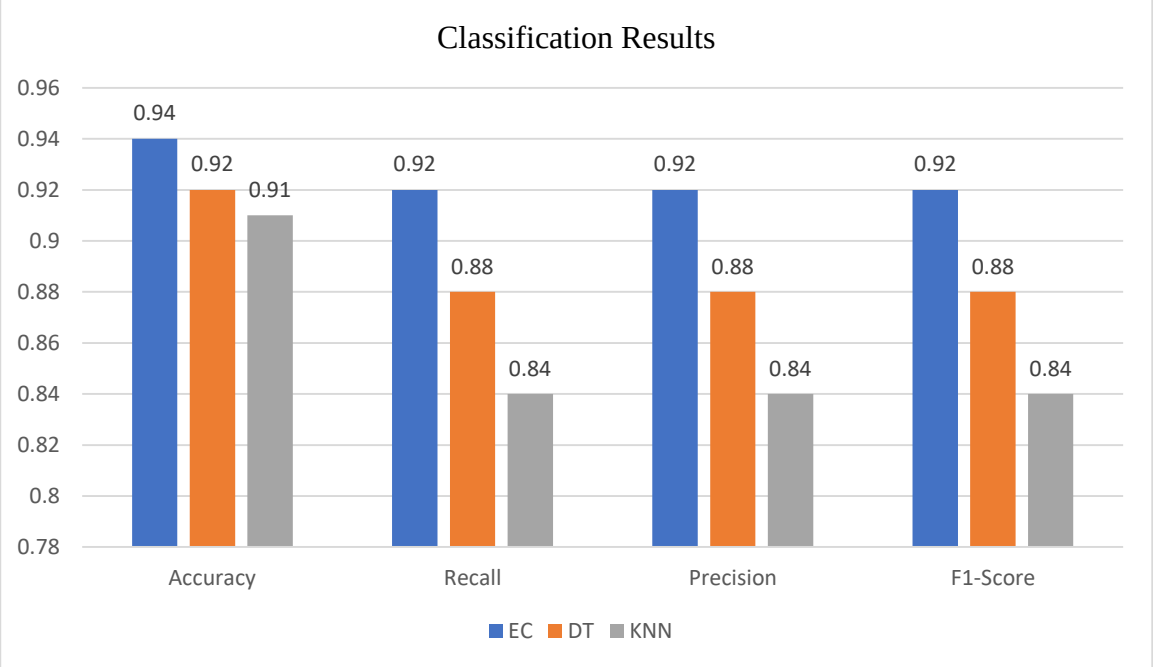
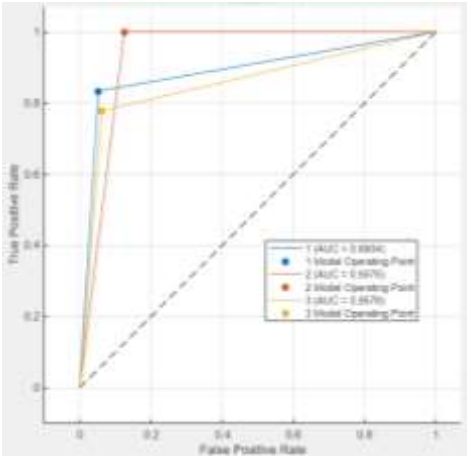


Figure4. The results of Accuracy, Precision, Recall and F1-scor for KNN, DT and EC algorithms

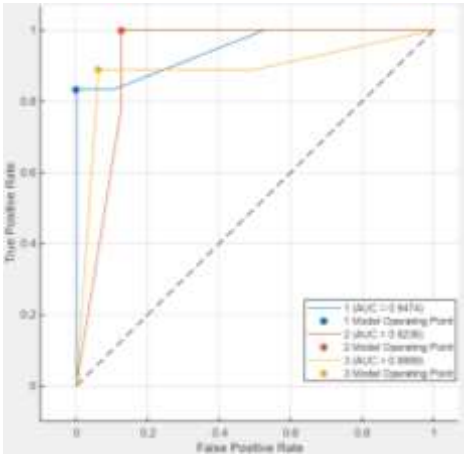
On the other hand, the ROC (Receiver Operating Characteristics) curve is utilized to assess classifier output quality. It is one of the most critical evaluation criteria for determining the effectiveness of any classification technique. Unlike most other metrics, it gives a graphical depiction of a classifier's performance rather than only one value. Classifiers on the ROC graph can be assessed simply by examining their location. ROC curves are commonly used in binary classification to investigate a classifier's performance. The ROC curve is extended for multi-class classification, and the output must be binarized.

In this study, four ROC curves can be created, one for each label, as shown in **Figure 5**. True positive numbers are on the Y-axis, whereas false positives are on the X-axis. The area under each

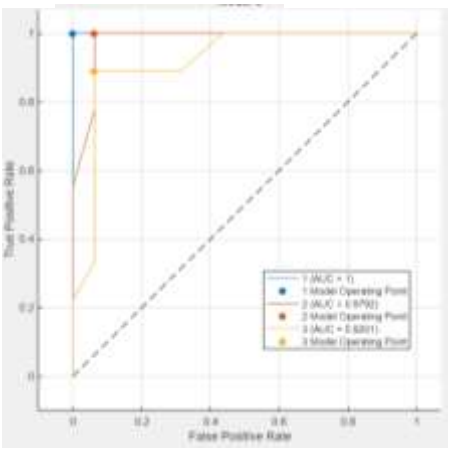
algorithm's ROC curves was used to compare its classification performance on a synthetic set of data. If a classifier has one value of area under the algorithm's ROC curve, then the classifier is very precise. Classifiers that produce curves closer to the top-left corner perform better. If the curve approaches the ROC's 45-degree diagonal space, the classifying data becomes less exact. **Figures (5a-5c)** show the ROC curves for the three algorithms of machine learning: KNN, DT, and EC, respectively. **Figure (5c)** reflects the ROC curve of the EC model, which is considered the ideal classifier because it has 1 value of area under the ROC curve, so the model is more accurate than the other models of classification.



(a) ROC Curve for KNN



(b) ROC Curve for DT



(c) ROC Curve for EC

Figure5. ROC curves for (a) KNN (b) DT (c) EC

Based on the results from **Figures 4 and 5**, it can be concluded that the EC model gained the best outcomes among all other models utilizing various extracted characteristics in terms of the confusion matrix and ROC curve.

4. Conclusion

This paper presents research on the study of reading rates by building models for categorizing readers into four category levels. For training and testing the dataset, Matlab 2021 was used for this purpose. The number of errors, the number of words read correctly, and the total reading time were used as input parameters to predict the output, which was the four various levels of WCPM. The four levels of WCPM were skilled reader, moderate reader, non-fluent reader 2, and non-fluent reader 1. Three machine learning algorithms (KNN, DT, and EC) were used to predict the levels of reading accuracy in WCPM. The dataset was split into 70% for training and 30% for testing. The classification outcomes were assessed utilizing the techniques of classification

assessment called Accuracy, Precision, Recall, F1-Score, and ROC Curve. Experimental results illustrated that ensemble classifier (EC) has performed better at predicting levels of reading rate WCPM than other machine learning classifiers, with accuracy reaching 0.94 and precision, recall, and F1-score reaching 0.92 for all of them. In terms of the ROC curve, EC also achieved the highest performance.

References

1. Grabe, W. Fluency in reading—thirty-five years later. *Reading in a Foreign Language* **2010**, *22*, *1*, 71-83.
2. Nation, P. Developing fluency. *In Exploring EFL Fluency in Asia*; Muller T., Adamson, J., P. Brown, S. Herder., Eds.; Springer: Palgrave Macmillan, **2014**, 11-25.
3. Rasinski, T. V. Assessing reading fluency. *Pacific Resources for Education and Learning (PREL)* **2004**.
4. Morrison, T.G.; Wilcox, B. Assessing expressive oral reading fluency. *Education Sciences* **2020**, *10*, *59*, 1- 13.
5. Williams, J. L.; Skinner, C. H.; Floyd, R. G.; Hale, A. D.; Neddenriep, C.; Kirk, E. Words correct per minute: The variance in standardized reading scores accounted for by reading speed. *Psychology in the Schools* **2011**, *48*, 87–101.
6. Mahendra, M. W.; Marantika, I. M. Y. The Phonological Interference in EFL Reading. *ELLITE: Journal of English Language, Literature, and Teaching* **2020,** *5*, *1*, 27-34.
7. Rahmawati, A.; Rosmalina, I.; Anggraini, H.W. Prosodic reading and reading comprehension in university. *EduLite: Journal of English Education, Literature, and Culture* **2020**, *5*, *1*, 89- 108.
8. Sonuç, E. Thyroid Disease Classification Using Machine Learning Algorithms. In *Journal of Physics: Conference Series* **2021,** *1963,1*, 012140.
9. Sajja, G. S.; Mustafa, M.; Ponnusamy, R.; Abdufattokhov, S. Machine learning algorithms in intrusion detection and classification. *Annals of the Romanian Society for Cell Biology* **2021**, *25*, *6*, 12211-12219.
10. Ahmad, A.; Ostrowski, K. A.; Maślak, M.; Farooq, F.; Mehmood, I.; Nafees, A. Comparative study of supervised machine learning algorithms for predicting the compressive strength of concrete at high temperature. *Materials* **2021,** *14,15*, 4222.
11. Bolaños, D.; Cole, R. A.; Ward, W. H.; Tindal, G. A.; Hasbrouck, J.; Schwanenflugel, P. J. Human and automated assessment of oral reading fluency. *Journal of educational psychology* **2013**, *105*, *4*, 1142.
12. Kim, H.; Hannah, L.; Jang, E. E. Using acoustic features to predict oral reading fluency of students with diverse language backgrounds. In *Collated Papers for the ALTE 7th International Conference, Madrid* **2021**,198.
13. Day, R. R. & Harsch, K. *Cover to cover 2 reading comprehension and EFL reading. English Language, Literature, and Teaching fluency*. Oxford University Press, **2008**.
14. Adithiyaa, T.; Chandramohan, D.; Sathish, T. Optimal prediction of process parameters by GWO-KNN in stirring-squeeze casting of AA2219 reinforced metal matrix composites. *Materials Today: Proceedings* **2020**, *21*, 1000-1007.
15. Miranda-Vega, J. E.; Rivas-Lopez, M.; Fuentes, W. F. K-nearest neighbor classification for pattern recognition of a reference source light for machine vision system. *IEEE Sensors Journal* **2020,** *21,10*, 11514-11521.

16. Wibowo, A. H.; Oesman, T. I. The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency. In *Journal of Physics: Conference Series* **2020**, February,*1450,1,* 012076.

17. Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends* **2021**, *2, 1,* 20-28.

18. Lee, C. S.; Cheang, P. Y. S.; Moslehpour, M. Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences* **2022**, *26,1,* 1-29.

19. Li, X.; Yi, S.; Cundy, A. B.; Chen, W. Sustainable decision-making for contaminated site risk management: A decision tree model using machine learning algorithms. *Journal of Cleaner Production* **2022**, *371*, 133612.

20. Matloob, F.; Ghazal, T. M.; Taleb, N.; Aftab, S.; Ahmad, M.; Khan, M. A.; ...; Soomro, T. R. Software defect prediction using ensemble learning: A systematic literature review. *IEEE Access* **2021**, *9*, 98754 - 98771

21. Alamir, M. A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers. *Applied Acoustics* **2021**, 175, 107829 and Culture, *5, 1,* 89-108

22. Mishra, P. K.; Yadav, A.; Pazoki, M. A. novel fault classification scheme for series capacitor compensated transmission line based on bagged tree ensemble classifier. *IEEE Access* **2018**, *6*, 27373-27382.