Stuttering disfluency detection using machine learning approaches

Stuttering disfluency detection using machine learning approaches

Abedal-Kareem Al-Banna, Department of Computer Science, Loughborough University, UK, A.Al-Banna@lboro.ac.uk

Eran Edirisinghe, School of Computing & Mathematics, Keele University, UK, e.edirisinghe@keele.ac.uk

Hui Fang, Department of Computer Science, Loughborough University, UK, H.Fang@lboro.ac.uk

Wael Hadi, Department of Information Security, University of Petra, whadi@uop.edu.jo

Abstract

Stuttering is a neurodevelopmental speech disorder wherein people suffer from disfluency in speech generation. Recent research has applied machine learning and deep learning approaches to stuttering disfluency recognition and classification. However, these studies have focused on small datasets, generated by a limited number of speakers and within specific tasks, such as reading. This paper rigorously investigates the effective use of eight well-known machine learning classifiers, on two publicly available datasets (FluencyBank and SEP-28k) to automatically detect stuttering disfluency using multiple objective metrics, i.e., prediction accuracy, recall, precision, F1-score, and AUC measures. Our experimental results on the two datasets show that the Random Forest classifier achieves the best performance, with an accuracy of 50.3% and 50.35%, a recall of 50% and 42%, a precision of 42% and 46%, and an F1 score of 42% and 34%, against the FluencyBank and SEP-28K datasets, respectively. Moreover, we show that the machine learning based approaches may not be effective in accurate stuttering disfluency evaluation, due to diverse variations in speech rate, and differences in vocal tracts between children and adults. We argue that the use of deep learning approaches and Automatic Speech Recognition (ASR) with language models may improve outcomes, specifically for large scale and imbalanced datasets.

1. Introduction

Stuttering or stammering are interchangeable terms used for the same speech disfluency. Stuttering is recognised in records of language pathology as one of the oldest speech impediments [1]. The World Health Organisation (WHO) describes stuttering as "Speech that is characterised by frequent repetition or prolongation of sounds or syllables or words, or by frequent hesitations or pauses that disrupt the rhythmic flow of speech. It should be classified as a disorder only if its severity is such as to markedly disturb the fluency of speech. " [2].

Around 1% of the global population suffer from noticeable stuttering, which means approximately 70 million people worldwide. Around 25% of all stuttering cases are female [3], i.e. the male to female ratio is 3:1. Moreover, 5% of all children between four to six years old go through a period of stuttering in their lives, which may remain for approximately six months or may develop into a chronic disorder [4]. While there are several types of stuttering, such as Neurogenic, Developmental and Psychogenic, they may have common symptoms, and the same disfluency events[5] as tabulated in table 1. In speech pathology, disfluency events include the following signs:- repetition, interjection, prolongation and block [6]–[8].

Table 1 Stuttering disfluency events

Event	Description	Example
Prolongation	Long time word pronunciation	Hoooow are you?
Block	Uncontrolled pause and interruptions of speech	I need (Stop) you
Phrase repetition	The duplication of a whole phrase	I need I need your help
Sound repetition	Duplication of the first letter of the word	f-f-f-facebook
Word repetition	Duplication of a whole word	Hi Hi my friend
Interjection	Adding extra sounds	Um,uh

Only few previous research attempts have investigated the use of artificial intelligence (AI) approaches, e.g., machine learning and deep learning algorithms, in disfluency detection [9], [10]. However, these approaches have been used in general applications within speech audio research, for e.g., in detection, classification and recognition tasks. Most of these studies employed deep learning and machine learning in speech interaction applications for blind people, elderly, and limited hand dexterity applications. Stuttering disfluency detection is a widely recognised challenge cited in literature due to diverse determinants such as speech rate, vocal tracts for children and adults, the lack of transcribed and annotated training data [7], [8]. Despite the good results reported in previous stuttering classification research, these studies conducted analysis on selected samples of public datasets, involving limited number of speakers. Therefore the performance and reliability of these models cannot be reliably concluded. [11].

In this work, eight widely used data mining algorithms are employed for automatic detection of stuttering disfluency, against two publicly available datasets [12], [13]. It is expected that the outcomes of this paper may support speech pathologist decisions by the early of detecting stuttering disfluency, mainly in the case of children who suffer from developmental stuttering.

The rest of this paper is organised as follows: Section 2 summarises literature on stuttering disfluency recognition. Section 3 presents the proposed methodology; Section 4 evaluates the experimental results; whilst Section 5 summarises the main results from this study and concludes the work.

2. Literature review

Only limited research work exist in literature that investigates stuttering disfluency detection and classification. This section reviews these previous studies.

The founders of the UCLASS dataset [14], introduced the first model in stuttering classification (SC). A fully connected artificial neural network ANN was applied on two stuttering events prolongation and repetition. Autocorrelation, spectral features and envelope parameters were extracted from the audio signal and used as input to the network. The reported model accuracy was 80%, in the prolongation events.

Howell et al. [15], appropriated a neural network ANN model to classify repetition and prolongation events. To achieve increased accuracy, they recorded speech samples from 12 speakers.
Moreover, manual segmentation of linguistic units and disfluency categorisation for each segment were followed before feature extraction. Furthermore, fragmentation measures, spectral measures, energy, and duration features of an acoustic signal were captured and used in training the model. The results determined that the most suitable parameters for SC were spectral fragmentation whole-word events and the duration and Supra-lexical for part-words event.

The authors of [16], suggested an ANN with formant frequencies and its amplitude to detect vowel prolongations, syllable repetitions and stopgaps. An empirical study was conducted on six fluent speakers and six people with stuttering. The accuracy achieved was 78.1%.

Several research works [17]–[19] used HMMs in stuttering classification and recognition. The authors of [19] developed Malay Speech Therapy Assistance Tools that help pathologists diagnose and train children who stutter. The system model was trained on 35 stuttering samples on certain words and the model's reported accuracy was 96% with MFCC feature extraction had been utilised. Additionally, the authors of [18] proposed two SC techniques of prolonged fricative phonemes using MFCC and HMM on 38 speech samples. The reported accuracy of their model was 80%. Furthermore, the authors of [17], utlised task-oriented finite state transducer (FST) lattices to detected revision, sound and word repetitions disfluencies. They also used amplitude thresholding methods to detect prolongations in stuttering speech. These techniques emerged in an average 37% error rate across the four various kinds of disfluencies.

Mahesha et al. [1], demonstrated a disfluency classification model based on the Gaussian mixture model (GMM) and MFCC feature extraction. The model trained on 50 selected samples from the UCLASS dataset for four disfluency events (prolongation, syllable repetition, word repetition, and interjection). The research concluded that the classification of disfluencies could be a pattern recognition problem, and the model's accuracy depends on the number of mixture components and the MFCC coefficients because it may allow the best-fit data representation. The average accuracy reported was 96%.

Villegas et al. [4], measured respiratory and heart rate biosignals of 68 participants to separate them into two-classes, stutter or not-stutter, with the study focused in reading tasks. The Multilayer perceptron (MLP) with 40 hidden layers with statistical features such as mean, standard deviation were employed to make a binary classification of the data. The reported model accuracy was 82.6%. Dash et Dash et al. [20], suggested a method to correct and recognise stuttering events within a give time. Amplitude thresholding within was applied on speech samples to extract prolongation events. Additionally, Text to speech (TTS) was implemented to remove repetition events. The model achieved 86% accuracy.

The authors of [9], proposed a deep neural network model based on Bidirectional long short-term memory (BI-LSTM) and residual network of six blocks and 18 convolutional layers, to classify six disfluencies. The average error rate reported was 10.03% on the UCLASS dataset. A recurrent network with BLSTM followed by Integer linear programming (ILP) as post-processing step was employed in the work of Zayats et al. (2016). This research proposed a different method in disfluency classification for word sequences. Pattern matching features were designed to decrease the vocabulary size in training, which improved the word sequence's performance; the miss rate reported was 19.4% over each repetitions disfluency.

Lea et al. [12] create a new dataset SEP-28K with a 32k audio clip, 28k collected from public podcasts and 4k clip from the Fluencybank dataset. The data were annotated with five disfluency events, i.e. sound repetitions, word repetitions, blocks, prolongations , and interjections. In addition, they proposed a ConvLSTM stuttering detection model based on LSTM and a Convolutional Neural Network, using CCC as the loss function. The F1 score reported for this model was 66.2 on SEP-28K dataset and 75.8 on Fluencybank dataset.

Table 2 summarises some of the previous work in stuttering classification and recognition.

Table 2 : Previous Works in stuttering classification and recognition

Author	**year**	**Classifier**	**Features**	**Dataset size**	**Disfluency event**	Approximate **Result**
[14]	1995	ANN	Autocorrelation (ACF), spectral features and envelope parameters	N/A Custom Dataset	Prolongation, repetition	Best Acc:80%

[15]	1997	ANN	Duration, energy peaks	12 sample	Repetitions, prolongations	Best Acc: 95% of fluent, Best Acc: 78% of disfluent
[16]	2003	ANN	Formant frequencies and its amplitude	12 speech sample contains 6 fluent and 6 with stopgaps	Stop-gaps, syllable, repetitions, vowel prolongations	Acc: 78.1%
[19]	2007	HMM	MFCC	20 sample of fluent speech,15 of artificial shuttering speech.	Repetition, prolongation, blocks	Acc: 96% for normal speech, 90% for artificial shuttering speech.
[21]	2009	SVM	MFCC	15 samples of speech	Repetition, non-repetition	90.35%
[22]	2012	k-NN	LPC, LPCC, WLPCC	39 Samples from UCLASS	Prolongation, repetition	WLPCC 97.06% LPCC 95.69% LPC 93.14%
[23]	2014	SVM	MFCC	16 Samples from UCLASS	Word repetition, prolongation	98.00%
[24]	2016	BLSTM	Word Embedding	Switchboard Corpus	Repetition	F1: 85.9 MR: 19.4
[1]	2016	GMM	MFCC	50 UCLASS samples	Word repetition, syllable repetition, interjection and prolongation.	96%
[17]	2018	Finite State Transducer, Amplitude and Time Thresholding	Word Lattice	129 samples UCLASS	Sound, word, part-word and phrase repetition, and revision, prolongation	Avg. MR: 37% false positive rate FPR 0.89%

[20]	2018	STT, Amplitude Thresholding	Amplitude	60 Speech Samples		Acc.: 86%
[9]	2019	FluentNet	Spectrogram	25 speech sample UClass	Phrase repetition, sound repetition, word repetition, revision, prolongation, interjection	MR. 9.35 Acc .91.75 On UClass 13.03 86.70 On LibriStutter
[7]	2020	(BI-LSTM) and residual network	Spectrogram	25 speech sample UClass	Phrase repetition, sound repetition, word repetition, revision, prolongation, interjection	The average miss rate 10.03% on UCLASS dataset Acc :91.15%
[12]	2021	LSTM and CNN	Filterbank features. Pitch articulatory features	SEP-28k 28000 audio clip Fluencybank 4000 audio clip	Interjection, sound repetition, word repetition, prolongation	F1 score on SEP-28K is 66.2 F1 score on Fluencybank 75.8

Due to the lack of datasets that can be used for stuttering disfluency analysis, previous related work has been limited to the UCLASS dataset or artificially generated data [7]–[9], [12]. Despite the acceptable results presented by these studies, most of them investigate stuttering detection on adults rather than on children and for reading related speech, not for the ideal, spontaneous speech. This paper will examine eight machine learning algorithms on a new dataset provided by Apple [12], which contains reading and spontaneous speech for children and adults who stutter. This dataset is aligned better with real data related to stuttering speech.

3. Methodology

This section goes through the methodology used to investigate the stuttering disfluency classification models in this paper. Our methodology is based on the Cross-industry standard process for data mining (CRISP-DM) [25], which is one of the most commonly used data mining process models. Figure 1 shows a block diagram of our methodology. The methodology is divided into five stages, as seen in the diagram: problem understanding, data understanding, data preparation, modelling, and evaluation. The following subsections detail the key stages of this methodology.

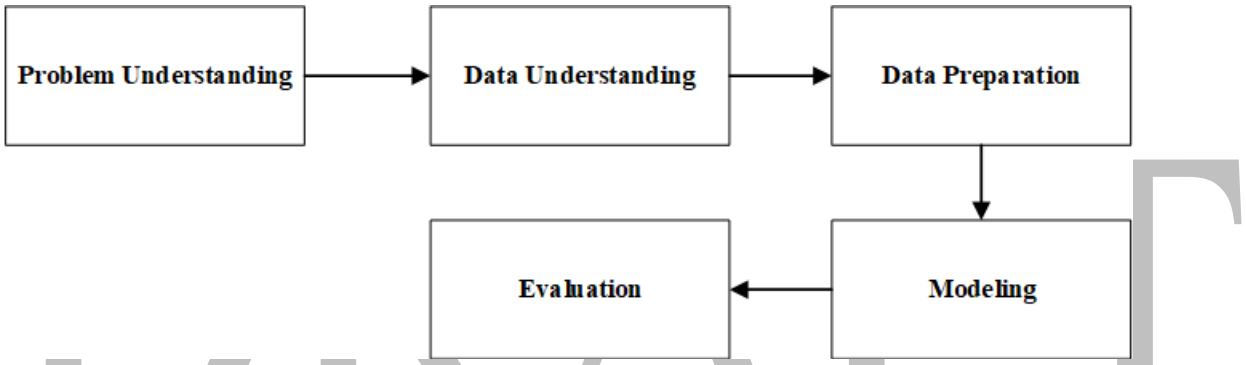

Figure 1: Proposed methodology

3.1.1 Problem understanding

Stuttering is a disorder that is relatively easy to treat in many cases, especially in the early life (pre-school) stage of an individual. However, it is a difficult problem to deal with in other cases, particularly in the adolescent stage. Moreover, for an individual, the impact of stuttering goes well beyond creating difficulties in free communication with others. It could affect the academic competency, social-emotional functioning, and independent functioning of children who stutter. Therefore, attention must be paid to any problems in the development of speech of children and should not be neglected.

Preliminary diagnosis and early detection of stuttering severity can reduce the risk of chronic disorder and adversely affect children who stutter and their families [26], [27]. However, the early stuttering detection and diagnosis are not affordable for many, because the assessment sessions are costly, difficult to find in some countries / communities and time-consuming for pathologists and the individuals affected [1].

The stuttering severity evaluation is a time consuming process that requires significant effort. The speech pathologist (SP) needs to track and observe all stuttering events such as disfluency, physical

concomitants, and speech's naturalness. However, the key activity in stuttering assessment is disfluency observation.

Given the above reason, automated, computer based stuttering disfluency detection systems could play a vital role in addressing the above practical challenges.

3.1.2 Data understanding

We conducted our experiment on (SEP-28K) [12] and fluency bank datasets [13]. The SEP-28K dataset contains 28,177 audio clips derived from 385 episodes of 8 stuttering YouTube podcasts. For each episode, 40-250 3-second intervals near the speech pause segment had been extracted and annotated. Further, in [12], 3.5 hours audio clips taken from the fluency bank dataset [13]were annotated and validated using Fleiss Kappa inter-annotator agreement [12]. In our experiment, six stuttering events each, in both datasets, have been investigated. Figure 2 shows these events and the number of occurrences of each event, in each dataset.

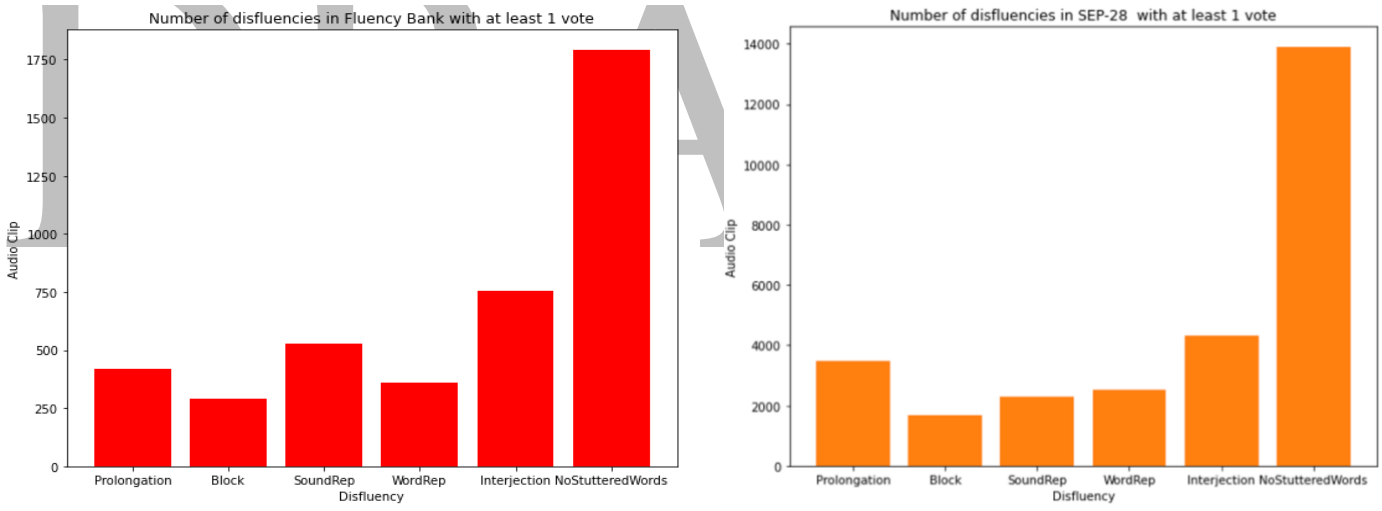

Figure 2: Number of disfluencies in Fluency Bank and SEP-28 datasets

The figure presents the total number of disfluency events in the Fluency bank and SEP-28K datasets for five different disfluency classes and one fluent class. The disfluency events include: - Prolongation, Block, Sound repetition, word repetition, and Interjection. Overall, we can observe that the Non-stuttering event is the most frequent event in both datasets. It is also clear that the distribution of data is imbalanced amongst events.

3.1.3 Data preparation

Acoustic feature extraction is vital in audio classification. The most widely used acoustic feature in stuttering disfluency classification is Mel-Frequency Cepstral Coefficients (MFCC) [1], [9]. The MFCC is a quasi-logarithmic frequency scale that emulates the auditory system in a human. In our experiment, we converted the time-domain audio clip into frequency domain using the Short-Time Fourier Transform (STFT). Subsequently, we extracted 40 MFCC vector representations for each disfluency audio clip. Figure 3 illustrates the signal processing steps involved.

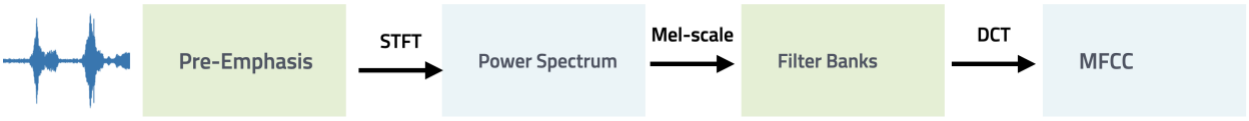

Figure 3: Signal processing steps

Firstly, the pre-emphasis filter was applied to the waveform signal to improve the signal to noise ratio and balance the frequency spectrum. This filter was applied to the time-domain signal (s) with α (=0.95) filter coefficient, as follows:

$$y(t) = s(t) - \alpha s(t-1) \quad (1)$$

Secondly, the filtered signal was divided into a short-time frame signal with a 25 ms frame size and 10 ms frame stride. After framing, hamming window function was applied as follows:

$$w[n] = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \quad (2)$$

Thirdly, the Short-Time Fourier-Transform (STFT) was applied on each frame with 512 NFFT and the power spectrum computed as the following equation.

$$P_s = \frac{|FFT(x_i)|^2}{N} \quad (3)$$

Fourthly, 40 triangular filters are applied as filter banks on a mel-scale to the power spectrum of the signal to extract frequency bands.

$$m = 2595\log_{10}(1+\frac{f}{700}) \quad (4)$$

$$f = 700(10^{m/I2595} - 1) \quad (5)$$

Finally, we applied the Discrete Cosine Transform (DCT) to decorrelate the filter bank coefficients to Mel-frequency cepstral coefficients. In our experiment, 13 cepstral coefficients are maintained. Figure 4 illustrates the waveform signal, Mel spectrogram and MFCC of the word-repetition disfluency.

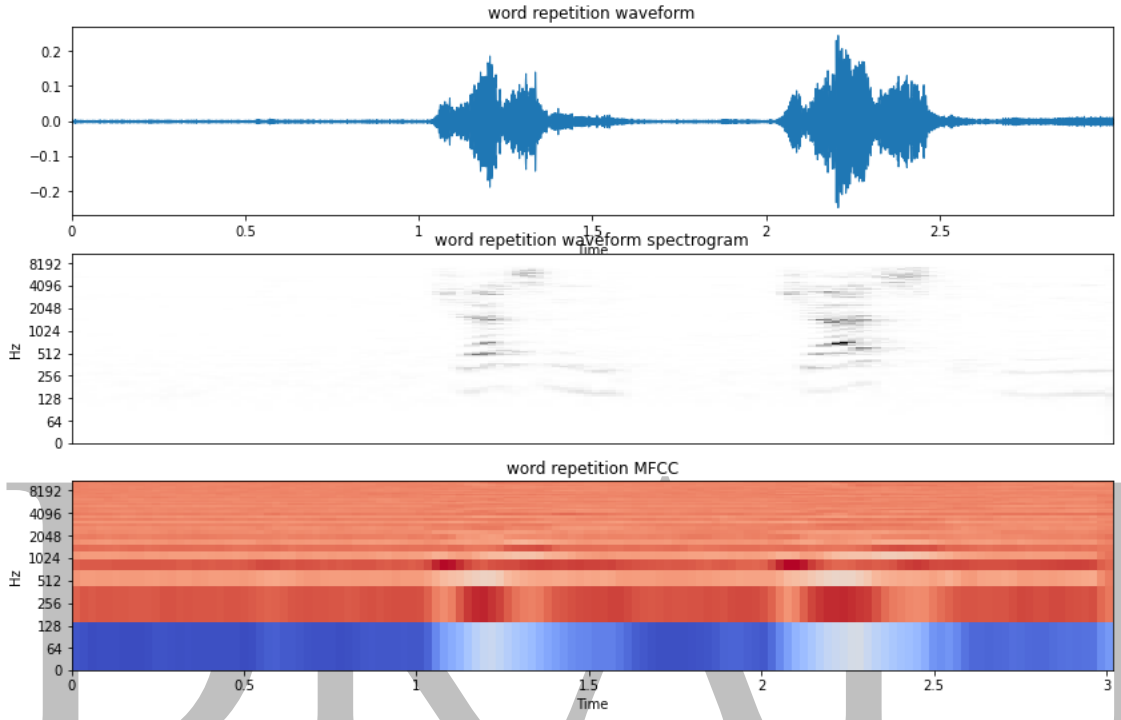

Figure 4: Waveform, spectrogram, and MFCC representations

3.1.4 Modelling

Eight widely used classifiers are utilized to identify stuttering disfluency during the classification process. The selected classifiers are SVM with a linear kernel, SVM with a RBF kernel, Random Forest, Decision Trees, AdaBoost, k-Nearest Neighbors, Gaussian Naïve Bayes, and Quadratic Discriminant Analysis. These classifiers have been used extensively in literature and are very simple to implement and use. The authors selected a public implementation for these classifiers in this study, which may make it easier for scholars to replicate and compare our findings. It should be noted that all the classifiers have been configured using SkLearn Library (Scikit Learn) in Python [28]. This stage's goal is to implement a classification model for identifying stuttering disfluency.

3.1.5 Evaluations

We utilize five well-known performance evaluators in the literature for speech applications to assess the classification model's performance [29]. These evaluation metrics are: Prediction accuracy, Precision, Recall, F1, and AUC curve area (Area Under Curve). The authors used a

standard 10 folds cross-validation procedure to evaluate all five measures, which provides an objective measure of how well the model fits and how well it will generalize to new data [30]. The aim of this stage is to figure out which classifier is the most effective in identifying stuttering disfluency.

Experimental results

Two publically available stuttering disfluency identifying datasets were investigated to evaluate the performance of the implemented classification models. The authors trained and tested all the classifiers upon all two datasets. Later we evaluated the classifiers by their prediction accuracy, precision, recall, AUC, and F1 measures on different datasets.

Table 3 shows the predication accuracy results. On the SEP-28K and FluencyBank datasets, the Random Forest classifier performed best according to prediction accuracy evaluator, and the decision tree classifier resulted in the worst performance. In particular, the Random Forest outperformed SVM with a RBF kernel, SVM with a linear kernel, k-nearest neighbors, AdaBoost, Quadratic Discriminant Analysis, Gaussian Naïve Bayes, and decision trees against FluencyBank dataset by 0.48%, 2.89%, 5.06%, 5.43%, 6.99%, 13.75%, and 14.84%, respectively. On the SEP-28K dataset, the Random Forest outperformed SVM with a RBF kernel, SVM with a linear kernel, AdaBoost, k-nearest neighbors, Quadratic Discriminant Analysis, Gaussian Naïve Bayes, and decision trees by 0.37%, 1.54%, 1.73%, 5.02%, 5.37%, 8.14%, and 17.29%, respectively.

Table 3: Prediction accuracy results

Classifiers	FluencyBank	SEP-28K
SVM with a RBF kernel	49.82%	49.98%
SVM with a linear kernel	47.41%	48.81%
decision trees	35.46%	33.06%
AdaBoost	44.87%	48.62%
k-nearest neighbors	45.24%	45.33%
Random Forest	50.30%	50.35%
Quadratic Discriminant Analysis	43.31%	44.62%
Gaussian Naïve Bayes	36.55%	42.21%

This finding is compatible with prior study [30], [31], which have found that the Random Forest is the best classifier when the prediction accuracy is considered. The random forest constructs multiple decision trees during the training phase, which enhances classification.

Recall results of the eight classifiers against the SEP-28K and FluencyBank datasets are shown in Figure 5. The Random Forest classifier achieves the best performance, with a recall of 50% and 50% on the SEP-28K and FluencyBank datasets, respectively. The decision trees classifier obtains the lowest value among the eight classifiers, with a recall of 33% and 35% on the SEP-28K and FluencyBank datasets, respectively.

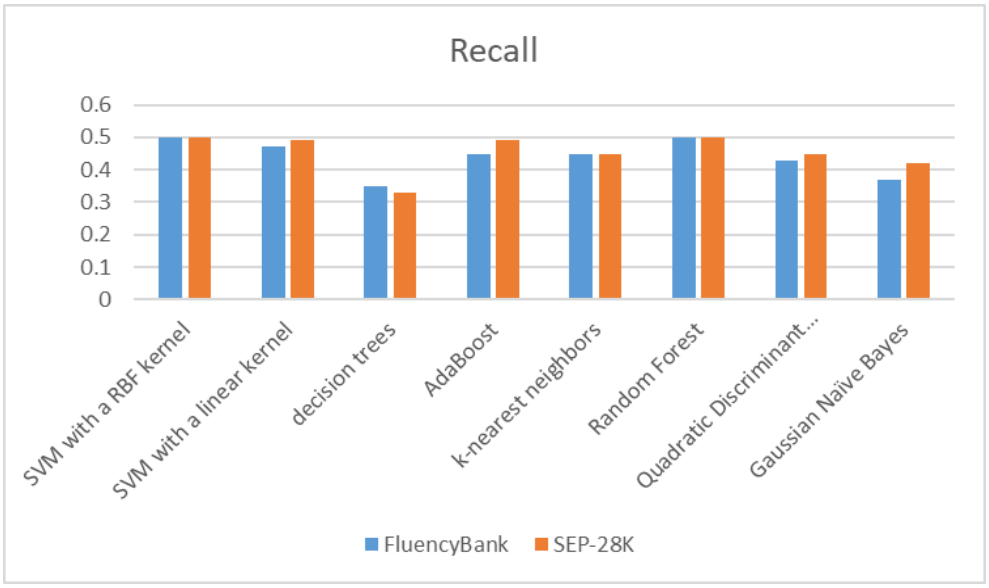

Figure 5: Recall results of the eight classifiers against the SEP-28K and FluencyBank datasets

Figure 6 shows the precision results of the eight classifiers against the SEP-28K and FluencyBank datasets. The Quadratic Discriminant Analysis classifier achieves the best performance, with a precision of 44% on the FluencyBank dataset. The SVM with a RBF kernel classifier achieves the best performance, with a precision of 48% on the SEP-28K dataset. The SVM with a linear kernel classifier obtains the lowest value among the eight classifiers, with a precision of 24% and 29% on the SEP-28K and FluencyBank datasets, respectively. On the other hand, the Random Forest obtains comparable results with the Quadratic Discriminant Analysis and the SVM with a RBF kernel classifiers.

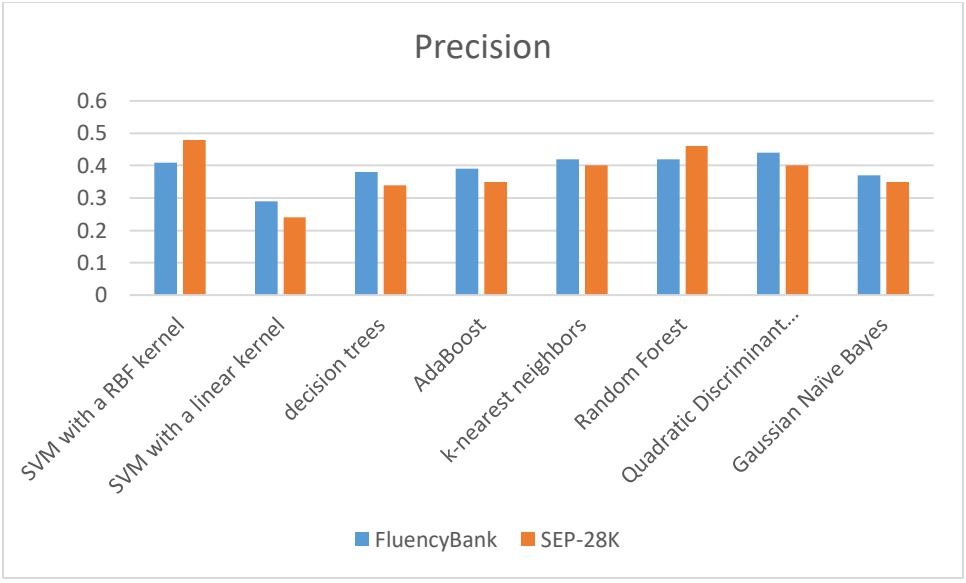

Figure 6: Precision results of the eight classifiers against the SEP-28K and FluencyBank datasets

Further, the F1 results of the eight classifiers against the SEP-28K and FluencyBank datasets are shown in Figure 7. The Quadratic Discriminant Analysis classifier achieves the best performance, with a F1 of 44% on the FluencyBank dataset. The k-nearest neighbor classifier achieves the best performance, with a F1 of 42% on the SEP-28K dataset. The SVM with a linear kernel classifier obtains the lowest value among the eight classifiers, with a F1 of 32% and 34% on the SEP-28K and FluencyBank datasets, respectively.

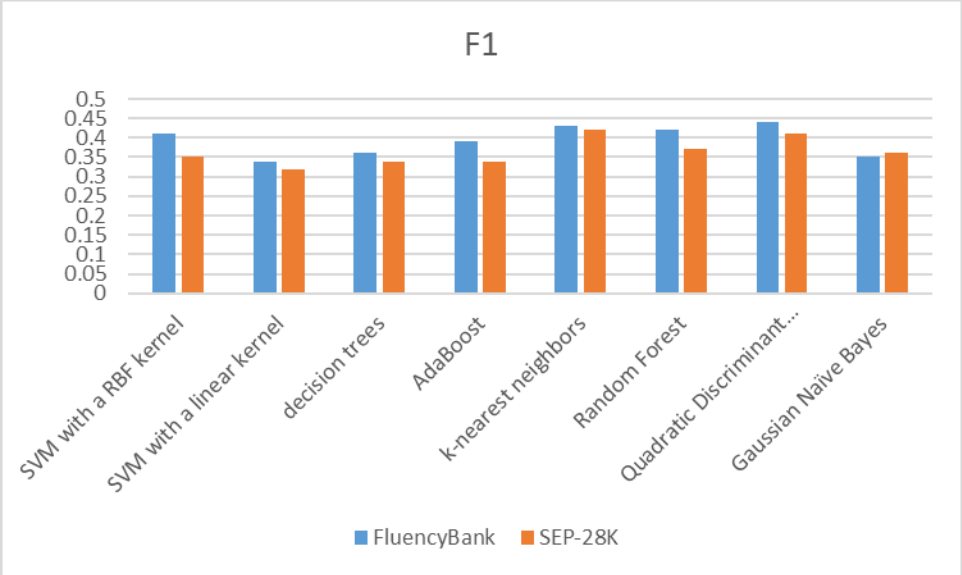

Figure 7: F1 results of the eight classifiers against the SEP-28K and FluencyBank datasets

AUC curves of the eight classifiers on FluencyBank and SEP-28K datasets are shown in Tables 4 and 5. According to the results, we observe that the prediction performance of the eight classifiers are relatively similar. However, all the classifiers present acceptable AUC curves and the ability to detect people with or without stuttering (AUC > 0.50). Meanwhile, the Random Forest classifier achieves the best prediction performance among the seven classifiers.

Table 4: AUC weights of the eight classifiers on FluencyBank dataset

Classifiers	Prolongation	Block	SoundRep	WordRep	Interjection	NoStutteredWords
SVM with a RBF kernel	0.57	0.58	0.58	0.58	0.66	0.58
SVM with a linear kernel	0.55	0.61	0.58	0.56	0.66	0.62
decision trees	0.50	0.51	0.53	0.52	0.53	0.54
AdaBoost	0.56	0.59	0.59	0.57	0.62	0.57
k-nearest neighbors	0.58	0.54	0.58	0.56	0.62	0.60
Random Forest	0.64	0.60	0.66	0.62	0.70	0.65
Quadratic Discriminant Analysis	0.61	0.61	0.64	0.62	0.67	0.62
Gaussian Naïve Bayes	0.55	0.60	0.57	0.57	0.60	0.56

It should be noted that we achieve lower results in comparison with previous studies [1], [15], [19] because these studies conducted the analysis on a small dataset with limited speakers where the accuracy and reliability of the model may not be confidently concluded. Further, the lower results we obtained are due to the fact that the FluencyBank, and SEP-28K are imbalanced datasets and the disagreement between the human evaluators involved in the annotation process in [12]. Finally, from all experiments, we can conclude that all eight classifiers produce acceptable classification accuracy and AUC rates. However, these results may not be effective in the stuttering evaluation process due to diverse determinants such as speech rate, vocal tracts for children and adults. Therefore, other deep learning approaches and end to end ASR may enhance these results for large scale datasets.

Table 5: AUC weights of the eight classifiers on SEP-28K dataset

Classifiers	Prolongation	Block	SoundRep	WordRep	Interjection	NoStutteredWords
SVM with a RBF kernel	0.64	0.54	0.64	0.56	0.72	0.59
SVM with a linear kernel	0.47	0.53	0.57	0.58	0.58	0.66
decision trees	0.53	0.54	0.49	0.52	0.56	0.58
AdaBoost	0.63	0.64	0.62	0.55	0.62	0.62
k-nearest neighbors	0.58	0.65	0.62	0.61	0.68	0.64
Random Forest	0.68	0.65	0.70	0.65	0.76	0.73
Quadratic Discriminant Analysis	0.67	0.69	0.68	0.65	0.71	0.70
Gaussian Naïve Bayes	0.61	0.59	0.65	0.61	0.65	0.63

4. Conclusions and future implications

Stuttering may affect the quality of life of children and adults and prevent them from freely participating in daily activities. This research work will support speech-language pathologists automatically (computer-based) detect stuttering events, in stuttering severity assessments as these assessments are usually managed manually. Moreover, the proposed, machine learning based fluency enhancement and evaluation system may help cure stuttering, especially in children of preschool age, by enabling early intervention.

This paper investigated the use of eight widely used machine learning classifiers on recently published datasets, SEP-28K and FluencyBank, to identify stuttering disfluency. Five evaluation measures were used in the investigation conducted. Experimental results and analysis proved that the Random Forest classifier outperforms the remaining seven classifiers in prediction accuracy, while KNN and Quadratic Discriminant Analysis outperform other classifiers in the F1 score. In particular, the Random Forest classifier obtains the best performance, with an accuracy of 50.3%, a recall of 50%, a precision of 42%, F1 score of 42% against the FluencyBank dataset.

Based on these results, we conclude that the machine learning classifiers can be useful and appropriate for addressing the stuttering disfluency classification on small datasets and low dimensionality frequency domain features such as MFCC. Accordingly, other deep learning approaches and end to end ASR may be more reliable to classify stuttering disfluencies. Therefore, we intend to investigate new disfluency detection approaches and employ other time-domain features such as autocorrelation function and zero-crossing rate for specific disfluency types, which may enhance the classification result. Further, we will investigate employing language

models for specific disfluency events such as word repetitions and sound repetitions. Moreover, the generative adversarial networks could be used to mitigate the effect of imbalanced datasets.

5. Acknowledgement

The authors would like to thank Dr Saadi Abadi and Mohammad Arafah for their support and helpful feedback during this research.

References

[1] P. Mahesha and D. S. Vinod, "Automatic segmentation and classification of dysfluencies in stuttering speech," in *ACM International Conference Proceeding Series*, Mar. 2016, vol. 04-05-Marc, doi: 10.1145/2905055.2905245.

[2] World Health Organisation, "ICD-10 Version:2010," *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Version for 2010*. 2010, Accessed: Feb. 06, 2021. [Online]. Available: https://icd.who.int/browse10/2010/en#/F98.5.

[3] G. Manjula, M. Shivakumar, and Y. V. Geetha, "Adaptive optimization based neural network for classification of stuttered speech," in *ACM International Conference Proceeding Series*, Jan. 2019, pp. 93–98, doi: 10.1145/3309074.3309113.

[4] B. Villegas, K. M. Flores, K. Pacheco-Barrios, and D. Elias, "Monitoring of respiratory patterns and biosignals during speech from adults who stutter and do not stutter: A comparative analysis," in *International Symposium on Medical Information and Communication Technology, ISMICT*, May 2019, vol. 2019-May, doi: 10.1109/ISMICT.2019.8743844.

[5] D. Iimura, N. Asakura, T. Sasaoka, and T. Inui, "Abnormal sensorimotor integration in adults who stutter: A behavioral study by adaptation of delayed auditory feedback," *Front. Psychol.*, vol. 10, no. OCT, 2019, doi: 10.3389/fpsyg.2019.02440.

[6] K. B. G Riley, "SSI-4: Stuttering Severity Instrument - Fourth Edition KIT Glyndon D. Riley : PRO-ED Inc. Official WebSite." 2009, Accessed: Feb. 02, 2021. [Online]. Available: https://www.proedinc.com/Products/13025/ssi4-stuttering-severity-instrument--fourth-edition.aspx?bCategory=ola!flu.

[7] T. Kourkounakis, A. Hajavi, and A. Etemad, "FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning," *arXiv Prepr. arXiv2009.11394*, 2020.

[8] S. Alharbi, M. Hasan, A. J. H. Simons, S. Brumfitt, and P. Green, "Sequence labeling to detect stuttering events in read speech," *Comput. Speech Lang.*, vol. 62, Jul. 2020, doi: 10.1016/j.csl.2019.101052.

[9] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2020, vol. 2020-May, pp. 6089–6093, doi: 10.1109/ICASSP40776.2020.9053893.

[10] S. C. Pravin and M. Palanivelan, “A Hybrid Deep Ensemble for Speech Disfluency Classification,” *Circuits, Syst. Signal Process.*, Feb. 2021, doi: 10.1007/s00034-021-01657-1.
[11] S. Khara, S. Singh, and D. Vir, “A Comparative Study of the Techniques for Feature Extraction and Classification in Stuttering,” in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, Sep. 2018, pp. 887–893, doi: 10.1109/ICICCT.2018.8473099.
[12] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham Apple, “SEP-28K: A DATASET FOR STUTTERING EVENT DETECTION FROM PODCASTS WITH PEOPLE WHO STUTTER.”
[13] N. Bernstein Ratner and B. MacWhinney, “Fluency Bank: A new resource for fluency research and practice,” *J. Fluency Disord.*, vol. 56, pp. 69–80, Jun. 2018, doi: 10.1016/j.jfludis.2018.03.002.
[14] P. Howell and S. Sackin, “AUTOMATIC RECOGNITION OF REPETITIONS PROLONGATIONS IN STUTTERED SPEECH Peter Howell and Stevie Sackin,” *Training*, no. October, 1995.
[15] P. Howell, S. Sackin, and K. Glenn, “Development of a Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children Who Stutter: II. ANN Recognition of Repetitions and Prolongations With Supplied Word Segment Markers Europe PMC Funders Group,” 1997.
[16] A. Czyzewski, A. Kaczmarek, and B. Kostek, “Intelligent processing of stuttered speech,” *J. Intell. Inf. Syst.*, vol. 21, no. 2, pp. 143–171, 2003, doi: 10.1023/A:1024710532716.
[17] S. Alharbi, M. Hasan, A. J. H. Simons, S. Brumfitt, and P. Green, “A lightly supervised approach to detect stuttering in children’s speech,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018-Septe, pp. 3433–3437, doi: 10.21437/Interspeech.2018-2155.
[18] E. Smołka, W. Kuniszyk-Jóźkowiak, M. Wiśniewski, and W. Suszyński, “Automatic detection of prolonged fricative phonemes with the Hidden Markov Models approach,” *J. Med. Informatics {\&} Technol.*, vol. 11, pp. 293–297, 2007.
[19] T. S. Tan, Helbin-Liboh, A. K. Ariff, C. M. Ting, and S. H. Salleh, “Application of Malay speech technology in Malay speech therapy assistance tools,” *2007 Int. Conf. Intell. Adv. Syst. ICIAS 2007*, pp. 330–334, 2007, doi: 10.1109/ICIAS.2007.4658401.
[20] A. Dash, N. Subramani, T. Manjunath, V. Yaragarala, and S. Tripathi, “Speech Recognition and Correction of a Stuttered Speech,” in *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, Nov. 2018, pp. 1757–1760, doi: 10.1109/ICACCI.2018.8554455.
[21] K. M. Ravikumar, R. Rajagopal, and H. C. Nagaraj, “An Approach for Objective Assessment of Stuttered Speech Using MFCC Features.”
[22] M. Hariharan, L. S. Chee, O. C. Ai, and S. Yaacob, “Classification of speech dysfluencies using LPC based parameterization techniques,” *J. Med. Syst.*, vol. 36, no. 3, pp. 1821–1830, Jun. 2012, doi: 10.1007/s10916-010-9641-6.
[23] J. Pálfy, “Analysis of Dysfluencies by Computational Intelligence,” 2014.
[24] V. Zayats, M. Ostendorf, and H. Hajishirzi, “Disfluency detection using a bidirectional LSTM,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2523–2527, 2016, doi: 10.21437/Interspeech.2016-1247.
[25] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” *J. Data*

Warehous., vol. 5, no. 4, pp. 13–22, 2000.
[26] R. Hayhow, A. M. Cray, and P. Enderby, “Stammering and therapy views of people who stammer,” *J. Fluency Disord.*, vol. 27, no. 1, pp. 1–17, 2002, doi: 10.1016/S0094-730X(01)00102-4.
[27] A. R. Craig and P. Calver, “Following up on treated stutterers: Studies of perceptions of fluency and job status,” *J. Speech Hear. Res.*, vol. 34, no. 2, pp. 279–284, 1991, doi: 10.1044/jshr.3402.279.
[28] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jan. 2012.
[29] A. S. AlAgha, H. Faris, B. H. Hammo, and A. M. Al-Zoubi, “Identifying β -thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine,” *Artif. Intell. Med.*, vol. 88, pp. 70–83, Jun. 2018, doi: 10.1016/j.artmed.2018.04.009.
[30] W. Hadi, N. El-Khalili, M. AlNashashibi, G. Issa, and A. A. AlBanna, “Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan,” *Comput. Biol. Med.*, vol. 114, p. 103474, Nov. 2019, doi: 10.1016/j.compbiomed.2019.103474.
[31] B. A. Al-Qatab and M. B. Mustafa, “Classification of Dysarthric Speech According to the Severity of Impairment: An Analysis of Acoustic Features,” *IEEE Access*, vol. 9, pp. 18183–18194, 2021, doi: 10.1109/ACCESS.2021.3053335.