Systematic Review of Machine Learning Approaches for Detecting Developmental Stuttering

Liam Barrett, *Member, IEEE,* Junchao Hu, Peter Howell

***Abstract*— A systematic review of the literature on statistical and machine learning schemes for identifying symptoms of developmental stuttering from audio recordings is reported. Twenty-seven papers met the quality standards that were set. Comparison of results across studies was not possible because training and testing data, model architecture and feature inputs varied across studies. The limitations that were identified for comparison across studies included: no indication of application for the work, data were selected for training and testing models in ways that could lead to biases, studies used different datasets and attempted to locate different symptom types, feature inputs were reported in different ways and there was no standard way of reporting performance statistics. Recommendations were made about how these problems can be addressed in future work on this topic.**

***Index Terms*— developmental stuttering, automatic speech recognition, machine learning, Vapnik–Chervonenkis dimension, language diversity.**

I. INTRODUCTION

Conservative estimates suggest that at least 5% of the population will be affected by developmental stuttering at some point in their life [1]. Many of the cases begin in childhood but approximately 80% of these recover by teenage. Whilst speech has to be affected during childhood for stuttering to be diagnosed, some adults who stuttered in early life no longer have speech symptoms but continue to report anxiety (covert stuttering). Automatic speech recognition (ASR) and machine learning (ML) procedures applied to the speech of people who stutter (PWS) could detect incidence of speech and anxiety for clinical and other purposes. They could provide objective and reliable biomarkers for PWS, healthcare professionals and researchers about status of stuttering at a given time or changes that occur over time (whether these happen spontaneously or as a result of interventions). The procedures would permit extensive screening for the speech symptoms indicative of stuttering for all children starting school [2][3], thereby allowing early referral of suspected cases of stuttering for speech therapy [4][5]. At present, prosthetic devices that improve the fluency of PWS make continuous changes to speech. Work on stuttering has called for biofeedback systems to be delivered just at moments of stuttering [6][7], and we suggest that biofeedback could be controlled by an ASR procedure that detects stuttering. Such targeted feedback would make the alterations delivered unobtrusive and could potentially lead to sustained improvements in fluency [8]. Availability of ASR schemes that address stuttering could make online platforms designed to recognize fluent speech more accessible to PWS.

Studies into auto-detection of stuttering began in 1995 [9]. Subsequently, a significant stimulus was the provision of freely-accessible online audio data with transcriptions and associated software [10]. A significant number of articles have published details of ASR and ML schemes for recognizing stuttering from audio recordings to warrant a systematic review. This systematic review focuses symptoms that affect single words/syllables rather than those that affect supralexical language units [11]. Supralexical dysfluencies were excluded as they are not currently regarded as discrete symptoms of stuttering [1][12]. ASR procedures have included video, as well as audio records [13][14]. Studies using video data were excluded in this review because required performance statistics were not reported and ethics and data protection procedures do not allow use of video recordings in some countries.

The systematic review was conducted according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) recommendations [15] and the Meta-Analysis Reporting Standards [16]. A full review of the ML methods generally is beyond the scope of the current article, see [17] for full description. As such, it is assumed that readers have appropriate background in machine learning as a full review of the ML methods is beyond the scope of the current article.

In this systematic review, searches were made for studies on automatic detection of stutters that used either statistical or ML approaches applied to speech data from PWS. Intentions were to determine best practice for preparing data, document the range of approaches adopted and their performance estimates (accuracy). Recommendations for reporting studies are made that will allow results to be replicated and to permit performance to be compared across studies

L. Barrett is a PhD student in the Department of Experimental Psychology at University College London: (e-mail: l.barrett.16@ucl.ac.uk).

J. Hu is a former master's student and currently research associate, both posts in the Department of Experimental Psychology at University College London: (e-mail: junchao.hu.19@ucl.ac.uk).

P. Howell (corresponding author) is Professor of Experimental Psychology in the Department of Experimental Psychology at University College London: (e-mail: p.howell@ucl.ac.uk).

II. Methods

Methods of data extraction, screening and analysis were published in advance, and research questions were formulated before the data extraction for systematic review began [15][16]. The project plan, search terms, exclusion criteria, search results and exclusion decisions were published on the Open Science Framework [18]. The meta-data from articles selected were used to examine the types and sources of available models and to index the performance of different modeling approaches. Key concepts were identified and generic limitations in the studies were noted. Essential next steps that were absent in the current evidence-base were documented.

2.1. Search strategy

Medline, Springer, EMBASE, ISI Web of Knowledge/Science and the Institute of Electrical and Electronics Engineers (IEEE) databases were searched for scientific peer-reviewed journal articles, pre-prints, chapters, and conference papers. Reviews and letters to editors were excluded. Google Scholar, GitHub, OpenGrey and OpenDOAR databases were also searched to allow unpublished work to be included. OpenGrey and OpenDOAR also identified 'Gray literature' [19].

Search terms used for examining the title and abstract of articles returned from the databases were: "machine OR deep learning" AND "stutter* OR stammer* OR *fluencies" AND "speech"; "classification OR detection OR recognition" AND "stutter* OR stammer* OR *fluencies" AND "speech". The wildcards in "stutter*" and "stammer*" allowed for "-ing" or "-er" endings. and the one in "*fluencies" allowed "dys-" and "dis" spellings to be returned.

The search term "UCLASS" (University College London Archive of Stuttered Speech) was included when the entire text of articles was examined [20]. UCLASS was the only publicly available archive of stuttered speech which has stutters transcribed and aligned against the audio signal as required when training models at the time the review was conducted[1].

2.2. Study selection

The search results were automatically screened for duplicates and a two-stage manual screening was undertaken: (1) Title/Abstract; and (2) Full text. J.H. and L.B. (the primary reviewers) independently screened all papers using the eligibility criteria for Title/Abstract screening given below. Disagreements were resolved by the third author (secondary reviewer) who was blind to the primary reviewers' decisions.

The full PICOS search strategy was as follows:

P Population was people who stutter.

I Intervention was the statistical or ML models used to automatically classify stuttered and fluent speech.

C The Comparisons of interest were the speech features, the dataset size and the type of model used for classifying stuttered speech.

O Outcome was model performance. Various metrics were reported including accuracy, precision, recall and Area Under the Curve (AUC). Studies had to report at least one of these metrics for the model to be forwarded to the next stage of the review.

S The Study (model) design was supervised machine learning models classifying stuttered speech from the audio signal.

The inclusion criteria applied to titles and abstracts of all publications identified were that: (1) a form of automatic learning model was used; (2) the model was trained on a dataset of speech from PWS; (3) Journal articles, chapters, conference papers and pre-prints were considered; (4) Studies were published up to May 2nd, 2021. No restrictions were placed on the language that the article was written in nor the language for the speech data itself. The exclusion criteria at this stage were: (1) The study was not peer-reviewed; (2) The study used non-human data (e.g. synthetic speech); (3) The study did not report results for a statistical or machine learning model; (4) The model was not built on, nor worked with, speech data from PWS; (4) The model did not address recognition of stuttered speech; (5) Models only used speech data from people who did not stutter. There were no constraints on the speech features used by the models.

Articles that passed the title/abstract review then underwent full-text screening. As well as meeting the previous criteria, studies had to comply with the following inclusion criteria: (1) Only supervised-learning models were considered; (2) Models had to provide the sample size of the training and test data sets. However, no constraints were set on the size of the sample nor its reported type (these varied in terms of length of recordings used, and the number of observations for stuttered versus fluent events). Studies were excluded if: (1) they were about neurogenic stuttering; (2) training and test datasets were not independent; (3) No "accuracy" metric was reported; (4) Sample size was not specified.

References in the selected paper were examined for any qualifying papers that had been missed during search-term screening. Twelve additional studies were identified at this stage. These were evaluated according to the two-stage screening procedure (title/abstract and full text review).

2.3. Data extraction

Full-text data were extracted from articles that passed title/abstract screening. These data reported accuracy and classification outcomes; type of signal pre-processing and feature extraction; model architecture, model training protocol, dataset size, participant demographics, dataset splits between training and test data, validation procedures applied. The following meta-data were also recorded on the data extraction form; full reference, country where the study was conducted, type of publication (journal article, conference proceeding or

[1] There are three additional databases for stuttered speech that future research could use: FluencyBank [21], Sep-28k [22] and LibriStutter [23]. FluencyBank was not included as a search term as, this open-access database does not have transcriptions timestamps that locate stutters. Thus, researchers would need to perform transcriptions themselves. Sep-28k was not included as it was released after this study commenced. LibriStutter was not used as, the stutters are simulated whereas, in the studies reviewed, the models classify actual stuttering from PWS. See section "5.1. Transparency about data used" for a discussion about the importance of stuttering definitions.

book chapter) and first author's affiliation. The full data extraction form is available at https://osf.io/ctzjk/.

When studies reported on more than one model, all data from the models were extracted but only the one with best performance was included in the quantitative analysis. For studies that reported outcome accuracies for separate stutter types (e.g., repetitions, prolongations, blocks), specific inputs to the model, model types and dataset size were extracted separately.

If data were missing from any papers, a request was made to authors to provide the missing data. If the requested data were not received, the paper was excluded as it lacked at least one inclusion criterion. The full-text data extracted independently by the first and second authors were compared and any residual disagreements were resolved by the secondary reviewer (third author).

2.4. Data extraction procedure

Data extraction was performed on Covidence (https://www.covidence.org) using a customized template. Author J. H. rechecked the final data form and completed consensus checks (i.e., rechecked the most appropriate response). Data quality was not formally assessed [24][25].

III. Results

4.1. Resultant dataset

Of the 1,372 results returned using the original search terms, 443 were duplicates leaving 929 papers to consider (Figure 1). The databases from which the papers were retrieved are indicated on the left of Figure 1. The two primary reviewers disagreed about 42 papers at the title/abstract stage. These were passed on to the secondary reviewer, who made final decisions. After title/abstract screening, 58 papers passed the inclusion criteria and were forwarded for full-text review. Kappa, a measure of inter-rater reliability, for the primary reviewers at this stage was 0.65, which is considered "substantial" [26].

The 58 selected papers were subjected to full text review. The primary reviewers disagreed about seven papers. Altogether, 31 paper were excluded. The inter-rater reliability between the primary reviewers at this stage (k = 0.76) was "substantial" [26]. The 27 papers that passed both review stages were included in the systematic review. Figure 2 summarizes the data selection and extraction procedure, shows points at which exclusions arose and the number of papers involved at these points. Table 1 provides details about each of the 27 papers including in-text citation, publication type, date of publication etc.

4.2. Sources of evidence

Table 1 gives a breakdown of data from the 27 studies. Thirteen of the studies that passed review, were peer-reviewed journal articles and book chapters, 12 were from conferences or proceedings and two were from 'Gray Literature' sources (7%). In Table 1 and the text, the 27 studies are indicated by Sn where n = 1 to 27 for ease of referral. There was a clear increase in published models over the 26-year period beginning with S1. Papers published in the last five, and ten, years were 9 (33%) and 19 (70%) respectively. Most studies were conducted in India (9), followed by Malaysia (6) and Poland (5). Languages used in samples were English (19 studies[S1, S2, S5-7, S9-16, S18, S20, S24-27]), Polish (5 studies[S3, S4, S8, S17, S19]), Hindi (one study[S21]), English and Hindi (one study[S22]) and one study[S23] did not state explicitly what language was used.

4.3. Sample details

Fourteen studies[S6, S9-16, S18, S20, S24, S25, S27] employed the publicly-available UCLASS dataset exclusively and another one[S22] from the All India Institute of Speech and Hearing, AIISH, used a private dataset and UCLASS (56% overall used UCLASS). One study [S26] used two other publicly available datasets (FluencyBank) and the recently released SEP-28 data set. Eleven studies[S1-5, S7, S8, S17, S19, S21, S23] used data that the researchers collected themselves. Use of private data sets alone prohibits comparison across models, and publicly-available datasets need to document reliability and validity of procedures used to label stuttering events. Upper and lower ages of the speakers were reported in 18 studies[S6, S8-15, S17-24, S27], three studies[S1, S2, S7] referred to participants as 'Adults' or 'Children', five studies[S3, S4, S16, S25, S26] did not report age and one study[S5] only reported mean age of the entire cohort.

All the studies reported the amount of speech data that was used to build the models. However, there was no consensus about how speech-sample size was reported for ASR of stuttered speech. Here the studies were classified into four descriptors of data size (Figure 3): S - number of speech samples, O - number of observations, C - number of recordings, and L - length of recordings. Note that the number of recordings and/or the length of the recordings provide no information about the actual number of stuttered events that the model was given. Consequently, the model's accuracy cannot be weighted by the amount of data the model was trained on. Number of speech samples provides a better indication in this respect since the frequency of each class along with its accuracy can be reported. For example, S21 provided the model with 28 fluent and 50 dysfluent samples of speech. Even so, this is not fully informative because what were actually provided as inputs to the model were features extracted from these samples, such as Mel-frequency cepstral coefficients, MFCCs. Although the window length (25 ms) and step size (10 ms) for feature extraction are provided[S21], the feature calculation and the results derived depend on multiple factors such as the computational methods (e.g., padding, smoothing) and the programing environment (e.g., Python/MATLAB/R libraries). Hence, it is not clear what the actual number of data observations that were input to the model was.

Fig. 1. PRISMA flow diagram of databases queried (EMBASE, MEDLINER, IEEE, Web of Science and Springer), the total results (1,372) and unique results (929) retrieved from the systematic search. The pie charts at right, show the split by database (original at top and after title/abstract screening at the bottom).

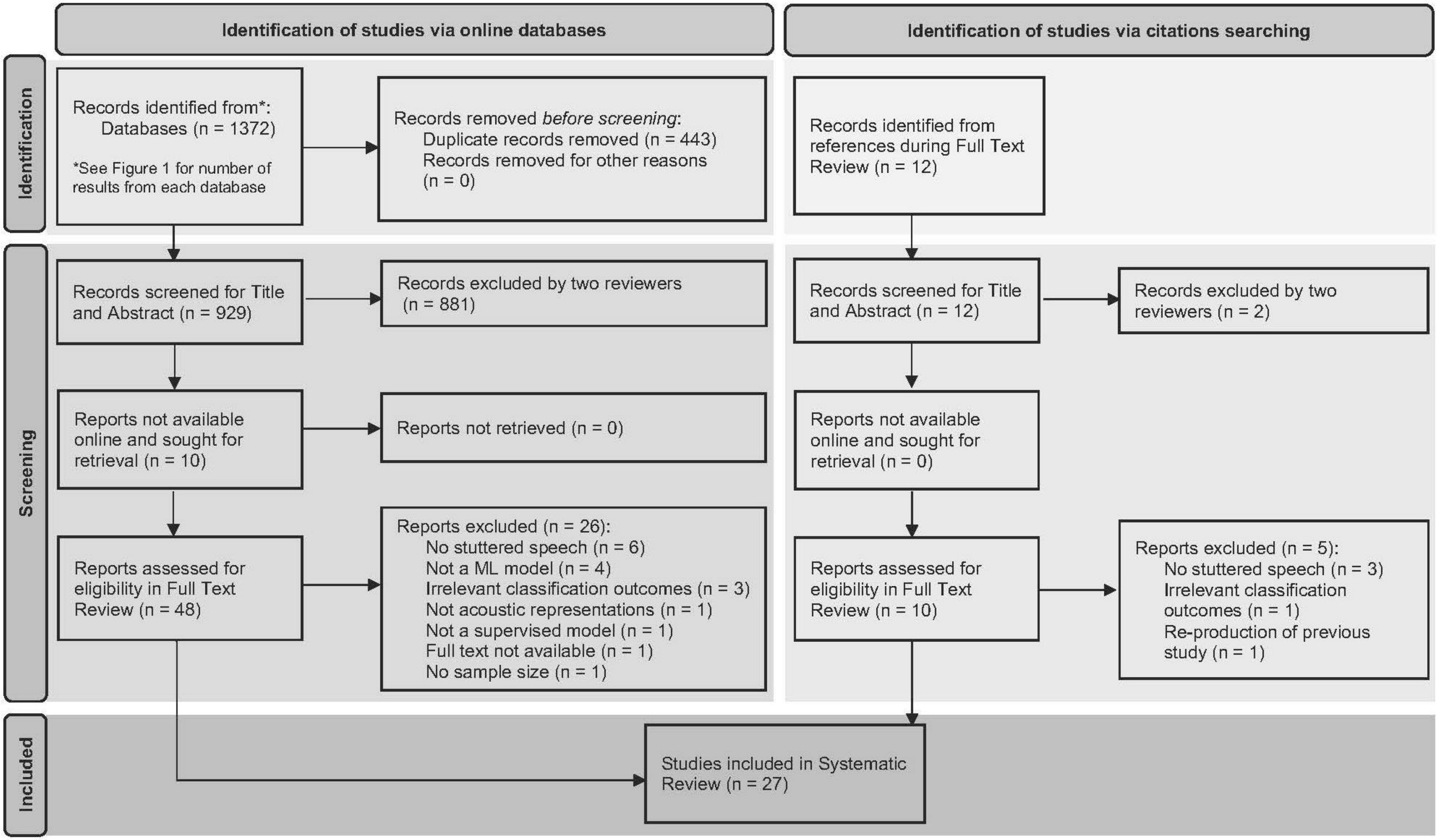

Identification of studies via online databases
Identification of studies via citations searching
Identification
Screening
Included
Records identified from*:
Databases (n = 1372)
*See Figure 1 for number of results from each database
Records removed before screening:
Duplicate records removed (n = 443)
Records removed for other reasons (n = 0)
Records identified from references during Full Text Review (n = 12)
Records screened for Title and Abstract (n = 929)
Records excluded by two reviewers (n = 881)
Records screened for Title and Abstract (n = 12)
Records excluded by two reviewers (n = 2)
Reports not available online and sought for retrieval (n = 10)
Reports not retrieved (n = 0)
Reports not available online and sought for retrieval (n = 0)
Reports assessed for eligibility in Full Text Review (n = 48)
Reports excluded (n = 26):
No stuttered speech (n = 6)
Not a ML model (n = 4)
Irrelevant classification outcomes (n = 3)
Not acoustic representations (n = 1)
Not a supervised model (n = 1)
Full text not available (n = 1)
No sample size (n = 1)
Reports assessed for eligibility in Full Text Review (n = 10)
Reports excluded (n = 5):
No stuttered speech (n = 3)
Irrelevant classification outcomes (n = 1)
Re-production of previous study (n = 1)
Studies included in Systematic Review (n = 27)

Fig. 2. PRISMA flow diagram of the systematic review process from initial search results from the databases as well as results from in-text reference searching, to screening processes at the Title and Abstract and Full-text stages through to the resultant dataset yielded by the systematic process (N = 27).

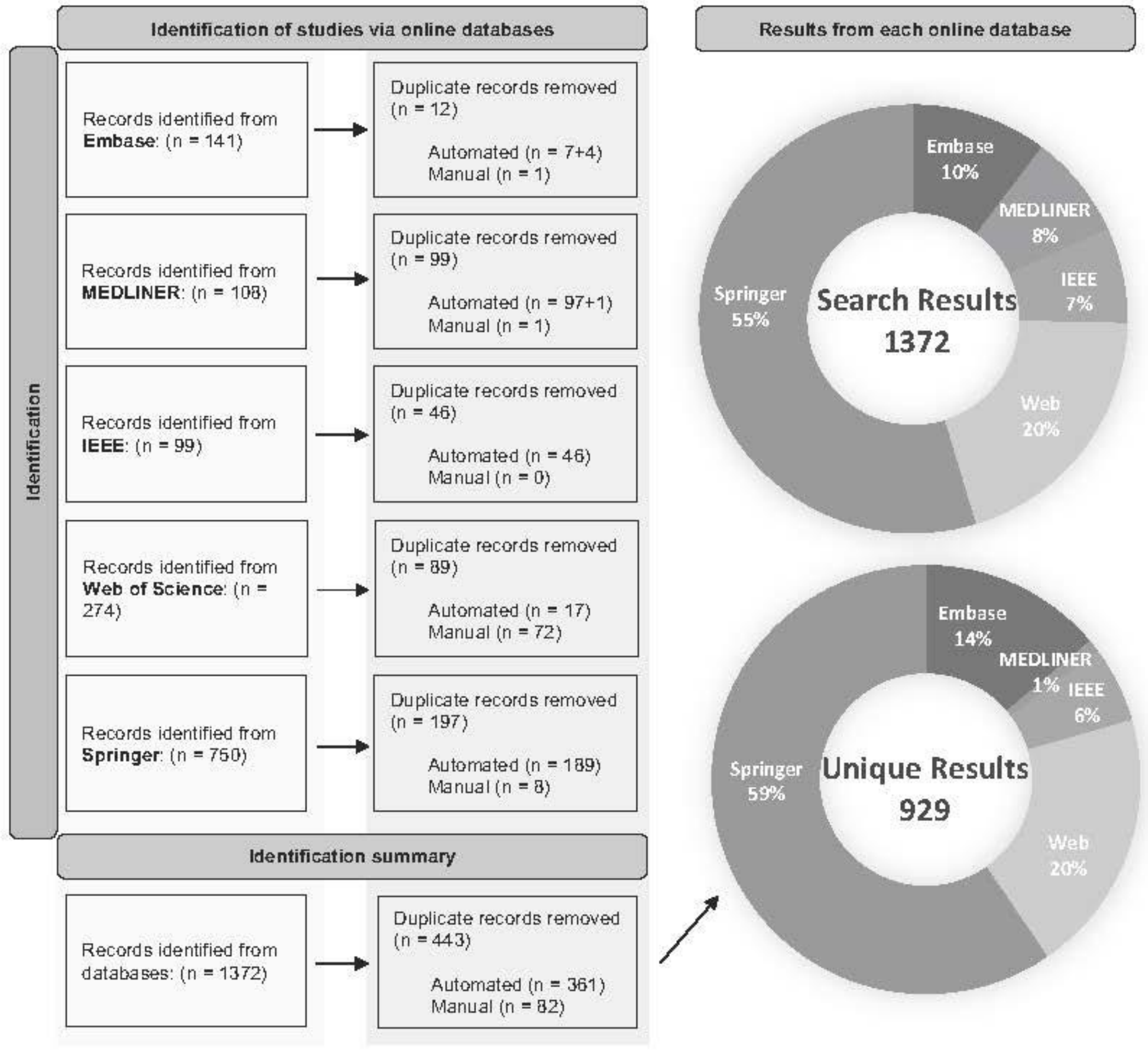

Identification of studies via online databases
Identification
Records identified from Embase: (n = 141)
Duplicate records removed (n = 12)
Automated (n = 7+4)
Manual (n = 1)
Records identified from MEDLINER: (n = 108)
Duplicate records removed (n = 99)
Automated (n = 97+1)
Manual (n = 1)
Records identified from IEEE: (n = 99)
Duplicate records removed (n = 46)
Automated (n = 46)
Manual (n = 0)
Records identified from Web of Science: (n = 274)
Duplicate records removed (n = 89)
Automated (n = 17)
Manual (n = 72)
Records identified from Springer: (n = 750)
Duplicate records removed (n = 197)
Automated (n = 189)
Manual (n = 8)
Identification summary
Records identified from databases: (n = 1372)
Duplicate records removed (n = 443)
Automated (n = 361)
Manual (n = 82)
Results from each online database
Search Results
1372
Embase
10%
MEDLINER
8%
IEEE
7%
Web
20%
Springer
55%
Unique Results
929
Embase
14%
MEDLINER
1%
IEEE
6%
Web
20%
Springer
59%

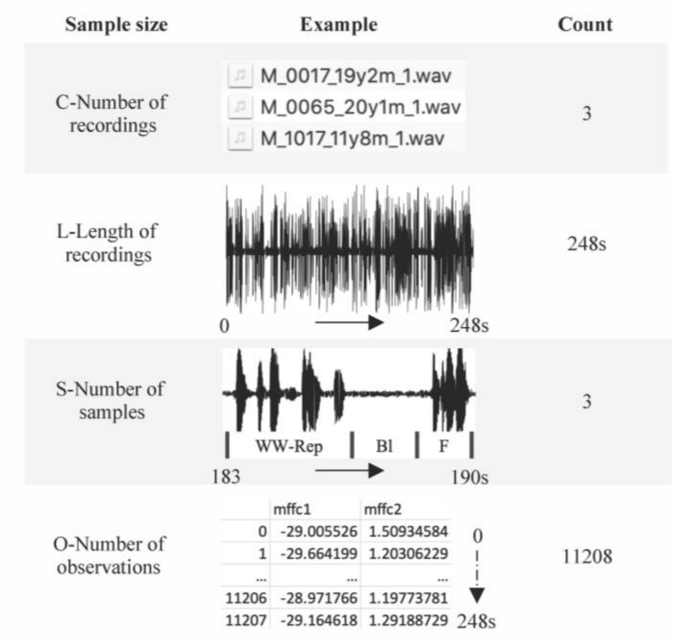
Table 1. Model- and Meta-data for all papers included in systematic review in chronological order.

ID[1]	Source[2]	Classification Outcomes[3]	Sample size[4]	Features[5]	Model Design[6]	Accuracy %
S1 Howell & Sackin, 1995 [9]	C/P	R, P & O	L: 2-min	20 Auto-correlation coefficients, 19 vocoder coefficients & the signal envelope	ANN	82
S2 Howell et al., 1995 [27]	PJ	R, P, F & DF	C: 12	32 features including whole word and part word duration, number of energy peaks, mean duration and SD	ANN	92
S3 Czyzewski et al., 2003 [28]	PJ	SG, P, R & F	S: 104	Cepstral coefficients, amplitude, frequency from F0-F3	Rough Set & Neural Network	Rough set (SG 91.67, P 97.22, R 96.67)
S4 Wisniewski et al., 2007 [29]	PB	P & R	S: 24	20 MFCCs & a distance metric based of these coefficients	HMM	70
S5 Ravikumar et al., 2008 [30]	G	F & DF	S: 251F, 67DF	12 MFCCs	ANN	83
S6 Chee et al., 2009 [31]	C/P	P & R	S: 110	MFCCs	kNN, LDA	kNN & LDA 90.91
S7 Ravikumar et al., 2009 [32]	G	F & DF	S: 251F, 67DF	12 MFCCs	ANN, SVM	SVM 98.3
S8 Swietlicka et al., 2009 [33]	PB	F & DF	S: 59F, 59DF	Winning Neuron of the Self-Organizing Map at each time step	ANN (MLP, RBF)	MLP 92
S9 Pálfy, 2011 [34]	C/P	R & F	S: 16F, 40DF	22 MFCCs	SVM	SVM 98
S10 Pálfy & Pospíchal, 2011 [35]	C/P	R & F	S: 16F, 40DF	22 MFCCs	SVM, ANN (Elman, MLP)	ANN-MLP 96.02, SVM with sigmoid kernel 99.05
S11 Ai et al., 2012 [36]	PJ	P & R	C: 39	21 LPCCs, 25 MFCCs	kNN, LDA	kNN 92.75
S12 Hariharan et al., 2012a [37]	C/P	P & R	C: 39	Entropy	LS-SVM	SVM 96.96
S13 Hariharan et al., 2012b [38]	PJ	P & R	C: 39	LPC, LPCC and WLPCC	kNN, LDA	kNN 98.24
S14 Fook et al., 2013 [39]	PJ	P & R	S: 171	MFCC, LPC, LPCC, wMPCC & PLP	kNN, LDA, SVM	SVM 96.20
S15 Hariharan et al., 2013 [40]	PJ	P & R	S: 171	DWT decomposition & Sample entropy	kNN, LDA, SVM	SVM 96.37
S16 Pálfy & Pospíchal, 2013 [41]	C/P	P-Rep & F	C: 12	MFCCs, short-time energy, SAX	SVM	97.6
S17 Swietlicka et al., 2013 [42]	PJ	Bl, S-Rep & P	S: 153F, 153DF	Winning Neuron of the Self-Organizing Map at each time step	ANN (MLP)	Bl 96 S-Rep 84 P 99

S18 Mahesha et al., 2015 [43]	PJ	S-Rep, WW-Rep & P	S: 200	39 features: 12 MFCCs, Energy, and their 1st and 2nd derivatives	GMM, GMM-SVM	GMM-SVM 98.24
S19 Kobus et al., 2016 [44]	C/P	S-Rep & F	S: 145F, 117DF	15 LPCs, Formants (not specified)	Average k-means distance	89
S20 Mahesha & Vinod 2016 [45]	PJ	S-Rep, WW-Rep, I & P	S: 200	12 MFCCs, 12 delta-MFCCs, 12 delta-delta-MFCCs & Average spectral energy	GMM	96.43
S21 Savin et al., 2016 [46]	C/P	R, P & F	S: 28F, 50DF	13 MFCCs, F0 and 3 Formants (F1, F2, F3), Maximum, Minimum, Mean & Variance of pitch, ZCR and Energy	ANN	88.29
S22 Mahesha & Vinod 2017 [47]	C/P	R, P & I	S: 750	39 LH-MFCCs	GMM	94.98
S23 Manjula et al., 2019 [48]	C/P	R, P & Bl	S: 92	MFCC	AOANN	R 92, P 94, Bl 92
S24 Gupta et al., 2020 [49]	PJ	F, P, S-Rep, WW-Rep & P_Rep	C: 20	14 WMFCCs (fusion of MFCC, delta and delta-delta)	ANN (Bi-LTSM)	96.67
S25 Kourkounakis et al., 2020 [50]	C/P	S-Rep, WW-Rep, P-Rep, Rev, I & P	S: 800	Spectrograms	ANN (Bi-LSTM)	91.15
S26 Lea et al., 2021 [22]	C/P	Bl, P, S-Rep, Phrase-Rep & I	O: 32321	40 MFCCs, pitch, delta-pitch, voicing-features, 8 articulatory features of vocal tract constriction variables & 41 phoneme probabilities	ANN (LSTM)	83.6
S27 Mishra et al., 2021 [51]	PJ	F & DF	O: 17545	39 MFCCs, Root Mean Square	ANN	86.67

Column content and abbreviations as follows: **1)** Study ID with numbers, Sn, where n = 1 to 27 in chronological order, along with the study's text and number citation. **2)** Reference type, C/P = Conference/Proceeding, PJ = Peer-reviewed journal, PB = Peer-reviewed book chapter, G = Gray literature. **3)** Symptoms recognized, R = Repetition (length unspecified), P = Prolongation, O = Other (undefined), F = Fluent, DF = Dysfluent, SG = Stop Gaps, P-Rep = Phrase repetition, S-Rep = Syllable Repetition, PW-Rep = Part-Word Repetition, WW-Rep = Whole Word Repetition, I = Interjection, Bl = Block, Rev = Revision. **4)** Sample size, L = Length of recordings, C = Number of recordings, S = Number of speech samples (F = Fluent, DF = Dysfluent), O = Number of observations. **5)** Features extracted, MFCC = Mel-Frequency Cepstral Coefficient, WMFCC = Weighted MFCC, LH-MFCC = Linear Prediction-Hilbert transform based MFCC, LPC = Linear Prediction Coefficient, LPCC = Linear Prediction Cepstral Co-efficient, F0 = Fundamental Frequency, Fn = Formant Frequency (value of n indicates which formant/s), PLP = Perceptual Linear Predictions, DWT = discrete wavelet transform **6)** Model type, ANN = Artificial Neural Network, AOANN = Adaptive Optimization based Artificial Neural Network, LSTM = long short-term memory, MLP = multi-layer Perceptron, SVM = support vector machine, HMM = Hidden Markov Model, GMM = Gaussian Mixture Model, k-NN = k-Nearest Neighbor, LDA = Linear Discriminant Analysis, SAX = Symbolic Aggregate Approximation, RBF = Radial-basis Function.

Fig. 3. Depiction of different ways or reporting sample size categories. Row one specifies three speech recordings from UCLASS. Each file contains continuous speech from one PWS and the length of these files varies. Row two gives the entire waveform of one speech recording (length specified in seconds). Note that the recording-length unit can vary between studies. Row three shows a 7-second audio snippet, containing three labelled speech samples: whole-word repetition (WW-Rep), block (Bl), and fluent speech (F). Depending on the segmentation and annotation procedure, length of the speech samples varied within, and between, studies. Row four summarizes a 2x11208 speech feature matrix that could be extracted from the 248-s sample in column two. The number in the first column is the frame number (0 – 11207) and rows two and three give values of two MFCC coefficients for that frame. The size of such matrices depends not only on how many features were computed and the length of the speech sample (248s), but also the frame size (25ms) and step size (10ms) used during feature extraction, and whether silence was removed from the speech recording.



Of the 27 studies, 18 gave the number of speech samples used to train their models, and eight studies[S5, S7-10, S17, S19, S21] indicated the number of fluent samples (e.g., S3 reported 104 samples overall whereas S5 gave the number of fluent and dysfluent samples as 251 and 67 respectively). On average, the studies had 179 dysfluent samples (SD = 224, N = 18), irrespective of dysfluency type (repetition, prolongation, block etc.). Of those reporting the differences between dysfluent/fluent speech, the average number of samples was 188 (SD = 122, N = 8) with 74 (SD = 40, N = 8) dysfluent and 115 (SD = 100, N = 8) fluent samples. Only two recent studies[S26, S27] gave the number of observations, i.e., the actual length of data matrix for model building. However, S26 and S27 did not give a breakdown of the observations per classification outcome, therefore it is unclear exactly how many fluent and dysfluent samples were used to train their models. For the remaining seven studies, six[S2, S11-13, S16, S24] reported the number of recordings and one[S1] gave the length of recording. Full details of sample size in each study are summarized in the project's online deposit (https://osf.io/ctzjk/).

Note there were also marked differences within each reported data type. For example, S3 and S6 had roughly the same number of samples (104 and 110, respectively), but it was not clear how many actual datapoints or observations each sample contained, which prevented comparison. On the other hand, studies S26 and S27, that used number of observations, had a better standard of reporting as they made the length of the data matrix clear, allowing the study to be replicated. Seven studies[S1, S2, S11-13, S16, S24], only reported the number, or length, of speech recordings.

Although it is difficult to gauge the effect of dataset size from cross-study comparisons, one study[S26] varied the amount of data input to the model and reported the effects on accuracy. S26 trained four Convolution Long Short-Term Memory, LSTM, on 5,000, 10,000, 20,000 and 28,000 observations. A clear increase in F1 score (the harmonic mean of the precision and recall of the model) on the test set was reported as the number of observations increased ($F1_{5k} = 71.7$; $F1_{10k} = 72.2$; $F1_{20k} = 73.4$; $F1_{28k} = 75.8$). This not only indicated that performance continued to increase with training set size but also that the models had not reached their maximum performance.

4.4. Machine leaning models

Studies[S3, S6, S7, S10, S11, S13-15, S18] reported more than one model design. The most popular model design was artificial neural network (ANN), which was used to classify stuttered speech in 13 studies[S1-2, S3, S5, S8, S10, S17, S21, S23-27]. Architectures used included Multi-layer perceptron (three studies[S8, S10, S17]) and LSTM (three studies[S24-26]).

Support vector machines (SVMs) were used in eight studies[S7, S9-10, S12, S14, S15-16, S18], k-Nearest Neighbors, (k-NN) were used in five studies[S6, S11, S13-15] Linear Discriminant Analysis (LDA) was used in five studies[S6, S11, S13-15] and Gaussian Mixture Models (GMM) were used in three studies[S18, S20, S22]. Models that were used in single studies included a Hidden Markov Model[S4], a k-Means Distance model[S19] and, a Rough Set model[S3].

Considering the differences in accuracy with respect to ML model design where there is more than one study, on average SVMs performed best with an average accuracy of 97.63% (SD = 0.9736; N = 8[S7, S9-10, S12, S14, S15-16, S18]). K-Nearest Neighbor models performed poorer on average with an accuracy of 93.12% (SD = 4.055; N = 5[S6, S11, S13-15]), Linear Discriminant Analysis models had an average accuracy of 92.15% (SD = 3.387; N = 5[S6, S11, S13-15]) and Gaussian Mixture Models performed on average at 95.07% (SD = 1.321, N = 3[S18, S20, S22]). Artificial Neural Networks had an accuracy of 89.12% (SD = 6.032; N = 13[S1-2, S3, S5, S8, S10, S17, S21, S23-27]).

Considering the lesser studied models with only one publication per model; k-Mean Distance Modelling had an accuracy of 89.00%[S19], Rough Set Modelling reported 84.63% accuracy[S3] and a Hidden Markov Model had 70% accuracy[S4].

It may be surprising that on average Neural Networks perform so poorly compared to other types of ML models given that of the nine papers published since 2016, six used a form of neural network to model stuttering [S21, S23-27]. This is not due to earlier studies that used ANNs pulling the average performance down since, when only the most recent neural net studies were considered, the average accuracy remained almost the same (89.84%, SD = 4.642; N = 6 [S21, S23-27]). However, the aims of the ML model must also be considered. Looking at two ML models of the same design that were trained on the same data, one[S27] attempted to classify dysfluent versus fluent speech from the audio signal whereas the other[S21] attempted to classify two different types of dysfluent speech as well as fluent speech (e.g., prolongation vs. repetition vs. fluent Speech). Since an extended number of outcomes was considered in the second model, its task-complexity was greater. This arises because the two models were set different tasks, despite the design, dataset and all other aspects being equivalent, it cannot be said that one is better than the other. It is necessary to consider how accuracy changes as a function of the various outcomes the models are tasked to identify. The field currently does not have an agreed baseline whose metrics could be used to determine whether performance of new models improves significantly.

Furthermore, the models all varied on the feature inputs supplied to the models. Some models[S12, S15] were trained on temporal features of the audio signal, others[S5-7] were trained on spectral features and others still[S20, S24] used meta-features such as delta-derivatives of the spectral features. Since performance depends on both model architecture and the feature inputs, their impacts cannot be separated. Indeed, in some instances the actual architecture of a model may vary as a function of the input features. For instance, the number of hyper-planes of an SVM model depends on the number of datapoints (features) in an observation.

4.5. Outcome classes

Table 1 (column labeled ‘Classification Outcomes’) shows that types of stutters identified varied. Out of the 27 studies, seven[S4, S6, S11-15] classified prolongations and repetitions, four[S5, S7, S8, S27] classified fluent and dysfluent speech, four[S9, S10, S16, S19] classified repetitions and fluent speech. The remaining 14 studies all defined different outcome classes.

As discussed in the introduction, the symptoms considered as stutters at supralexical levels are not universally agreed to be stutters. This also applies at the word/syllable level, where some researchers included whole-word (WW) repetitions as a symptom of stuttering [12] whilst others argue that they are a typical dysfluency that occurs in the speech of both PWS *and* fluent speakers [52]. Given the disagreement about whether WW repetitions are or are not stutters, more consideration about what symptoms ASR and ML models recognize is warranted. Twenty-two studies[S1-4, S6, S9-26] explicitly stated that they attempted to classify repetitions from the audio signal, with variation across studies in what was considered ‘repetition’ (syllable/sound, whole word or phrase repetition as summarized in Table 2). Fifteen studies[S1-4, S6, S9-15, S21-S23] simply referred to ‘repetition’ with no indication about the criteria applied for classifying repetitions. Of the seven remaining studies, S17 and S19 considered only syllable or ‘sound’ repetitions; S16 classified phrase repetitions alone; S18 and S20 classified both syllable/sound-repetitions as well as WW repetitions; S26 defined both sound/syllable and phrase repetition for use in the model; S24 and S25 defined and classified all three forms of repetition (syllable/sound-, WW- and phrase-repetitions). Some studies simply addressed ‘dysfluency’ as a class which could include some or all forms of repetition.

Table 2. Studies which have included repetition in their classification scheme split by type/s of repetition.

ID[1]	Syllable	Whole word	Phrase	General
S1[9]				X
S2[27]				X
S3[28]				X
S4[29]				X
S6[31]				X
S9[34]				X
S10[35]				X
S11[36]				X
S12[37]				X
S13[38]				X
S14[39]				X
S15[40]				X
S16[41]			X	
S17[42]	X			
S18[43]	X	X		
S19[44]	X			
S20[45]	X	X		
S21[46]				X
S22[47]				X
S23[48				X
S24[49]	X	X	X	
S25[50]	X	X	X	
S26[22]	X		X	

[1] Study ID with numbers, Sn, where n = 1 to 27 in chronological order, along with the study’s number citation.

When studies used different numbers of all symptom outcome classes, performance was relatively unaffected (Table 3). Hence, the number of classes does not seem to degrade the performance of ML models for classifying dysfluent speech in PWS.

Table 3. Studies split by the number of classes of speech given to the model, the number of studies, the average accuracy and the standard deviation are given in columns 1-4.

Number of Classes	**Number of Studies**	**Average Accuracy (%)**	**Standard Deviation**
2	15	92.43	7.706
3	6	92.23	6.012
4	3	87.78	9.564
5	2	90.14	9.242
6	1	91.15	NA

4.6. Features

A full list of features for each of the studies is given in Table 1, column labeled ‘Features’. Some studies included more than one of the features listed in the caption. Many papers (18) included MFCCs as inputs to the model[S4-7, S9-11, S13-14, S16, S18, S20-

[24, S26-27]. However, the number of MFCCs extracted and their frequency bands varied across studies. This applied to studies that used raw MFCCs or ones that transformed them, such as delta-derivatives and log transforms. The modal number of MFCCs extracted was 12 ([S5, S7, S18, S20]). When considering all studies with MFCCs included as inputs to the ML models, the average accuracy was 91.98% (SD = 7.476) as compared with 90.47% (SD=6.586) for models that did not extract MFCCs.

Eight studies extracted some form of energy feature from the signal such as the signal's envelope[S1-3, S16, S18, S21], the Root Mean Square Energy[S27] or other energy measures[S1-3, S16, S18, S21, S27]. Studies that used energy of the audio signal had an average performance accuracy of 89.92% (SD = 7.470).

Three studies extracted Linear Prediction Coefficients (LPCs)[S13-14, S19] and three extracted Linear Prediction Cepstral Coefficients (LPCCs)[S11, S13-14]. These studies had an average accuracy of 94.39% (SD = 4.766) and 94.73% (SD = 4.202) respectively.

Finally, two papers extracted formant frequencies[S19, 21] and two extracted entropy[S12, S15]. The average accuracy for models using formant frequencies was 88.65% (SD = 0.502) and for entropy, 96.82% (SD = 0.205). Again, on the basis of these pooled averages, one might be inclined to conclude that models performed better when entropy was an input. However, many exogenous variables differed between studies that make this claim violable.

4.7. Optimization and Validation

Most studies (70%) gave the exact training/testing split of their data, either in percentages[S2, S6, S7, S9-13, S16, S18, S20-22, S24], number of samples[S4, S5, S19], number of observations[S26] or number of speakers[S1]. One study[S23] only gave the number of samples for testing. The remaining seven studies[S3, S8, S14, S15, S17, S25, S27] did not report any details on data split.

Twelve studies[S1, S3, S5-7, S9, S10, S12, S14, S15, S22, S25] reported a clear validation split for model assessment. Amongst these, five studies[S6, S12, S14, S15, S22] used 10-fold, one study [S9] 4-fold and one study[S10] 16-fold cross-validation. Two studies[S5, S7] split the test data into two folds for validation, two studies[S3, S25] used leave-one-out test, and one study[S1] tested the model on speech of the same speaker used for training and that of five other speakers unfamiliar to the model. Three other studies[S8, S17, S23] reported data were split over training, test, and validation, but the size of the validation set was unclear. The remaining 12 studies[S2, S4, S11, S13, S16, S18-21, S24, S26, S27] did not describe any validation approach for assessing the model.

Moreover, the studies varied in the way that they optimised the model's performance. Of the 27 studies, only ten[S6, S9, S11-13, S18-20, S23, S24] reported hyper-parameter tuning for their models. Of these three studies[S6, S11, S13] varied the number of neighbors (k) for the k-NN model; two studies varied the gamma [S12] and C [S9] parameter in SVM models; two studies changed the mixture weights[S18] and model order[S20] of the GMM; one study[S19] varied the number of means between clusters; one study[S23] used Artificial Fish Swarm Optimization to tune the neural network; and one study[S24] used sequential grid search to adjust the learning rate, batch size, number of epochs, and number of hidden units in its bi-directional LTSM. Among these, two studies[S11, S13] described feature optimization by varying the frame length, frame overlap, and pre-emphasis used for feature extraction.

IV. Discussion

This review showed that there is a developing body of research into the automatic classification of stuttered speech. One general suggestion is that it would be useful for studies to state explicitly what application/s their recognizers address. Thus, modeling approaches that have significant, but sub-optimal, accuracy, that work rapidly, might be appropriate for real-time applications. However, they would not be suitable for clinical purposes where performance has to be at maximum, but results may not be required immediately.

The review showed that progress has been made into: (1) developing the amount of speech data that is available to researchers; (2) enhancing the architectures of the ML models to address the complexity of the classification problem faced; (3) providing ML models with a range of representative symptoms of stuttered speech; and (4) selecting features in the speech signal that look promising candidates for separating different classes of fluent and dysfluent speech.

There are, however, numerous issues that need to be addressed: The lack of clear definitions of sample size and outcome classes has resulted in models potentially being under-powered or, of more concern for all applications, classifying non-stuttered speech as stuttered. Looking across studies, it was not possible to identify which modeling approach performed best because studies varied in the dataset used, modeling algorithms, features used in the models and because several different outcome measures were used to report 'accuracy'. Three main areas emerged from the systematic review where attention is needed in future work to facilitate comparison of model performance across studies. These are discussed below and realistic recommendations are offered for future work. It might not be realistic, for instance, to recommend that only publicly-available databases are used where groups are developing commercial products.

5.1. Transparency about data used

A. Language diversity

A general problem is that we know very little about how stuttering symptoms vary across languages [53] and even less about how this impacts on ASR of stutters. Whilst ASR studies have used languages other than English, there is usually only a single report for these languages, details such as data set descriptions are often incomplete. Collection of data on additional languages that meet all recommended standards once these are agreed should be encouraged.

B. Labels

Continuous audio data need valid annotation labels irrespective of the type of symptoms models are designed to recognize. Labels can be, minimally, fluent versus dysfluent or involve separate types of symptoms. Nine of the papers[S1-3, S9, S10, S16, S20, S25, S26] provided descriptions of what dysfluencies were considered but had no indication about efforts made to validate the labels provided to the model. The level of fidelity that researchers give about models needs to be improved.

Mislabeling can lead to ML models learning parameters for the wrong category of speech. Incorrect alignment by labelers degrades performance. Similarly, windowing applied to recordings degrades classification: For example, a standard SVM model that has 12 MFCCs as input with a 30ms moving window and 15ms overlap, has overlaps between a fluent utterance and a stuttered utterance at some points that make the classifier's task indeterminate.

Few studies have reported on whether the manual labels that they employ with private datasets are reliable and valid. An exception is the extensive SEP-28 dataset used by S26. However, the manual annotations that SEP-28 provides are limited in several ways: The labels were added manually to 3-second clips by three non-clinician judges. The length of the clips makes the granularity of changes in speech coarse, as speech can change in fluency several times within a 3-second clip. Additionally, the interval selection method is vulnerable to inaccurate and unreliable annotations due to noise levels, type of stutter and position of stuttering within the temporal window [54]. It is not stated how clips were marked as, for example, prolonged (e.g., based on one, two or three of the judges). Reliability and validity of annotators was not at an acceptable standard (Fleiss' Kappa for different symptoms varied between 0.11 and 0.62).

Automatic labeling procedures could speed up labeling. However, this is problematic. The auto-aligners use fluent models and this biases them to work optimally for fluent speech. Thus, if these auto-aligners are used, the labels for stutters are more compromised than those for fluent speech, leading to a perpetual disadvantage for recognizers trained on jeopardized stuttered speech sample labels. In support, [55] showed that automatic labels were not as accurate as manual ones when applied to Huntington's disease patients' speech.

C. Sample size considerations

There is considerable variation concerning how overall sample sizes are reported (Table 1, column five), nor have considerations about power adequacy been discussed. Sample size affects model performance in two ways: (1) If the dataset is small so that there are more features than observations, the model fits the data completely merely by allowing a feature to code for a given observation and its class. This leads to an inflation of the model's perceived efficacy; (2) The model does not have sufficient data to fully '*learn*' the parameters that define a class. This would lead to an under-estimation of the model's ability to classify stuttered and fluent speech. Although there is no absolute rule about the required dataset size for a given problem, learning graphs can provide insight into whether the current models are reaching equilibrium for a given dataset or, if more data would continue to improve performance. The discussion of S26 in the Sample Detail section provides an example where performance had not plateaued over the parameter space explored. There is also the problem that what looks like studies with similar sized datasets often provide insufficient detail to check whether this is the case (e.g., S3 and S6 discussed in the Sample Details section). Recordings were selected from datasets on occasions[S6, S9-16, S18, S20, S24, S25, S27]. Selections made of severe cases, by gender, by age etc. all bias models in particular ways making them less generally applicable.

D. Recommendations

Stutters in datasets need to be annotated at word or syllable levels and audio data should be made available where possible in order that annotation procedures can be checked.

When unpublished, or new, samples are used, all annotated outcomes should be explicitly defined and meet acceptable standards of reliability and validity.

A way of comparing new data sets with standard ones, such as UCLASS, is required. A recommended approach would be to have available a simple ML model, apply it to the standard and new datasets and report the results on all datasets for comparison. New datasets could include ones that use auto alignments. Comparison with the standard would allow reports to be made about whether the new data perform better or worse than the standard.

5.2. Symptoms identified

A. Impact on Classifiers

Models vary in the number of dysfluencies they attempt to classify, as indicated in the 'outcome classes' section and in Tables 2 and 3. Most models aim to distinguish prolonged, repeated and fluent speech whereas others do not distinguish stuttering symptoms (train for binary fluent/stuttered outcomes) as shown in Table 2. Whilst aggregated model performance was similar across studies that identified different numbers of symptoms, any influences of symptom type was masked by differences in data set used, model architecture and features used as inputs. As noted, there are two positions [12][22] concerning whether WW repetitions are or are not stutters. If WW repetitions are stutters, then they are a legitimate target outcome for recognizers. The recognition outputs for studies that target repetitions usually do not distinguish whole-word, from part-word, repetitions. Thus, whilst 23 studies[S1-4, S6, S9-26] targeted repetitions (various classification forms such as repetition, phrase-repetition, syllable-repetition), only two[S24, S25] distinguished WW from part-word repetition. Hence, it is not known whether grouping types of repetitions affects performance adversely. Findings of [56] suggest that whole-word repetition has different brain activity patterns from that which occur for part-word repetitions, prolongations and breaks in words which are symptoms which are always considered to be stutters. If brain activity differences are preserved in the audio signal, then WW repetition and part-word repetition should be recognized as separate outcomes for maximum performance. This test remains to be made.

B. Recommendations

Symptom outcomes should be clearly defined. In particular, reports should be explicit about whether repetitions were sub-divided into WW and part-word types.

Inclusion of different selections of speech symptoms in studies is legitimate, but to allow this, a comprehensive set of annotations needs to be included in the data. This would allow people with different perspectives about symptoms to work with the data and to confirm the status (stuttered or not) of whatever symptom groups they use [1]. This proposal does not restrict investigation to a single set of symptoms.

5.3. Model architecture and features used

A. Train/test split

Care also needs to be taken when reporting how data were split between training and test sets. The size of the split data is frequently under-defined. Often research defines the splits as the number of segments in each category (i.e., 50 cases of prolongation and 50 cases of repetition, e.g., as in S9). But this does not provide an explicit number of *observations* that are input to the model.

A confusion matrix along with the shape of the data matrix (i.e., number of features or columns and number of observations or rows) should be provided. This would allow for full investigation of a model's power given the dataset it was trained on to be determined as well as making it possible to weight a model by its dataset in cross-study comparisons.

B. Model architecture

Although cross-study comparisons of model design are limited because extraneous factors vary between studies, within-study comparison of both ML model type and architecture is possible. For ANNs, S8 investigated the effect of the activation function in the hidden nodes, testing both the Multi-layer perceptron (MLP) and Radial-Basis function (RBF) approaches. S8 reported a marginal improvement for MLP (92% vs. 91% accuracy). However, it was not reported whether this difference between models was significant.

S25 compared a unidirectional LSTM against a bi-directional LSTM and reported an increase in accuracy of 0.09% in preference for the bi-directional LSTM. S25 varied hyper parameters of the bi-directional LSTM including learning rate, batch size and number of epochs and number of hidden units. The model that was set to a batch size of 16,100 epochs and 100 hidden units had the optimal learning rate of 0.01. When learning rate was set to 0.01 and epochs and hidden units to 100, the optimal batch size was 8. When the optimal learning rate and batch size were used with the hidden nodes set to 100, the optimal number of epochs was 50. Finally, when the established optimal parameters were used and the number of hidden nodes was varied, 100 nodes was optimal. By varying these parameters, S25 reported differences in accuracy from ~70% up to 96.67%. Clearly optimization of these parameters led to major changes in the model's ability to learn what separates the various types of speech fluency. Finally, S26 compared a uni-directional LSTM with a convolution LSTM and again found convolution improved performance accuracy by approximately 1.8%. Again, whether this was a *significant* increase or not was not established.

Of the studies that employed SVM models, three compared various intra-model and inter-model performances[S10, S14-15]. S10 compared the effect of changing the SVMs kernel (linear and RBF kernels). Better performance was reported for the linear kernel (98.00%) as compared with the RBF (96.13%). S14 compared performances of an SVM against k-NN and LDA classifiers. SVM outperformed both models in almost all runs when feature extraction parameters were varied. On average the performances of SVMs was approximately 95% compared to the k-NN's 90% and the LDA's 90%. The superiority of the SVM over the k-NN and LDA was supported by a further study from the same group (S15) that reported maximum accuracy of 96.14% for the SVM, 94.39% for the k-NN and 91.87% for the LDA.

Although intra-study model comparisons suggest that there are differences in ability to accurately classify stutters, comparisons across studies are needed to substantiate these claims. Although the currently-available data do not permit this, future research could employ meta-analytic techniques to bridge this knowledge gap (see below).

C. Features used

Currently, the way that performance depends on features used is not apparent across the studies in this review. However, it is possible to make inferences about feature importance when there are intra-study variations of features. The feature inputs to the same model, trained on the same dataset with the same outcome classes, was done in eight studies[S2, S11-15, S21, S24] for various features and parameters.

S2 varied features by inputting various combinations of features to their ANN and determined which set yielded the highest accuracy. Combinations of: (1) Word duration; (2) Word-length fragmentation measures; (3) Word-length spectral measures; (4) Syllable durations; (5) Syllable-length energy; (6) Syllable-length fragmentation measures; and (7) Syllable-length spectral measures were investigated. Using various combinations of these feature groups, a set of feature groups that had between two and five features per set yielded the greatest overall accuracy (92.00%). The researchers also permuted the combinations of feature sets to obtain accuracy for all combinations. Considering only the top 10 ANNs, feature 4 (syllable durations) was present in nine, feature 5 (syllable-length energy) was present in eight, and feature 6 (syllable-length fragmentation measures) was present in six. This suggests that features at the syllabic level (~200ms) tended to perform better than features extracted at the word level, for ANNs at least. The decrease between a model with features 4 and 5 (Accuracy = 91.04) and a model with features 4, 5 and 7 (Accuracy = 90.90) was due to the inclusion of the additional feature (7).

In S2, the researchers varied the feature input size to the model. More can be learned about feature importance by tuning the feature parameters such as window length and time-step. Hariharan's group[S11-15] has investigated extensively feature-parameter tuning with respect to MFCCs, LPCs and LPCCs. Varying the window length between 10ms and 50ms, S11 reported that a k-NN and LDA classifier with 25 MFCCs as input performed optimally with 20ms and 40ms windows, respectively. For a 21-dimensional LPCC, however, the k-NN's optimal performance was with a 30ms window whilst it was 20ms for the LDA classifier. The study also varied the pre-emphasis filter cut-off (α), with the LDA performing optimally for both MFCCs and LPCCs, with a pre-emphasis of 0.93. Whereas the k-NN performed best at α = 0.96 and α = 0.98 for MFCCs and LPCCs, respectively. Finally, the time-step between windows or 'overlap' was varied. For models trained on MFCCs, the optimal overlap was 0 for LDA and 50% for k-NN. For LPCCs, the optimal overlap was 50% for the LDA, and both 33.33% and 75% overlaps performed optimally for the k-NN. By permuting all parameter values and combinations, MFCCs extracted from a 20ms window with a 10ms time-step and an $\alpha = 0.9375$ yielded accuracies of 92.55% for the k-NN and 88.82% for the LDA. In comparison, LPCCs extracted from a window of 30ms and a time-step of 7.5ms and an $\alpha =$

0.98 resulted in a 94.51% accuracy for the k-NN and a 90.00% accuracy for the LDA. S13 also varied these parameters of the LPCCs and used k-NN and LDA ML models. However, for the k-NN the optimal window length was 20ms not 30ms as in S11. Whereas the LDA performed optimally at both 10ms and 30ms window lengths, S11 reported an optimum at 20ms. Clearly then, the optimal window length, and indeed optimal parameters for feature extraction generally, also depend on factors such as dataset size, number and types of classes given to the model, model architecture etc. As yet, no study has attempted to test whether tuning the feature extraction parameters leads to *significantly* better classifiers and how robust this is to changes in other aspects of the ML model.

D. Recommendations

The reporting standards for sample size as well as train-test splits are often unclear and vary significantly between publications. A simple way to resolve this problem is to publish the frequency of observations by classes as in an example from our work shown Table 4. This would allow a range of accuracy measures to be calculated by other researchers.

Table 4. The number and percentage of observations for each class of speech (fluent and stuttered).

Type	**Number of Observations**	**Percentage of total observations**
Fluent	994,912	62.6%
Stuttered	594,779	37.4%
Total	1,589,691	100.0%

Clear documentation of the matrix shapes of the dataset is required.

Given that model architectures and feature inputs vary dramatically across studies, it would be desirable to have openly-available source code of a standard model available as has been done for Huntington's disease [57] where results for the reference data set are known. A logistic regression with specified MFCC features as inputs and fluent/stuttered as outcomes is one possible standard model. Availability of a standard model would provide a basis for comparison when other data sets and symptoms, and model architectures and features are used.

5.4. Model performances

A range of performance measures was used in studies, making model comparisons impossible. Model performance depends on the complexity of the classification problem as well as the complexity of the model being trained. The Vapnik–Chervonenkis (VC) dimension [58] is one such measure of a model's complexity. Technically, the VC dimension of a model is the largest set of points that the model can *'shatter'*. The power of the dataset is directly related to the VC dimension, given by the Equation[2]:

$$N = F\left(\frac{VC\ +\ \ln\frac{1}{d}}{\epsilon}\right)$$

Where VC is the VC dimension, d is the probability of failure and epsilon is the learning error. As noted in [60], the amount of data needed for learning depends on the complexity of the model.

A. Validation and metrics

A full description of model comparison as well as null hypothesis testing within a ML framework is beyond the scope of the current review. For further details see [61] and [62].

Model validation and performance assessment show that there are many options to consider. One could report metrics such as, accuracy, specificity, precision, recall or sensitivity, fallout, or false positive rate (FPR) measures, ROC Area, R^2, and Root mean square error. One could also report performance in terms of model quality such as the Akaike information criterion or Bayes information criterion.

If a full confusion matrix is given (Table 5), all of the above metrics and performance assessments are computable.

Table 5. Confusion matrix for a binary classification model which allows several accuracy statistics to be computed.

Type	**Observation Predicted as Fluent**	**Observation Predicted as Stuttered**	**Total**
Observation Labelled as Fluent	700	100	800
Observation Labelled as Stuttered	50	150	200
Total	750	250	1000

Most studies do not report such data. Instead, they simply report a performance metric with no context. There is a distinct lack of quantification of the models' effect sizes, and some procedure is required to assess whether one model *significantly* outperforms another. As discussed in the 'Machine Learning Models' section, although some studies compared different models on the same dataset, no tests were implemented to assess whether this difference was large enough to conclude that one model is better than the other. A method for calculating the effect size and its confidence intervals has been proposed [63] and this would allow claims about which model performs best to be assessed. Adoption of this approach would allow results to be accumulated and meta-analytic techniques applied to them that, in turn, would allow optimal model design and feature importance for the automatic classification of stuttered speech to be determined.

Choice of the various measures used for assessing performance require that different aspects of model performance depend on the aims of the model and the characteristics of the classification problem. For instance, in the classification of a rare disease, a false positive is not as detrimental as a false negative. In the case of false positives, further tests could be run to support the diagnosis. However, in the case of false negatives the patient continues without

[2] This provides the bounds of the data required but, for a space of binary functions. Alternate formulations of the power of the sample can be calculated for other models such as neural networks [59].

treatment. In this example Specificity may be preferred over Sensitivity for report.

Considering stuttering, given that stuttered speech is relatively infrequent compared to fluent speech even in PWS, missing a stutter (Precision) may be of greater consequence than incorrectly predicting a stutter during fluent speech (false positive). This relates to a general issue within the field of automatic stuttering recognition, given that stuttering occurs on approximately 1-5% of words in PWS. Were an ML model to predict all speech was fluent, it would be correct 95-99% of the time. Clearly, however, this model does not recognize stutters.

B. Recommendations

A complete range of performance statistics is required, or data in the format of Table 5 that allow all metrics to be computed should be supplied. These should also be available for any reference model (section 5.3.E) and a standard data set (section 5.1.D) for comparison.

To determine the expected benefit when a large dataset is used, studies should publish either the VC-dimension itself with the estimated upper and lower bounds of a dataset needed to adequately train the proposed model or provide the information necessary to calculate the VC dimension.

V. Summary

This systematic review evaluated models for the automatic recognition of stuttered speech. Twenty-seven publications met the inclusion criteria. It is clear that although this is a nascent field, there is promise that models could improve research and therapeutic applications. The issues in current ML models are limited because: (1) the current models are constrained and, in some cases, biased by what data are employed; (2) it is not definitely known what features from the audio signal provide the greatest information to separate fluent from dysfluent speech and even different forms of dysfluent speech; and (3) it is not known which modeling approach is best suited for the symptom classifications being made.

Work conducted after the closing date of this systematic review suggests that deep learning techniques may further improve automatic classification of stuttering [23][64]. Further work is needed to confirm whether modern deep learning methods can allow advances in performance as seen in other fields such as computer vision [65].

The systematic review has made some preliminary conclusions as well as clear recommendations for future work within the field. Furthermore, it is highly recommended that researchers consider the generic reporting standards set out in [66]. This should prevent further confusion within the field and allow future work to consider a meta-analysis of ML studies.

References

[1] P. Howell, *Recovery from stuttering*. New York, NY, USA: Psychology Press, 2010.

[2] P. Howell, "Screening school-aged children for risk of stuttering," *Journal of Fluency Disorders*, vol. 38, no. 2, pp. 102–123, 2013.

[3] P. Howell, K. Tang, O. Tuomainen, S. K. Chan, K. Beltran, A. Mirawdeli, and J. Harris, "Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds," *International Journal of Language & Communication Disorders*, vol. 52, no. 5, pp. 595–611, 2017.

[4] P. Howell, L. Y. Chua, K. Yoshikawa, H. H. Tang, T. Welmillage, J. Harris, and K. Tang, "Does working-memory training given TO Reception-class children improve the speech of children at risk of Fluency difficulty?," *Frontiers in Psychology*, vol. 11, 2020.

[5] R. Alsulaiman, J. Harris, S. Bamaas, and P. Howell, "Identification of Stuttering and Word-finding Difficulty in Samples of Arabic and English Preschool Children," *Frontiers in Pediatrics*, to be published.

[6] P. Howell, "Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering," *Stammering research: an on-line journal published by the British Stammering Association*, 2004 [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18259594/, Accessed on: Aug. 17, 2021.

[7] L. Barrett and P. Howell, "Impact of Altered Sensory Feedback on Speech Control in Fluent Speakers and Speakers who Stutter," in Manual of Clinical Phonetics, 1st ed., Routledge, 2021, pp. 461-479.

[8] P. Reed and P. Howell, "Suggestions for improving the long-term effects of treatments for stuttering: A review and synthesis of frequency-shifted feedback and operant techniques," *European Journal of Behavior Analysis*, vol. 1, no. 2, pp. 89–106, 2000.

[9] P. Howell and S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech," *Proceedings of the First World Congress on Fluency Disorders*, vol. 2, pp. 372-374, 1995.

[10] P. Howell and M. Huckvale, "Facilities to assist people to research into stammered speech," *Stammering research: an on-line journal published by the British Stammering Association*, 2004 [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18418475/, Accessed on: Aug. 17, 2021.

[11] P. Howell, J. Au-Yeung, S. Sackin, K. Glenn, and L. Rustin, "Detection of supralexical dysfluencies in a text read by children who stutter," *Journal of Fluency Disorders*, vol. 22, no. 4, pp. 299–307, 1997.

[12] E. Yairi and N. G. Ambrose, *Early childhood stuttering: For clinicians by clinicians*. Austin, TX: PRO-ED, 2005.

[13] M. Dzienkowski, "Eye Tracking Study of Visual Attention during Visual-auditory Diagnosis of Speech Non-fluencies," *8th International Conference of Education, Research and Innovation (ICERI)*, pp. 2759–2753, 2015.

[14] S. Yildirim and S. Narayanan, "Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 2–12, 2009.

[15] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Annals of Internal Medicine*, vol. 151, no. 4, p. 264, 2009.

[16] American Psychological Association, "Reporting standards for research in psychology: Why do we need them? What might they be?" *American Psychologist*, vol. 63, no. 9, pp. 839–851, 2008.

[17] Bishop, C., 2016. *Pattern recognition and machine learning*. New York: Springer.

[18] L. Barrett, J. Hu, and P. Howell, "Systematic Review, Meta-Analysis and Meta-Synthesis of Machine Learning Models Classification of Stuttered Speech," https://osf.io/ctzjk/, 2021.

[19] A. P. Siddaway, A. M. Wood, and L. V. Hedges, "How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses," *Annual Review of Psychology*, vol. 70, no. 1, pp. 747–770, 2019.

[20] P. Howell, S. Davis, and J. Bartrip, "The University College London archive of Stuttered Speech (UCLASS)," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 2, pp. 556–569, 2009.

[21] N. B. Ratner and B. MacWhinney, "Fluency Bank: A new resource for fluency research and practice," *Journal of fluency disorders*, vol. 56, p. 69-80.

[22] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[23] T. Kourkounakis, A. Hajavi, and A. Etemad, "FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2986-2999, 2021.

[24] J. C. Valentine and H. M. Cooper, "Can we measure the quality of causal research in education?" *Empirical Methods for Evaluating Educational Interventions*, pp. 85–111, 2005.

[25] P. Jüni, "The hazards of scoring the quality of clinical trials for meta-analysis," *JAMA*, vol. 282, no. 11, p. 1054, 1999.

[26] J. R. Landis and G. G. Koch, “The measurement of Observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, p. 159, 1977.

[27] 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,” *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 5, pp. 1085–1096, 1997.

[28] A. Czyzewski, A. Kaczmarek, and B. Kostek, “Intelligent Processing of Stuttered Speech,” *Journal of Intelligent Information Systems* vol. 21, 143–171, 2003.

[29] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński, “Automatic detection of disorders in a continuous speech with the hidden markov models approach,” *Advances in Soft Computing*, pp. 445–453, 2007.

[30] K. Ravikumar, B. Reddy, and R. Rajagopal, H. Nagaraj, “Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies,” *International Journal of Electrical and Computer Engineering*, vol. 2(10), pp. 2142 – 2145, 2008.

[31] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, “MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA,” *2009 IEEE Student Conference on Research and Development (SCOReD)*, 2009.

[32] K. Ravikumar, R. Rajagopal, and H. Nagaraj, “An Approach for Objective Assessment of Stuttered Speech Using MFCC Features,” *Digital Signal Processing Journal*, vol. 9, no. 1, pp. 1687-4811, 2009.

[33] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and E. Smołka, “Artificial neural networks in the disabled speech analysis,” *Advances in Intelligent and Soft Computing*, pp. 347–354, 2009.

[34] J. Pálfy, “ANN and SVM based recognition of the dysfluencies of speakers with stuttering,” *Mendel 2011: 17th International Conference on Soft Computing*, pp. 440-447, 2011.

[35] J. Pálfy and J. Pospíchal, “Recognition of repetitions using Support Vector Machines, Signal Processing Algorithms,” *Architectures, Arrangements, and Applications* pp. 1-6, 2011.

[36] O. Ai, M. Hariharan, S. Yaacob, and L. Sin Chee, “Classification of Speech Dysfluencies with MFCC and LPCC Features,” *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157–2165, 2012.

[37] M. Hariharan, V. Vijean, C. Y. Fook, and S. Yaacob, “Speech stuttering assessment using sample entropy and least square support vector machine,” *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, 2012.

[38] M. Hariharan, L. S. Chee, O. C. Ai, and S. Yaacob, “Classification of Speech Dysfluencies Using LPC Based Parameterization Techniques,” *Journal of Medical Systems*, vol. 36, no. 3, pp. 1821–1830, 2012.

[39] C. Y. Fook, H. Muthusamy, L. S. Chee, A. H. Adom, and S. B. Yaacob, “Comparison of speech parameterization techniques for the classification of speech disfluencies,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, pp. 1983–1994, 2013.

[40] M. Hariharan, C. Y. Fook, R. Sindhu, A. H. Adom, and S. Yaacob, “Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy,” *Digital Signal Processing*, vol. 23, no. 3, pp. 952–959, 2013.

[41] J. Pálfy and J. Pospíchal, “Algorithms for dysfluency detection in symbolic sequences using suffix arrays,” *Text, Speech, and Dialogue*, pp. 76–83, 2013.

[42] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and E. Smołka, “Hierarchical ANN system for stuttering identification,” *Computer Speech & Language*, vol. 27, no. 1, pp. 228–242, 2013.

[43] P. Mahesha and D. S. Vinod, “Support vector machine-based stuttering dysfluency classification using GMM supervectors,” *International Journal of Grid and Utility Computing*, vol. 6, no. 3/4, p. 143, 2015.

[44] A. Kobus, W. Kuniszyk-Jóźkowiak, and I. Codello, “Automatic syllable repetition detection in continuous speech based on linear prediction coefficients,” *Advances in Intelligent Systems and Computing*, pp. 295–304, 2016.

[45] P. Mahesha and D. S. Vinod, “Gaussian mixture model based classification of stuttering dysfluencies,” *Journal of Intelligent Systems*, vol. 25, no. 3, pp. 387–399, 2016.

[46] P. S. Savin, P. B. Ramteke, and S. G. Koolagudi, “Recognition of repetition and prolongation in stuttered speech using ANN,” *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, pp. 65–71, 2016.

[47] P. Mahesha and D. S. Vinod, “LP-Hillbert transform based MFCC for effective discrimination of stuttering dysfluencies,” *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017.

[48] G. Manjula, M. Shivakumar, and Y. V. Geetha, “Adaptive optimization based neural network for classification of stuttered speech,” *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy - ICCSP '19*, 2019.

[49] S. Gupta, R. S., R. K., and R. Verma, “Deep learning Bidirectional LSTM based detection of Prolongation and repetition in Stuttered speech using Weighted MFCC,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020.

[50] T. Kourkounakis, A. Hajavi, and A. Etemad, “Detecting multiple Speech Disfluencies using a Deep Residual network with bidirectional Long short-term memory,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[51] N. Mishra, A. Gupta, and D. Vathana, “Optimization of stammering in speech recognition applications,” *International Journal of Speech Technology*, 2021.

[52] M. E. Wingate, “SLD is not stuttering,” *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 381–383, 2001.

[53] R. AlSulamainan, “Arabic Fluency assessment: Procedures for Assessing Stuttering in Arabic Preschool Children,” Ph.D. dissertation, Dept. Experimental Psychology, University College London, London, UK, 2021

[54] P. Howell, A. Staveley, S. Sackin, and L. Rustin, “Methods of interval selection, presence of noise and their effects on detectability of repetitions and prolongations,” *Journal of the Acoustical Society of America*, vol. 104, pp. 3558-3567, 1998.

[55] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A. Roberts, and E. Mower Provost, “Classification of Huntington disease using acoustic and lexical features,” *Interspeech 2018*, 2018.

[56] J. Jiang, C. Lu, D. Peng, C. Zhu, and P. Howell, “Classification of types of stuttering symptoms based on brain activity,” *PLoS ONE*, vol. 7, no. 6, 2012.

[57] R. Riad, H. Titeux, L. Lemoine, J. Montillot, J. H. Bagnou, X.-N. Cao, E. Dupoux, and A.-C. Bachoud-Lévi, “Vocal markers from sustained phonation in huntington’s disease,” *Interspeech 2020*, 2020.

[58] V. N. Vapnik and A. Y. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities,” *Measures of Complexity*, pp. 11–30, 2015.

[59] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press, 2019, pp. 67-78.

[60] B. Juba and H. S. Le, “Precision-recall versus accuracy and the role of large data sets,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4039–4048, 2019.

[61] D. Berrar, “Confidence curves: An alternative to null hypothesis significance testing for the comparison of classifiers,” *Machine Learning*, vol. 106, no. 6, pp. 911–949, 2017.

[62] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[63] D. Berrar and J. A. Lozano, “Significance tests or confidence intervals: Which are preferable for the comparison of classifiers?” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 25, no. 2, pp. 189–206, 2013.

[64] V. Mitra, Z. Huang, C. Lea, L. Tooley, S. Wu, D. Botten, ... and J. Bigham, “Analysis and Tuning of a Voice Assistant System for Dysfluent Speech,” *arXiv preprint arXiv:2106.11759*, 2021.

[65] N. O’Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, ... and J. Walsh, “Deep learning vs. traditional computer vision,” In *Science and Information Conference*, Springer, Cham, 2019, pp. 128-144.

[66] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for MODEL REPORTING,” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.