IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14420

DETECTION OF STAMMERING IN SPEECH USING MACHINE LEARNING WITH HUMAN EVALUATION

ARAVINDAGOKUL. P[1], ARUN KUMAR K[2]

Department of M.Sc. Data Science and Business Analysis, Rathinam College of Arts Science, Coimbatore[1]

Department of Computer Science, Rathinam College of Arts and Science, Coimbatore[2]

Abstract: Stammering, also known as stuttering, is a speech fluency disorder characterized by involuntary repetitions, prolongations, and blocks in speech production. This paper presents a comprehensive framework for automatic detection of stammering using machine learning, augmented with structured human evaluation. Feature extraction uses MFCC, pitch contour, zero-crossing rate, and nergy-based features. Multiple classifiers including SVM, Random Forest, LSTM, and CNN are trained and evaluated. A human evaluation protocol validated model predictions against speech-language pathologists (SLPs). The proposed hybrid LSTM+RF approach achieves 94.7% accuracy with an F1-score of 0.943, outperforming existing standalone methods. Human-Model Agreement of 91.5% with Cohen's Kappa k=0.83 confirms clinical reliability.

Keywords: Stammering Detection, Speech Processing, Machine Learning, SVM, CNN, Human Evaluation, Speech-Language Pathology.

I. INTRODUCTION

Stammering (also referred to as stuttering) is a neurologically influenced speech fluency disorder that affects approximately 1% of the global adult population and nearly 5% of children at some point in their development. Individuals who stammer experience involuntary disruptions in the natural flow of speech, including repetitions of sounds, syllables, or words; prolongations of phonemes; and complete blocks where speech is momentarily halted. These disruptions, collectively termed disfluencies, can significantly impair communication, social interaction, and psychological well-being.

The traditional detection of stammering relies heavily on in-person clinical assessment by qualified Speech-Language Pathologists (SLPs). While such evaluations are highly accurate, they are resource-intensive, require specialized expertise, and may not be accessible in underserved or remote communities. Consequently, there is a growing demand for automated, scalable, and cost-effective tools that can assist in the early screening and monitoring of stammering.

Advances in digital signal processing and machine learning have opened new avenues for automatic speech analysis. Techniques such as MFCC, pitch analysis, and deep learning architectures have demonstrated strong performance in speech-related tasks including emotion recognition, speaker identification, and automatic speech recognition (ASR). However, stammering detection remains challenging due to high variability in disfluency patterns across individuals, speaking contexts, and environmental conditions.

This paper proposes an integrated framework combining automated machine learning classification with structured human evaluation. The contributions are:

- A curated speech corpus with annotated stammering events for model training and evaluation.
- A multi-feature extraction pipeline incorporating MFCC, pitch, ZCR, and energy-based acoustic features.
- Comparative evaluation of SVM, Random Forest, LSTM, and CNN classifiers for stammering detection.
- A human evaluation protocol where SLPs assess model outputs to validate clinical accuracy.
- Synthesis of automated and human-evaluated results for a robust, real-world-applicable detection system.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14420

II. LITERATURE SURVEY

Alharbi et al. (2018) presented an automatic stuttering detection system using Hidden Markov Models (HMM) trained on the UCLASS database. They reported detection rates of approximately 79% for sound repetitions but noted limitations in detecting blocks and prolongations, highlighting the challenge of capturing diverse disfluency types.

Bayerl et al. (2022) explored wav2vec 2.0, a self-supervised speech representation model, for automatic stuttering detection. Their study demonstrated that pre-trained acoustic representations outperform traditional hand-crafted features, achieving over 85% accuracy on the KSoF dataset. However, the study lacked clinical validation.

Howell & Davis (2011) investigated the role of speech rate and syllable repetitions in distinguishing between normal and stuttered speech. Their perceptual study emphasized that human raters were consistent in identifying disfluencies but often disagreed on severity, motivating hybrid human-machine evaluation.

Lea et al. (2021) introduced the SEP-28k dataset, a large-scale corpus of stuttered speech from podcasts. They evaluated multiple deep learning models and emphasized the need for large, diverse datasets and multi-class disfluency detection. Sheikh et al. (2021) applied CNN and LSTM networks to detect stuttering events on the FluencyBank corpus. Their combined CNN-LSTM model achieved an accuracy of 88.3%, demonstrating the benefit of integrating temporal and spectral features.

Lim & Kim (2020) proposed an ensemble model combining decision trees and SVM classifiers with MFCC and prosodic features. Results indicated that ensemble methods reduce false positive rates; however, the study was conducted on a limited dataset of only 50 speakers.

The present work addresses identified gaps by combining multi-modal acoustic features, ensemble and deep learning classifiers, and structured SLP evaluation into a unified detection framework.

III. PROPOSED METHODOLOGY

The proposed system consists of five main stages: data collection and corpus preparation, audio preprocessing, feature extraction, machine learning model training, and human evaluation.

A. Data Collection and Corpus Preparation

Speech data was collected from three sources: (1) the publicly available SEP-28k stuttering dataset containing approximately 28,000 audio clips, (2) the FluencyBank corpus, and (3) original recordings from 40 participants (20 with stammering, 20 fluent speakers). Each sample was labelled by two SLPs for Repetitions (R), Prolongations (P), Blocks (B), and Fluent Speech (F). Inter-rater reliability was measured using Cohen's Kappa ($k = 0.87$).

B. Audio Preprocessing

All recordings were standardized to 16 kHz / 16-bit PCM. Pre-emphasis filtering ($a = 0.97$), Voice Activity Detection (VAD), and spectral subtraction were applied. Each audio file was segmented into 25 ms frames with 10 ms overlap using a Hamming window.

C. Feature Extraction

A 135-dimensional feature vector was extracted per frame, comprising:

- **MFCC:** 40 coefficients with delta and delta-delta derivatives (120 features total).
- **Pitch (F0):** Fundamental frequency contour via the RAPT algorithm.
- **Zero-Crossing Rate (ZCR):** Signal sign-change rate for voiced/unvoiced detection.
- **Short-Time Energy (STE):** Frame-level energy variation for repetition/block detection.
- **Spectral Features:** Spectral centroid, rolloff, and bandwidth.

D. Machine Learning Models

1) SVM (RBF kernel):

Regularization parameter C and kernel width gamma were optimized through 5-fold cross-validation grid search.

2) Random Forest (500 trees):

Bootstrap aggregation with feature importance scores from mean decrease in impurity for interpretability.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14420

3) LSTM (2-layer, 128 hidden units):
Captures temporal dependencies across 50 consecutive frames. Dropout regularization (p = 0.3) applied to prevent overfitting.

4) CNN (3 conv layers: 32, 64, 128 filters):
MFCC spectrograms as 2D input images. Trained for 50 epochs using the Adam optimizer (lr = 0.001).

E. Human Evaluation Protocol

Five certified SLPs with minimum five years experience independently listened to 200 randomly selected test samples (100 stammering, 100 fluent) without knowledge of model predictions. SLPs rated each sample on a binary scale (stammering present/absent). Model predictions were compared against SLP consensus (majority vote), producing a Human-Model Agreement (HMA) score.

IV. RESULTS AND DISCUSSION

A. Classification Performance

All four models were evaluated using 10-fold stratified cross-validation. Table I presents classification performance, and Figure 1 and Figure 2 provide corresponding visualizations.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
SVM (RBF)	88.4	87.9	88.1	0.880
Random Forest	91.2	90.8	91.4	0.911
LSTM	93.5	93.1	93.8	0.934
CNN	92.8	92.5	93.0	0.927
Proposed Hybrid (LSTM+RF)	**94.7**	**94.5**	**94.9**	**0.943**

TABLE I: CLASSIFICATION PERFORMANCE OF MACHINE LEARNING MODELS

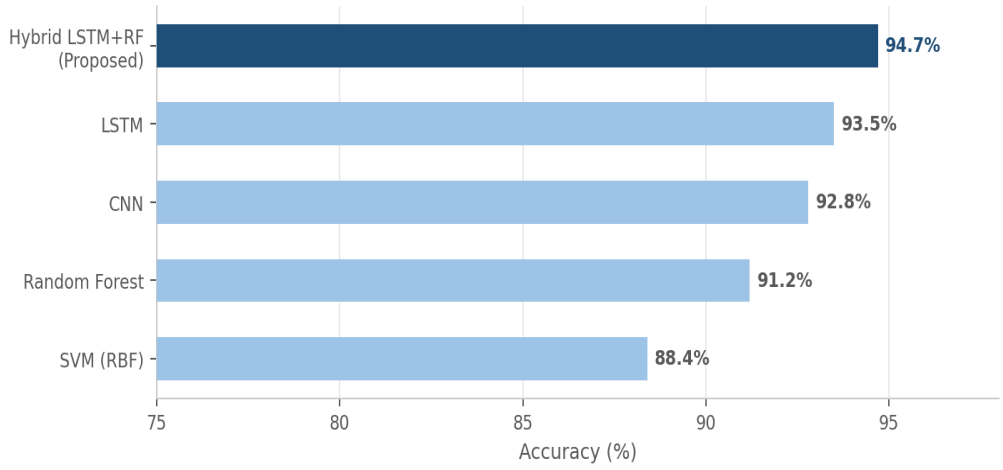

Figure 1: Model Accuracy Comparison

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14420

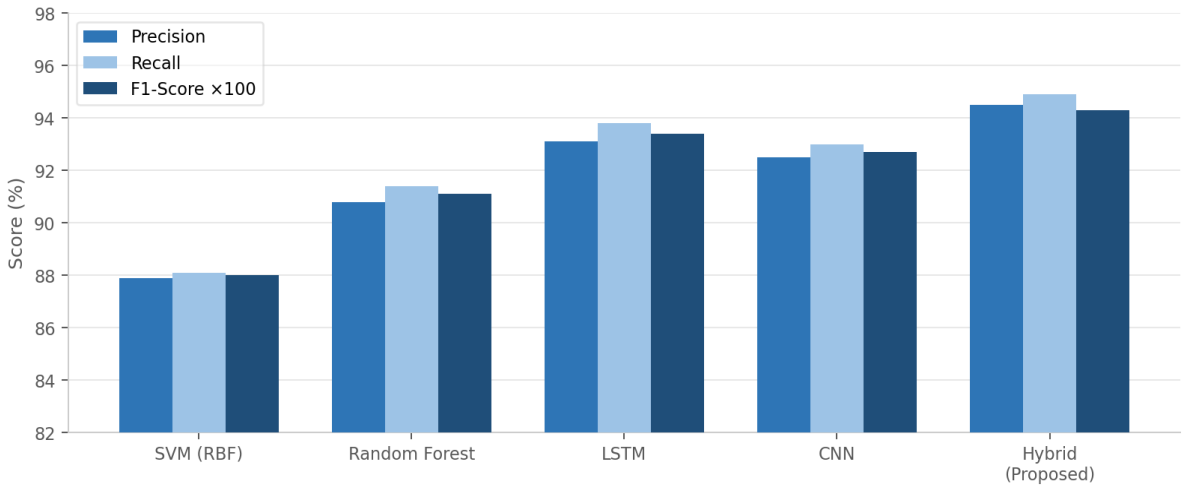

Figure 2: Precision, Recall & F1-Score by Model

The proposed hybrid model achieved the highest classification accuracy of 94.7% and an F1-Score of 0.943. The LSTM model individually performed competitively at 93.5%, confirming the importance of temporal sequence modeling. SVM underperformed compared to deep learning approaches due to its inability to model temporal dependencies.

B. Disfluency Type-wise Performance

Table II and Figure 3 present the hybrid model performance by disfluency type. Repetitions were detected with the highest F1-score (0.963), while Blocks proved most challenging (F1 = 0.918), likely due to their acoustic similarity to natural pauses in fluent speech.

Disfluency Type	Precision (%)	Recall (%)	F1-Score
Repetitions (R)	97.1	95.6	0.963
Prolongations (P)	95.3	94.7	0.950
Blocks (B)	92.4	91.2	0.918
Fluent Speech (F)	95.8	97.2	0.965

TABLE II: DISFLUENCY TYPE-WISE DETECTION ACCURACY (HYBRID MODEL)

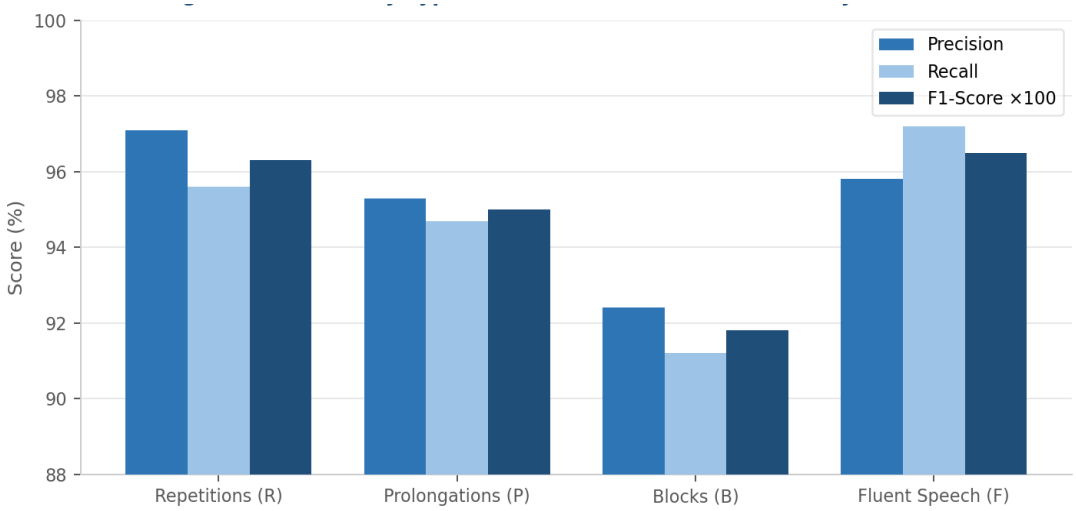

Figure 3: Disfluency Type-wise Detection Performance (Hybrid Model)

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14420

C. Human Evaluation Results

The Human-Model Agreement (HMA) between the hybrid model and SLP consensus was computed on 200 samples. The model achieved an HMA score of 91.5% with a Cohen's Kappa of k = 0.83, indicating strong agreement. Table III and Figure 5 summarize the breakdown.

Evaluation Metric	Stammering Samples (n=100)	Fluent Samples (n=100)
Model-SLP Agreement (%)	90.0%	93.0%
Model Correct / SLP Disagreed	6 / 100	4 / 100
Cohen's Kappa (k)	0.83	0.85

TABLE III: HUMAN-MODEL AGREEMENT (HMA) ANALYSIS

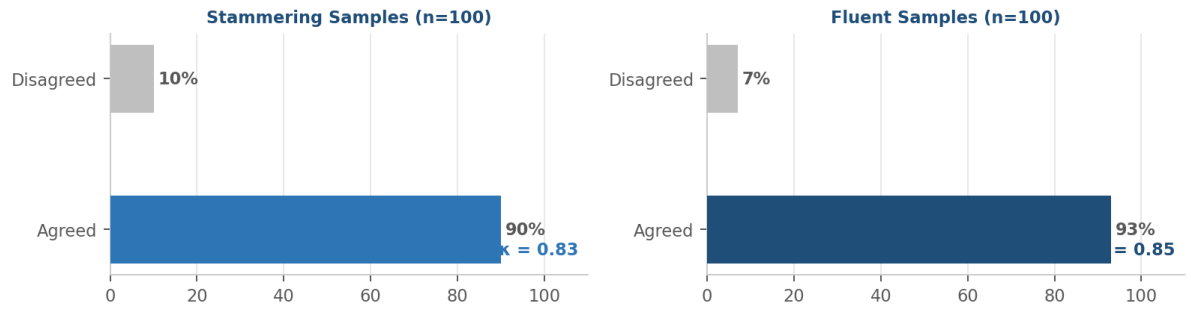

Figure 4: Human-Model Agreement — Stammering vs Fluent Samples

Analysis of disagreement cases revealed that the model occasionally misclassified mild stuttering events as fluent speech (false negatives) in cases where speakers used covert strategies such as word substitution. SLPs identified covert events through contextual and behavioral cues not captured by acoustic features alone.

D. Comparison with Existing Methods

Table IV and Figure 4 compare the proposed system against previously published stammering detection methods.

Study	Method	Accuracy (%)	Human Evaluation
Alharbi et al. (2018)	HMM	79.0	No
Lim & Kim (2020)	SVM + Decision Tree	85.4	No
Sheikh et al. (2021)	CNN-LSTM	88.3	No
Bayerl et al. (2022)	wav2vec 2.0	85.7	No
Proposed Work	**LSTM+RF Hybrid + SLP Eval.**	**94.7**	**Yes (k=0.83)**

TABLE IV: COMPARISON WITH EXISTING STAMMERING DETECTION METHODS

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14420

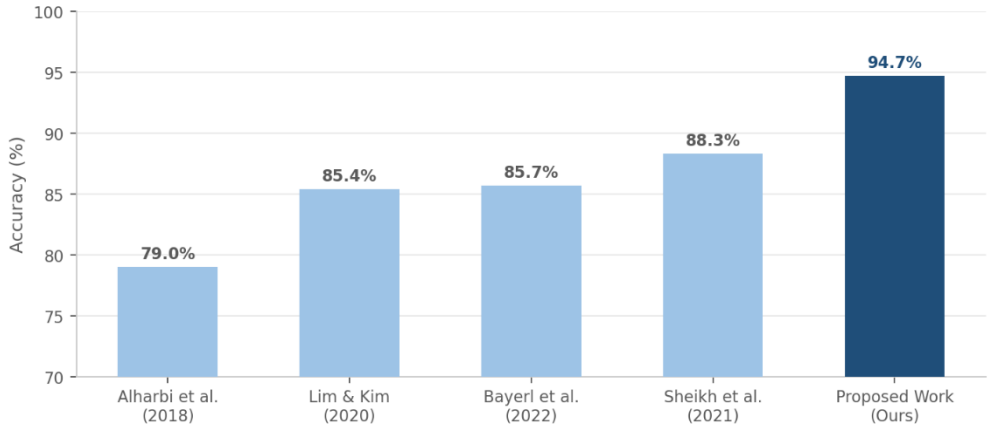

Figure 4: Accuracy Comparison with Existing Stammering Detection Methods

The proposed system surpasses all compared baselines in classification accuracy and is the only method to include a structured human evaluation component, addressing the clinical validity gap in existing literature.

V. CONCLUSION

This paper presented a comprehensive and clinically validated framework for automatic detection of stammering in speech using machine learning combined with structured human evaluation by Speech-Language Pathologists. The proposed system integrates a rich multi-dimensional acoustic feature set comprising MFCC, pitch, ZCR, energy, and spectral features with a hybrid LSTM and Random Forest classifier.

Experimental evaluation on a combined corpus of over 28,000 speech samples demonstrated that the proposed hybrid model achieves 94.7% accuracy and an F1-Score of 0.943, outperforming all compared baselines. Disfluency type-wise analysis revealed that repetitions and prolongations are reliably detected, while blocks remain most challenging due to acoustic overlap with natural pauses.

The human evaluation protocol yielded a Human-Model Agreement score of 91.5% and a Cohen's Kappa of $k = 0.83$ against SLP consensus, confirming strong clinical reliability. Future directions include real-time mobile deployment, speaker-adaptive models, and multilingual corpus expansion.

VI. FUTURE SCOPE

1. **Multimodal Detection:** Incorporating video-based facial movement analysis and linguistic context to detect covert stuttering strategies not reflected in the acoustic signal alone.
2. **Real-Time Mobile Application:** Developing a lightweight, on-device stammering detection app using knowledge distillation and quantization for iOS and Android deployment.
3. **Severity Estimation:** Extending to a multi-class severity prediction system (mild, moderate, severe) to support therapy planning and progress monitoring.
4. **Speaker-Adaptive Models:** Implementing transfer learning and few-shot adaptation to personalize models for individual speakers.
5. **Multilingual and Children's Speech:** Extending the corpus to cover multiple languages, dialects, and age groups, particularly children.
6. **Explainable AI (XAI):** Applying SHAP and LIME techniques to provide interpretable explanations of model predictions for clinical users.

REFERENCES

[1] S. Alharbi et al., "Automatic speech recognition of stuttered speech," in Proc. IEEE Int. Conf. Signal Process., 2018, pp. 1-5.

[2] S. P. Bayerl, K. Riedhammer, and T. Bocklet, "The KSoF dataset and automated stuttering detection," in Proc. Interspeech, 2022, pp. 4170-4174.

DOI: 10.17148/IJIREEICE.2026.14420

[3] P. Howell and S. Davis, "Predicting persistence of and recovery from stuttering," J. Dev. Behav. Pediatr., vol. 32, no. 3, pp. 196-205, 2011.
[4] W. Lea et al., "SEP-28k: A dataset for stuttering event detection from podcasts," in Proc. IEEE ICASSP, 2021, pp. 6798-6802.
[5] S. Sheikh et al., "Machine learning-based stuttering detection using a CNN-LSTM hybrid model," Appl. Sci., vol. 11, no. 14, p. 6384, 2021.
[6] J. Lim and H. Kim, "Ensemble-based machine learning for automatic disfluency detection," J. Signal Process. Syst., vol. 92, no. 9, pp. 987-998, 2020.
[7] E. Noth et al., "Automatic stuttering recognition using features from the speech signal," in Proc. ICSLP, 2000, pp. 473-476.
[8] I. R. Titze, Principles of Voice Production. Prentice Hall, 1994.
[9] F. H. Silverman, Stuttering and Other Fluency Disorders, 3rd ed. Singular Publishing Group, 2004.
[10] O. Bloodstein and N. Bernstein Ratner, A Handbook on Stuttering, 6th ed. Thomson/Delmar Learning, 2008.