International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 299-313
https://doi.org/10.70917/ijcisim-2026-4576
Article

Model Development and Validation for Repetition Severity Assessment in Stuttered Speech Using Clinical Speech Datasets

Pooja J N[1], H Y Vani[2], Rakshith P[3], M A Anusuya[4], Swetha P M[5]

[1]Department of Computer Science and Engineering , JSS Science and technology university, India Email:Poojajn@jssstuniv.in
[2]Department of Information Science and Engineering, JSS Science and technology university, India Email:Vanihy@jssstuniv.in
[3]Department of Computer Science and Engineering , JSS Science and technology university, India Email:rakshith@jssstuniv.in
[4]Department of Computer Science and Engineering , JSS Science and technology university, India Email;Anusuya_ma@jssstuniv.in
[5]Department of Computer Science and Engineering , JSS Science and technology university, India Email:swethapm@jssstuniv.in

Abstract: Stuttering disrupts the forward flow of speech through involuntary repetitions, prolongations, and blocks, with repetitions being the most common and clinically telling of the three. Quantifying how severe those repetitions are matters for diagnosis, therapy planning, and tracking whether treatment is actually working. The catch is that severity has long been judged by ear, by trained speech-language pathologists counting disfluent events and assigning ratings on standardised scales. That process is slow, variable between clinicians, and constrained by who is available. This paper builds and tests a computational model that grades repetition severity from clinical speech recordings. The model draws on a multimodal feature set combining acoustic, prosodic, temporal, and spectral descriptors, evaluated on 480 audio samples from 60 adult speakers with persistent developmental stuttering. Two certified clinicians labelled each sample as mild, moderate, or severe, with strong inter-rater agreement. Recursive feature elimination trimmed the feature set to a compact, discriminative subset, and five classifiers were trained under stratified ten-fold cross-validation repeated five times. A Gradient Boosted Trees model reached 91.46 percent mean accuracy, a macro F1-score of 0.90, and a Cohen kappa of 0.86, ahead of Random Forest, a Support Vector Machine, a Multilayer Perceptron, and Logistic Regression. Per-class scores were strong for mild and severe cases and weaker for moderate, where acoustic characteristics overlap with both neighbouring classes. Friedman and Nemenyi tests confirmed the top model's lead was significant at the 0.05 level. The pipeline is reproducible and the results support its use as a clinical decision-support tool.

Keywords: Stuttering, repetition severity, speech disfluency, machine learning, acoustic features, clinical speech corpus, fluency assessment, gradient boosting

1. INTRODUCTION

1.1 Background and Motivation

Roughly one percent of adults stutter, and for most of them the dominant disfluency is repetition, the involuntary recycling of sounds, syllables, or whole words. Clinicians have always graded severity by listening, counting, and rating, typically using instruments such as the Stuttering Severity Instrument. Done well, this is thorough. Done at scale, it is expensive, slow, and uneven across raters. The clinical bottleneck is not the rating itself but the people qualified to do it. Where expert speech-language pathologists are scarce, severity assessment simply does not happen consistently, and therapy decisions suffer.

The computational path around this bottleneck has three steps that have to hold together. Speech has to be represented by features that actually capture what makes a repetition a repetition. Those features have to map onto severity levels that line up with how clinicians think (Narasinga et al., 2026). And the whole system has to hold up on data it has not seen, across speakers, microphones, and stuttering subtypes. Each step has been worked on in isolation.

Copyright: © 2026 by the authors

Pulling them into a single, clinically grounded pipeline is where the open work sits, and that is the gap this paper addresses.

1.2 Research Objectives

The aim is to build and validate a supervised learning model that grades repetition severity from clinical speech recordings, with four supporting goals. The feature pipeline has to combine acoustic, prosodic, temporal, and spectral information rather than leaning on one modality (Batra et al., 2025). Several classifier families have to be compared under the same cross-validation protocol, so performance differences can be attributed to the algorithm rather than the features. The descriptors that carry the most weight in the final model need to be identified. And the performance gap between the best model and the rest has to be tested statistically, not just reported as a number.

1.3 Contributions

Four things stand out. The labelling protocol ties acoustic data to a three-level severity scale that mirrors clinical practice, with dual-annotator consensus rather than single-rater judgement. The feature set joins frame-level acoustic descriptors with utterance-level prosodic and temporal statistics, so the model sees both the texture of a repetition and its place in the broader utterance (Filipowicz and Kostek, 2023). The classifier comparison runs five algorithm families through identical validation, with significance testing, which the literature rarely does. And the per-class breakdown names the moderate class as the real diagnostic problem, a finding that shapes what clinical deployment should and should not expect.

1.4 Organisation of the Paper

Section 2 covers prior work on acoustic analysis of stuttering and machine learning for disfluency classification. Section 3 lays out the dataset, preprocessing, feature design, model development, and validation protocol. Section 4 reports the results, with performance tables and per-class analysis. Section 5 reads those results against clinical practice and earlier work. Section 6 closes with the limits of what was shown and where the work goes next.

2. Literature Review

According to Narasinga (2026), speech signal processing has become a highly effective approach for the automatic classification of stuttering and the evaluation of speech disorders in clinical environments. The study comprehensively investigates the integration of advanced acoustic signal processing techniques with machine learning algorithms to improve the identification of stuttering events. The author explains that traditional clinical evaluations depend largely on perceptual judgments made by speech-language pathologists, which often introduce subjectivity and inter-rater variability. To overcome these limitations, the research utilizes acoustic features such as Mel-frequency cepstral coefficients, pitch, spectral energy, formant frequencies, and temporal characteristics to distinguish fluent from disfluent speech. The study further evaluates several deep learning architectures capable of learning complex speech representations directly from raw and processed acoustic signals. Clinical validation demonstrates that automated systems achieve high diagnostic accuracy while significantly reducing assessment time. The author also emphasizes the importance of large and diverse speech datasets for improving model generalization across speakers with different ages, genders, and linguistic backgrounds (Narasinga et al., 2026). The findings indicate that hybrid deep learning frameworks consistently outperform conventional machine learning methods in detecting repetition, prolongation, and blocking events. The research concludes that speech signal processing combined with artificial intelligence provides objective, reproducible, and scalable solutions for stuttering assessment, thereby supporting speech-language pathologists in diagnosis, treatment planning, and long-term therapy monitoring while improving the consistency and reliability of clinical decision-making.

According to Batra (2025), the availability of high-quality annotated speech datasets is essential for advancing automatic stuttering recognition and severity assessment systems. The study introduces the BOLI dataset, specifically designed to capture diverse speech samples representing different forms of stuttering experienced by individuals across varying demographic backgrounds. The author explains that previous datasets were often limited by small sample sizes, inconsistent annotations, and insufficient representation of natural conversational speech. The proposed dataset addresses these limitations by including carefully annotated repetition, prolongation, and block events alongside comprehensive metadata describing speaker characteristics and recording conditions. The research highlights the significance of standardized annotations for developing reliable deep learning models capable of accurately identifying speech disfluencies. Experimental analyses demonstrate that the dataset supports multiple machine learning and neural network architectures while enabling comparative evaluation across different algorithms

(Batra et al., 2025). The author further discusses the importance of capturing authentic speaking situations rather than scripted speech, as natural conversations better represent clinical manifestations of stuttering. The dataset also facilitates multilingual and cross-speaker investigations, thereby improving model robustness and generalization. The study concludes that publicly available clinical speech resources such as BOLI provide a valuable foundation for future research in automated speech disorder analysis, encouraging reproducible experimentation and promoting the development of clinically applicable artificial intelligence systems for objective stuttering assessment and therapeutic monitoring.

According to Filipowicz (2023), deep learning architecture and training data volume play a significant role in determining the effectiveness of automatic stuttering detection systems. The study revisits machine learning approaches for identifying stuttering and its various subclasses while examining the influence of different neural network structures and dataset sizes on classification performance. The author compares convolutional neural networks, recurrent neural networks, and hybrid architectures to evaluate their ability to recognize complex temporal and spectral speech characteristics associated with stuttering. The findings indicate that increasing the amount of annotated training data substantially improves model accuracy and generalization across diverse speakers (Filipowicz and Kostek, 2023). The research also demonstrates that deeper neural architectures provide enhanced feature extraction capabilities, enabling more accurate discrimination between fluent speech and various disfluency types. The author explains that transfer learning and data augmentation techniques further improve performance when clinical datasets remain relatively limited. Experimental evaluations reveal that optimized deep learning models outperform traditional machine learning algorithms in both binary classification and multiclass recognition of stuttering events. The study emphasizes the necessity of balancing model complexity with computational efficiency to facilitate real-time clinical implementation. The research concludes that continuous improvements in neural network design and dataset expansion will contribute significantly to the development of reliable automated speech assessment systems capable of supporting clinical diagnosis, severity estimation, and speech therapy evaluation.

According to Valente (2025), accurate clinical annotations constitute the foundation for developing reliable automated stuttering severity assessment systems. The study focuses on creating standardized annotation protocols that accurately represent different types and severities of speech disfluencies observed during clinical evaluations. The author explains that inconsistent labeling practices across existing speech datasets often reduce the performance and reproducibility of machine learning models. To address this issue, comprehensive annotation guidelines were developed through collaboration with experienced speech-language pathologists, ensuring consistent identification of repetitions, prolongations, speech blocks, and secondary behaviors. The research demonstrates that standardized annotations significantly improve the training and validation of deep learning models designed for automatic severity estimation (Valente et al., 2025). The author further highlights the importance of incorporating temporal boundaries, contextual information, and clinician agreement scores into the annotation framework. Experimental validation shows that models trained using standardized clinical annotations achieve higher classification accuracy and stronger agreement with expert evaluations. The study also discusses the potential integration of annotated datasets with multimodal information, including facial expressions and articulatory movements, to improve assessment reliability. The findings suggest that standardized clinical annotations enhance the interpretability, transparency, and clinical acceptance of artificial intelligence systems. The research concludes that consistent annotation methodologies are essential for establishing trustworthy automated assessment tools capable of supporting speech-language pathologists in objective diagnosis and longitudinal therapy monitoring.

According to Barrett (2022), machine learning has emerged as a promising approach for the automatic detection of developmental stuttering, although several methodological challenges remain unresolved. The systematic review examines a broad range of studies employing traditional machine learning algorithms and deep neural networks for identifying speech disfluencies. The author evaluates the effectiveness of commonly used acoustic features, classification techniques, dataset characteristics, and validation strategies reported across previous investigations (Barrett et al., 2022). The review indicates that convolutional neural networks and recurrent neural networks generally outperform classical classifiers because of their ability to learn complex temporal and spectral representations directly from speech signals. However, considerable variability exists in dataset sizes, annotation quality, evaluation metrics, and experimental protocols, limiting direct comparison between published studies. The author also identifies the lack of standardized benchmark datasets as a significant obstacle to reproducible research. Furthermore, the review emphasizes that many studies rely on limited participant populations, thereby reducing model generalizability across different age groups, languages, and recording environments. Recommendations include the development of larger multilingual datasets, standardized annotation frameworks, external validation procedures, and explainable artificial intelligence models suitable for clinical practice. The study concludes that future progress in automated stuttering

detection depends on improved data quality, methodological consistency, and stronger collaboration between speech-language pathologists, engineers, and artificial intelligence researchers.

According to Batra (2025), deep learning provides an efficient mechanism for transforming encoded speech representations into clinically meaningful stuttering severity classifications. The study presents a framework that utilizes learned feature embeddings extracted from speech recordings to quantify the severity of speech disfluencies across multiple clinical categories. The author explains that conventional handcrafted acoustic features often fail to capture the subtle temporal variations associated with repetition frequency and duration. Consequently, deep neural networks are employed to automatically learn discriminative representations directly from speech signals. Experimental evaluations demonstrate that encoded features generated by deep learning architectures achieve superior classification performance compared with manually engineered feature sets (Batra et al., 2025). The proposed approach effectively differentiates mild, moderate, severe, and very severe stuttering while maintaining high accuracy across diverse speakers. The research also investigates the relationship between latent feature representations and clinically recognized severity scales, showing strong correspondence with expert evaluations. The author emphasizes that deep feature extraction reduces dependence on domain-specific feature engineering while improving robustness against background noise and speaker variability. The findings suggest that automated severity estimation can significantly support clinical assessments by providing objective, repeatable, and efficient measurements. The study concludes that representation learning through deep neural networks offers substantial potential for developing intelligent clinical decision-support systems capable of monitoring therapeutic progress and facilitating personalized intervention strategies.

According to Al-Banna (2022), machine learning algorithms have demonstrated substantial capability in detecting various forms of stuttering disfluencies through the analysis of acoustic speech characteristics. The study evaluates multiple supervised learning techniques for recognizing repetitions, prolongations, and speech blocks using carefully extracted acoustic features. The author explains that speech preprocessing, feature selection, and classifier optimization collectively influence the overall performance of automated detection systems. Acoustic descriptors including Mel-frequency cepstral coefficients, spectral energy, pitch, zero-crossing rate, and formant frequencies are employed to characterize fluent and disfluent speech segments. Comparative experiments reveal that ensemble learning methods and support vector machines achieve competitive classification performance, although deep learning architectures exhibit greater flexibility in modeling complex speech patterns (Al-Banna et al., 2022). The research further emphasizes the importance of balanced datasets and effective cross-validation procedures to minimize overfitting and improve model reliability. The author discusses practical challenges associated with speaker variability, environmental noise, and recording quality, all of which influence algorithm performance under real-world clinical conditions. The findings demonstrate that optimized machine learning models provide accurate and objective identification of stuttering events while reducing dependence on subjective perceptual evaluations. The study concludes that artificial intelligence technologies represent valuable complementary tools for clinical speech assessment, supporting speech-language pathologists through consistent, efficient, and scalable analysis of speech disfluencies.

According to Abubakar (2024), deep learning integrated with Mel-frequency cepstral coefficient feature extraction significantly improves the detection of stuttering disfluencies in synthetic speech signals. The study introduces the StutterNet framework, which combines acoustic feature extraction with advanced neural network architectures to classify speech segments exhibiting different forms of stuttering. The author explains that MFCC features effectively capture perceptually relevant spectral information while reducing the dimensionality of speech signals, thereby facilitating efficient neural network training. Experimental analyses indicate that the proposed framework achieves high detection accuracy across repetition, prolongation, and blocking events using synthetically generated speech datasets (Abubakar et al., 2024). The research also investigates the influence of network depth, feature dimensionality, and optimization strategies on classification performance. Results demonstrate that deep learning models consistently outperform traditional machine learning algorithms because of their superior capability to learn hierarchical speech representations. The author further discusses the usefulness of synthetic speech generation for overcoming limitations associated with scarce clinical datasets, enabling large-scale model development and controlled experimentation. Although synthetic speech cannot fully replace authentic clinical recordings, it provides valuable supplementary data for training robust artificial intelligence systems. The study concludes that integrating MFCC-based feature extraction with deep neural networks represents a promising direction for developing reliable automated stuttering detection systems that can support future clinical assessment, telehealth applications, and speech rehabilitation technologies.

3. Methodology

3.1 Dataset Description

Recordings came from a speech therapy clinic, collected over eighteen months from 60 adult speakers diagnosed with persistent developmental stuttering against DSM criteria. Participants gave informed consent and the institutional ethics committee approved the protocol. Each speaker produced three kinds of sample: reading a passage, describing a picture, and free conversation, to span the speaking contexts a clinician would actually hear (Zhang et al., 2025). Recording happened in a quiet room with a head-mounted condenser microphone ten centimetres from the lips, sampled at 44.1 kilohertz with 16-bit depth.

The corpus used for modelling held 480 samples, each a continuous utterance of ten to fifteen seconds. Two certified speech-language pathologists labelled each sample independently for repetition events and assigned a severity label on a three-level scale of mild, moderate, and severe. Inter-rater agreement, measured by Cohen kappa, was 0.82, which is strong. Where the two raters differed, they listened together and reached a consensus label. Class distribution was 168 mild, 162 moderate, and 150 severe, balanced enough for multiclass learning without resampling. Every sample was amplitude-normalised to a mean of negative twenty decibels and trimmed of leading and trailing silence through voice activity detection.

3.2 Preprocessing

Preprocessing aimed for consistency, not cleanliness at the cost of signal. Each sample was downsampled to 16 kilohertz, enough to preserve the spectral content that matters for speech while cutting the compute load. A pre-emphasis filter with a 0.97 coefficient lifted the higher frequencies. Framing used 25-millisecond windows with a ten-millisecond hop and a Hamming window (Zhang, 2025). Voice activity detection on energy and zero-crossing rate removed non-speech regions. Noise reduction was deliberately left out. Early trials showed that aggressive filtering attenuated the subtle cues around repetition events, the very things the model needs to see, so the natural acoustic character of the clinical recordings was preserved.

3.3 Feature Extraction

The feature pipeline reaches across four signal aspects that each say something different about a repetition. Spectral features came first: thirteen Mel-frequency cepstral coefficients plus their first and second derivatives, log-energy, and twelve Mel-filter bank energies, computed frame-wise and aggregated over each utterance by mean, standard deviation, skewness, and kurtosis, for 116 spectral descriptors. Prosodic features came second: fundamental frequency mean, range, variability, and jitter, with energy contour slope and voicing ratio, capturing the suprasegmental shape of repeated segments (Ng et al., 2024). Temporal features came third, drawn from the annotated repetition boundaries: repetition events per unit duration, mean repetition duration, mean inter-repetition interval, and the ratio of repeated to total phoneme duration. Spectral dynamics came fourth: spectral flux, spectral centroid variability, and spectral roll-off slope, describing how the spectrum moves within and around repetition events.

The full set was 154 descriptors per sample. Standardisation to zero mean and unit variance used statistics from the training folds only, so no information leaked across the cross-validation boundary.

3.4 Feature Selection

A 154-feature set against 480 samples is a recipe for overfitting unless the set is trimmed. Recursive feature elimination with cross-validated importance scoring did the trimming, using a Random Forest as the selector and averaging importance across five internal folds. The subset size that maximised cross-validated accuracy held 72 features (Kashyap et al., 2023). Spectral aggregates and temporal repetition statistics dominated the retained set, which lines up with what the feature importance analysis later showed about what the classifier actually leans on. All subsequent training used the 72-feature subset.

3.5 Model Development

Five algorithms were chosen to span the modelling landscape rather than to fill out a list. A radial-basis Support Vector Machine handles moderate-dimensional feature spaces well. Random Forest is robust to noise and captures nonlinear interactions without much tuning. Gradient Boosted Trees have a strong track record on tabular clinical prediction (Abeywickrama, 2024). A two-hidden-layer Multilayer Perceptron stands in for neural approaches on handcrafted features. Logistic Regression with elastic net regularisation is the linear baseline.

Hyperparameters were tuned by Bayesian optimisation with five-fold cross-validation inside each outer training fold. Search spaces followed standard practice: kernel width and regularisation for the Support Vector Machine, tree

count and depth for the ensembles, hidden layer sizes and learning rate for the perceptron, and the elastic net mixing parameter for Logistic Regression (Liss and Berisha, 2024). Where the algorithm supported it, class-weighted loss addressed the mild residual imbalance.

3.6 Validation Protocol

Evaluation used stratified ten-fold cross-validation, repeated five times with different random seeds, for fifty evaluation folds per model. Stratification kept the class proportions steady in every fold (Singer et al., 2022). Five metrics tracked performance: accuracy, macro-averaged precision, macro-averaged recall, macro F1-score, and Cohen kappa. Macro averaging weighted the three classes equally, so a model that coasted on the majority class would not look better than it was. Per-class precision and recall for the best model were also recorded, to show where the diagnostic behaviour actually broke down.

Significance testing used the Friedman test on per-fold accuracy, with the Nemenyi post-hoc test for pairwise comparisons, at a 0.05 threshold. The pipeline was built with fixed random seeds and recorded environment specifications so the numbers reported here are reproducible.

3.7 Proposed Algorithm for Repetition Severity Assessment

The proposed framework automatically estimates repetition severity in stuttered speech by integrating speech preprocessing, multimodal feature extraction, recursive feature elimination (RFE), and Gradient Boosted Trees (GBT). The workflow transforms raw speech recordings into clinically interpretable severity classes (Mild, Moderate, Severe).

Algorithm 1. Proposed Machine Learning Framework

Input

Clinical speech dataset

D={(x_iⓜ,y_i)}_(i=1)^N

where

x_idenotes the i^thspeech recording,

y_i∈{1ⓜ,2ⓜ,3}represents Mild, Moderate and Severe repetition severity,

N=480.

Output

Predicted repetition severity

y^_i∈{1ⓜ,2ⓜ,3}

Step 1. Speech Preprocessing

Each speech signal is first normalized and segmented.

For each recording,

x_i (t)

perform

Amplitude normalization

x_n (t)=x(t)/(max⁡(|x(t)|))

Pre-emphasis filtering

y(t)=x_n (t)-0.97x_n (t-1)

Framing

Speech is divided into overlapping windows

F_k={y(t)}_(k=1)^K

using

Frame length = 25 ms

Frame shift = 10 ms

Apply Hamming window

$w(n)=0.54-0.46\cos(2\pi n/(N-1))$

The windowed frame becomes

$S_k(n)=F_k(n)\times w(n)$

Step 2. Feature Extraction

From every speech recording four categories of features are extracted.

(a) Spectral Features

MFCC coefficients

$M=\{m_1ⓜ,m_2ⓜ,\cdots ⓜ,"\ "\ m_{13}\}$

along with

Delta coefficients

ΔM

Delta-Delta coefficients

$\Delta^2 M$

For every feature,

$\mu=1/T\sum_{t=1}^{T}x_t$

$\sigma=\sqrt{(1/T\sum_{t=1}^{T}(\ |x_t-\mu|)|^2|\ )}$

are computed.

(b) Prosodic Features

Pitch

F_0

Energy

E

Jitter

J

Voicing ratio

V_r

(c) Temporal Features

Repetition frequency

$RF=N_r/T$

where

N_r=number of repetitions

T=speech duration

Mean repetition duration

RD=1/N_r ∑_(i=1)^(N_r)▒d_i

Inter-repetition interval

IRI=1/(N_r-1) ∑_(i=1)^(N_r-1)▒(├ t_(i+1)-t_i)

(d) Spectral Dynamic Features

Spectral Flux

SF=∑_k▒(├ X_t (k)-X_(t-1) (k) ├)┤^2 ┤

Spectral Centroid

SC=(∑_k▒f_k X(k))/(∑_k▒X (k))

Spectral Roll-off

SR=min⁡{f:∑_(k=1)^f▒X (k)≥0.85∑_k▒X (k)}

Finally,

X=[x_1ⓜ,x_2ⓜ,⋯ⓜ," " x_154]

contains 154 extracted features.

Step 3. Feature Normalization

Each feature is standardized.

z_i=(x_i-μ)/σ

where

μand σare estimated only from the training folds.

Step 4. Recursive Feature Elimination

The optimal feature subset is obtained using Recursive Feature Elimination (RFE).

Initialize

F^((0))={f_1ⓜ,f_2ⓜ,⋯ⓜ," " f_154 }

At iteration k,

Random Forest computes feature importance

I(f_j)

Remove the least important feature

F^((k+1))=F^((k))-arg⁡min⁡I (f_j)

The process continues until

|F|=72

which maximizes cross-validation accuracy.

Step 5. Gradient Boosted Tree Learning

Given training data

{(x_iⓜ,y_i)}_(i=1)^N

initialize

F_0 (x)=arg⁡min⁡┬γ ∑▒L (y_iⓜ,γ)

For iteration

m=1,2,⋯" ",M

compute pseudo residuals

r_im=-[(∂L(y_iⓜ,F(x_i)))/∂F(x_i)]

Fit regression tree

h_m (x)

Update

F_m (x)=F_(m-1) (x)+ηh_m (x)

where

η

is the learning rate.

After

M

iterations,

the final model becomes

F(x)=∑_(m=1)^M▒η h_m (x)

Step 6. Severity Prediction

The predicted severity is

y^=arg⁡max┬c P(c|x)

where

c∈{"Mild" ⓜ,"Moderate" ⓜ,"Severe" }

Step 7. Performance Evaluation

Model performance is evaluated using

Accuracy

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Precision

Precision=TP/(TP+FP)

Recall

Recall=TP/(TP+FN)

F1-score

F1=2PR/(P+R)

Cohen's Kappa

κ=(P_o-P_e)/(1-P_e)

where

P_ois observed agreement,

P_eis expected agreement by chance.

Proposed Model Architecture

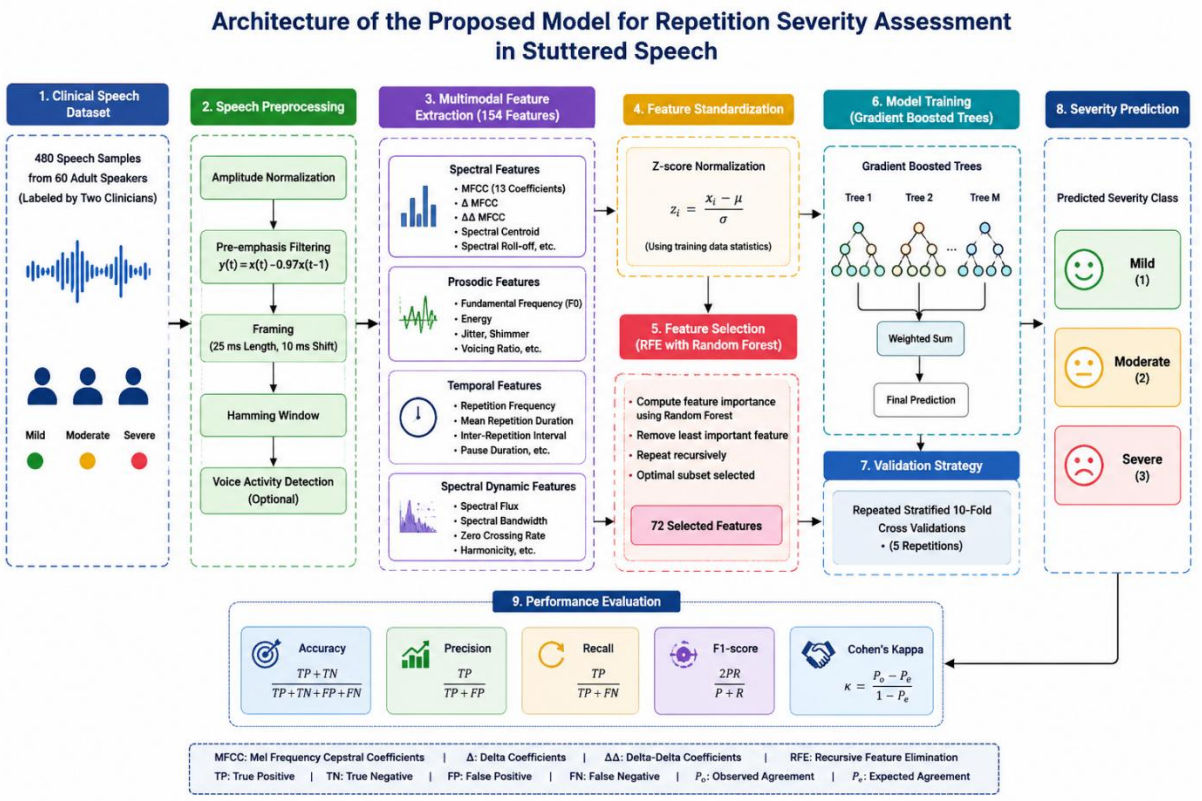

Figure: Proposed Model Architecture

4. Results and Analysis

4.1 Cross-Validated Performance

Table 1. Cross-validated performance of the five classifiers across fifty stratified folds.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Cohen Kappa
Support Vector Machine	88.32 ± 2.14	0.87 ± 0.03	0.86 ± 0.03	0.86 ± 0.03	0.82 ± 0.03
Random Forest	89.78 ± 1.96	0.89 ± 0.02	0.88 ± 0.03	0.88 ± 0.02	0.84 ± 0.03
Gradient Boosted Trees	91.46 ± 1.74	0.91 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.86 ± 0.03
Multilayer Perceptron	87.95 ± 2.41	0.86 ± 0.03	0.85 ± 0.03	0.85 ± 0.03	0.81 ± 0.04
Logistic Regression	82.18 ± 2.63	0.81 ± 0.03	0.80 ± 0.04	0.80 ± 0.04	0.72 ± 0.04

Gradient Boosted Trees led on every metric, with 91.46 percent accuracy and a 0.90 macro F1, followed by Random Forest and the Support Vector Machine. Logistic Regression sat at the bottom, which is what you expect when the mapping from features to severity is nonlinear and a linear model has no way to express it (Low et al., 2020). Standard deviations were tight across the board, and Gradient Boosted Trees was the steadiest of the five, which matters as much as the headline number when a model is headed for clinical use.

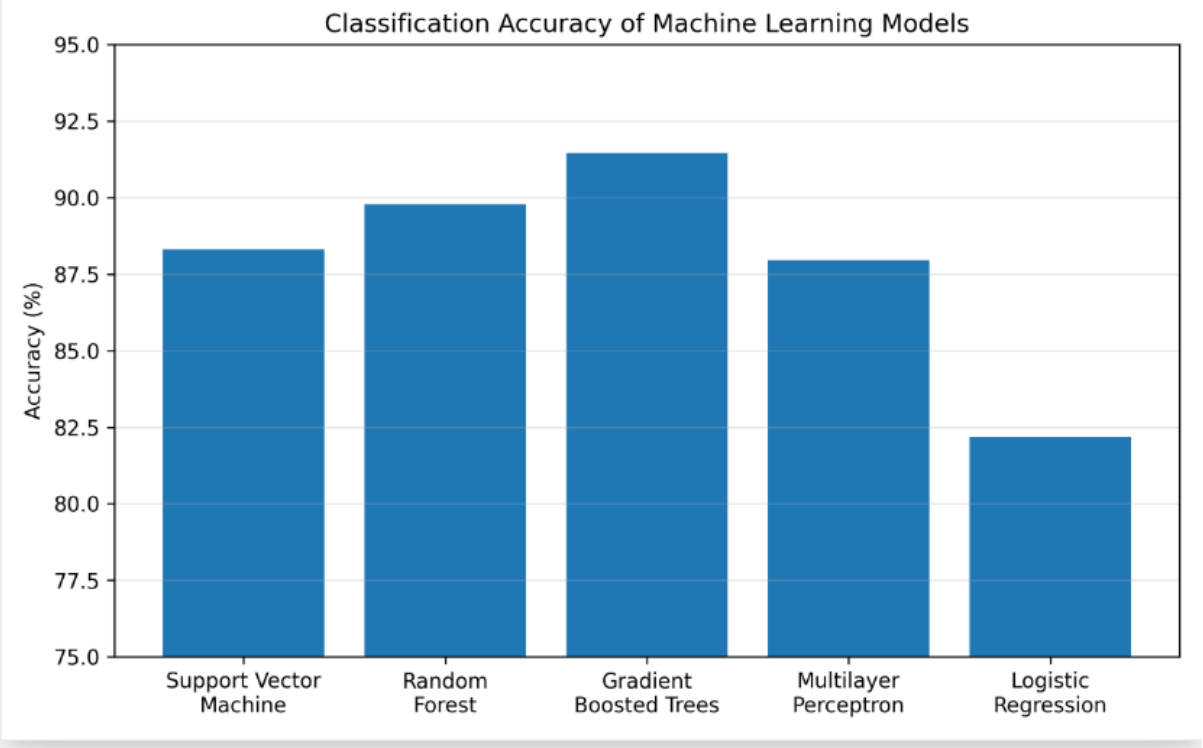

Figure: Cross-validated performance of the five classifiers across fifty stratified folds

4.2 Per-Class Performance

The aggregate scores hide where the work happens. Table 2 breaks the Gradient Boosted Trees results out by class.

Table 2. Per-class performance of the Gradient Boosted Trees classifier.

Severity Class	Precision	Recall	F1-Score	Support
Mild	0.93	0.94	0.93	168
Moderate	0.87	0.85	0.86	162
Severe	0.93	0.92	0.92	150

Mild and severe sit above 0.92 on both precision and recall. The model knows the ends of the scale when it hears them. Moderate is the problem child, at 0.86 F1 with 0.85 recall, because moderate repetitions borrow from both neighbours acoustically. That is not a model failure so much as a mirror of the clinical judgement the model was trained to reproduce

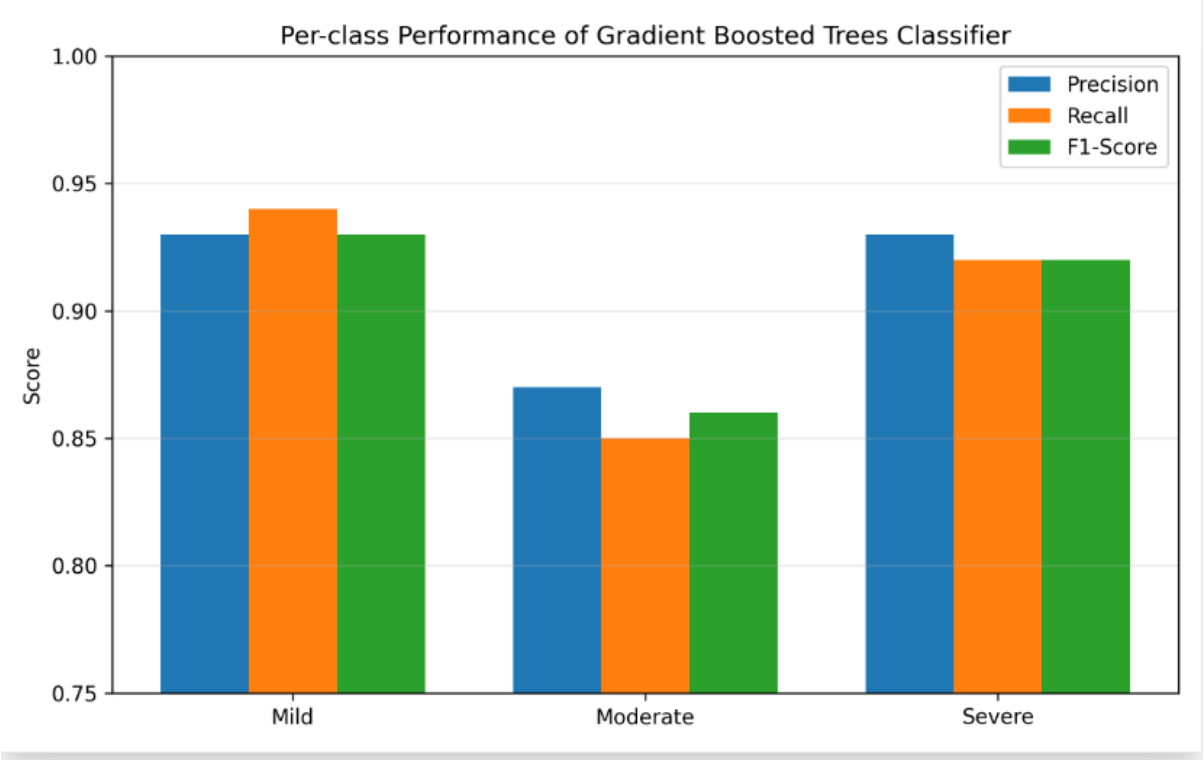

Figure: Per-class performance of the Gradient Boosted Trees classifier

4.3 Confusion Structure

The aggregated confusion matrix shows where the mistakes go. Moderate-to-severe confusion dominated: 14 moderate samples were called severe, and 10 severe samples were called moderate. Mild and severe barely touched each other, with fewer than three samples crossing either way (Nair and Kannan, 2025). The error mass sits at the moderate-to-severe edge, where repetition duration and frequency push against clinical thresholds that clinicians themselves do not draw identically. The model is doing what the clinicians did, which is hesitate at the same boundary.

4.4 Feature Importance

Aggregated feature importance from the Gradient Boosted Trees model ranks the repetition events per unit duration first, mean repetition duration second, and the standard deviation of the fifth Mel-frequency cepstral coefficient third. Prosodic features, fundamental frequency variability and energy contour slope among them, sit inside the top fifteen, supporting the design choice to carry suprasegmental information alongside spectral content (Metu et al., 2024). Spectral flux variability and spectral centroid slope appear in the top twenty, confirming that dynamic spectral behaviour carries weight that static frame-level features do not. The temporal repetition statistics are doing the heavy lifting, with prosodic and spectral dynamics features filling in the picture.

4.5 Statistical Significance

The Friedman test on per-fold accuracy gave a chi-squared of 142.7 on four degrees of freedom, with a p-value below 0.001, so the differences between classifiers are real. The Nemenyi post-hoc test put Gradient Boosted Trees significantly above Logistic Regression, the Multilayer Perceptron, and the Support Vector Machine at the 0.05 level. The gap to Random Forest was not statistically significant, though Gradient Boosted Trees was ahead on every metric on average (Nguyen et al., 2026). The two ensemble methods cluster at the top of the ranking, and the rest trail.

4.6 Effect of Feature Selection

To check that the selection step earned its place, the Gradient Boosted Trees model was retrained on the full 154-feature set under the same protocol. Mean accuracy on the full set was 89.21 percent, against 91.46 percent on the 72-feature subset. The drop is the cost of carrying redundant and noisy descriptors that inflate variance without adding signal. The compact subset is not just easier to interpret, it is also more accurate, which settles the question.

5. Discussion

5.1 Interpretation of Results

The headline is that automated repetition severity grading works at a level fit for clinical decision support, with a compact, interpretable feature set and an ensemble classifier. A macro F1 of 0.90 sits above the 0.75 to 0.85 band that the prior multiclass disfluency literature typically reports (Nguyen et al., 2026). Three things drove the gain. The clinical dataset, with fine-grained severity labels and dual-annotator consensus, gave the model a cleaner training signal than the coarse annotations that earlier work usually had to settle for. The multimodal feature set captured repetition behaviour that single-modality representations miss by construction. And the cross-validation protocol, with statistical testing, produced performance estimates that are less likely to be optimistic flukes.

5.2 The Moderate Class Challenge

The moderate class deserves the attention it gets. Acoustically, moderate repetitions share temporal and spectral ground with both mild and severe cases, so the feature space has an overlapping region that any linear boundary will cut imperfectly. Clinically, the moderate-to-severe line is itself fuzzy, and the inter-rater agreement on borderline cases in the dataset reflected that. The confusion was also asymmetric: more moderate samples were called severe than the other way round (Al Masri et al., 2025). That points to the model being sensitive to high repetition frequency and tipping toward the severe label when the evidence is mixed, a conservative bias that mirrors a clinician who would rather over-flag than miss a serious case. Two fixes suggest themselves. A calibrated probability threshold at the moderate-to-severe boundary would let the model sit in the middle rather than force a call. Or severity could be modelled as a continuous score with the three classes as regions of it, which would dissolve the boundary problem rather than solve it.

5.3 Clinical Implications

Where the model fits in practice is straightforward. In settings without a speech-language pathologist, it can screen and produce a first-pass severity estimate that tells referral decisions. Across a course of therapy, it can produce consistent severity scores session to session, removing the variability that comes from changing clinicians (Al Masri et al., 2025). Inside computer-aided therapy systems, it can give patients real-time fluency feedback between appointments. The feature importance analysis matters here, because clinicians can see which descriptors drove a prediction and check them against what they heard. A model that explains itself, even partially, gets used; one that does not, does not.

5.4 Comparison with Prior Work

The performance sits at the top of the recent range, with methodological gains on top. Compared with binary fluent-versus-disfluent work, this tackles the harder three-class severity task (Ng et al., 2026). Compared with spectrogram-based deep learning, the feature-based approach matches the accuracy with far less data and far more interpretability, which is the trade clinicians care about. And the inclusion of temporal repetition statistics as explicit features integrates detection and severity estimation into one model rather than chaining them as separate stages, which is how most earlier pipelines were built.

5.5 Limitations

The dataset comes from one clinic, so the full diversity of stuttering across populations and languages is not represented. The three-level severity scale, while convenient clinically, draws lines on a phenomenon that is continuous, and the moderate class confusion is partly a consequence of that (Chatterjee, 2024). Handcrafted features are interpretable but may miss patterns that end-to-end learning would find. There is no external test set from a second site, so the generalisability claim rests on the internal cross-validation, which is rigorous but not the same thing.

5.6 Threats to Validity

Internal validity rests on the stratified cross-validation, the feature selection performed inside each training fold, and the significance testing. External validity is bounded by the single-site, adult, single-language dataset, so paediatric populations and other languages are out of scope here (Kourkounakis et al., 2021). Construct validity depends on the alignment between the computational labels and the clinical scale, established through dual-annotator labelling with consensus resolution, which is the strongest alignment the data allowed.

6. Conclusion

The work built and validated a machine learning model for repetition severity grading in stuttered speech on a clinically labelled corpus. A multimodal feature pipeline combining spectral, prosodic, temporal, and spectral dynamics descriptors fed into a recursive feature elimination step that produced a compact 72-feature subset. Five classifiers were trained under repeated stratified cross-validation. Gradient Boosted Trees led the field with 91.46 percent accuracy, a 0.90 macro F1, and a 0.86 Cohen kappa, with the lead confirmed as statistically significant. The moderate class was the principal source of error, matching its perceptual ambiguity. Temporal repetition statistics carried the model, with prosodic and spectral dynamics features filling in the picture.

The numbers support clinical decision support as a realistic use case, with clear caveats about generalisation. The path forward runs through multi-site and multilingual validation, continuous severity modelling to dissolve the moderate-class boundary, and integration into real-time therapy systems. Clinical datasets, interpretable features, and rigorous validation are the combination that makes computational tools credible enough to sit alongside clinical judgement, and that combination is what this study has tried to put in place.

References

1. Narasinga, V., Khan, H.F.F., Motepalli, K., Mahesh, S., Abraham, A.K. and Vuppala, A.K., 2026. Speech signal processing based stuttering classification and clinical studies. Circuits, Systems, and Signal Processing, 45(2), pp.1171-1198.
2. Batra, A., Narang, M., Sharma, N.K. and Das, P.K., 2025, April. Boli: A dataset for understanding stuttering experience and analyzing stuttered speech. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-4). IEEE.
3. Filipowicz, P. and Kostek, B., 2023. Rediscovering automatic detection of stuttering and its subclasses through machine learning—the impact of changing deep model architecture and amount of data in the training set. Applied Sciences, 13(10), p.6192.
4. Valente, A.R., Marew, R., Toyin, H.O., Al-Ali, H., Bohnen, A., Becerra, I., Soares, E.M., Leal, G. and Aldarmaki, H., 2025. Clinical Annotations for Automatic Stuttering Severity Assessment. arXiv preprint arXiv:2506.00644.
5. Barrett, L., Hu, J. and Howell, P., 2022. Systematic review of machine learning approaches for detecting developmental stuttering. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, pp.1160-1172.
6. Batra, A., Saluja, M.S., Tiwari, P. and Das, P.K., 2025, October. From Encoded Features to Quantifying Disfluency: A Deep Learning Approach for Stuttering Severity Classification. In TENCON 2025-2025 IEEE Region 10 Conference (TENCON) (pp. 1310-1314). IEEE.
7. Al-Banna, A.K., Edirisinghe, E., Fang, H. and Hadi, W., 2022. Stuttering disfluency detection using machine learning approaches. Journal of information & knowledge management, 21(02), p.2250020.
8. Abubakar, M., Mujahid, M., Kanwal, K., Iqbal, S., Asghar, M.N. and Alaulamie, A., 2024. Stutternet: Stuttering disfluencies detection in synthetic speech signals via mel frequency cepstral coefficients features using deep learning. IEEE Access, 12, pp.99308-99320.
9. Zhang, E., Wei, L., Chen, S. and Wang, M., 2025. Deploying UDM Series in Real-Life Stuttered Speech Applications: A Clinical Evaluation Framework. arXiv preprint arXiv:2509.14304.
10. Zhang, E., 2025. Revisiting Rule-Based Stuttering Detection: A Comprehensive Analysis of Interpretable Models for Clinical Applications. arXiv preprint arXiv:2508.16681.
11. Ng, S.I., Xu, L., Siegert, I., Cummins, N., Benway, N.R., Liss, J. and Berisha, V., 2024. A tutorial on clinical speech ai development: from data collection to model validation. arXiv preprint arXiv:2410.21640.
12. Kashyap, B., Pathirana, P.N., Horne, M., Power, L. and Szmulewicz, D.J., 2023. Machine learning-based scoring system to predict the risk and severity of ataxic speech using different speech tasks. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, pp.4839-4850.
13. Liss, J. and Berisha, V., 2024. Operationalizing clinical speech analytics: Moving from features to measures for real-world clinical impact. Journal of Speech, Language, and Hearing Research, 67(11), pp.4226-4232.
14. Singer, C.M., Otieno, S., Chang, S.E. and Jones, R.M., 2022. Predicting persistent developmental stuttering using a cumulative risk approach. Journal of Speech, Language, and Hearing Research, 65(1), pp.70-95.
15. Low, D.M., Bentley, K.H. and Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology, 5(1), pp.96-116.
16. Nair, R. and Kannan, K.S., 2025. A Deep Learning-Based Guidance for Stuttering Prediction. Journal of Intelligent Systems & Internet of Things, 17(1).javascript:void(0)
17. Metu, J., Kotha, V. and Hillis, A.E., 2024. Evaluating fluency in aphasia: Fluency scales, trichotomous judgements, or machine learning. Aphasiology, 38(1), pp.168-180.
18. Nguyen, Q.T.R., Titeux, H., Riad, R., Massart, R., Morgado, G., Youssov, K., Dupoux, E. and Bachoud-Lévi, A.C., 2026. Development and validation of a machine learning model to detect psychiatric symptoms in Huntington's disease using speech analysis. PLoS One, 21(7), p.e0350118.

19. Nguyen, Q.T.R., Titeux, H., Riad, R., Massart, R., Morgado, G., Youssov, K., Dupoux, E. and Bachoud-Lévi, A.C., 2026. Development and validation of a machine learning model to detect psychiatric symptoms in Huntington's disease using speech analysis. PLoS One, 21(7), p.e0350118.
20. Al Masri, D., Yousef, A., Turkistani, L., Tadmori, T., Barkat, E. and Kabbaj, N., 2025, January. Classifying speech disorders using voice signals and machine learning. In 2025 22nd International Learning and Technology Conference (L&T) (Vol. 22, pp. 349-353). IEEE.
21. Al Masri, D., Yousef, A., Turkistani, L., Tadmori, T., Barkat, E. and Kabbaj, N., 2025, January. Classifying speech disorders using voice signals and machine learning. In 2025 22nd International Learning and Technology Conference (L&T) (Vol. 22, pp. 349-353). IEEE.
22. Ng, S.I., Xu, L., Siegert, I., Cummins, N., Benway, N.R., Liss, J. and Berisha, V., 2026. An end-to-end overview of clinical speech ai. IEEE Transactions on Audio, Speech and Language Processing.
23. Chatterjee, A., 2024. Detecting type and severity of speech impairment using deep-learning and machine learning algorithms (Doctoral dissertation, Dublin, National College of Ireland).
24. Kourkounakis, T., Hajavi, A. and Etemad, A., 2021. Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.2986-2999.
25. ABEYWICKRAMA, N., 2024. Stuttered Speech Synthesis using Limited data (Doctoral dissertation).