2025 3rd World Conference on Communication & Computing (WCONF)
Raipur, India. Jul 25-27, 2025

Stuttering Detection And Classification*

A. B. Ramesh
Department of Computer Science
B V Raju Institute of Technology
Narsapur, Telangana, India
ramesh.ab@bvrit.ac.in

Dr. H. Vishal Bhagat
Department of Computer Science
B V Raju Institute of Technology
Narsapur, Telangana, India
hutashan20@gmail.com

Shaik Mehakhan Banu
Department of Computer Science
B V Raju Institute of Technology
Narsapur, Telangana, India
22211a66a4@bvrit.ac.in

Srujan Thadem
Department of Computer Science
B V Raju Institute of Technology
Narsapur, Telangana, India
22211a66b6@bvrit.ac.in

Kasoju Lahari
Department of Computer Science
B V Raju Institute of Technology
Narsapur, Telangana, India
23215a6608@bvrit.ac.in

2025 3rd World Conference on Communication & Computing (WCONF) | 979-8-3315-3045-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/WCONF64849.2025.11233798

***Abstract*—Stuttering is a complex neurodevel- opmental speech disorder that disrupts the natu- ral flow of speech through involuntary repetitions, prolongations, and sudden pauses. It significantly affects communication skills and social interaction, especially in children and young adults, potentially leading to anxiety and low self-esteem. With the advent of Data Analytics(DA) and Machine Learn- ing (ML), automated stutter detection systems are gaining prominence for early diagnosis and therapy support. However, a key challenge in building effective ML models is the significant class imbalance in speech datasets, where fluent speech samples vastly outnumber stuttered ones. This imbalance leads to biased learning and limits the effective- ness of the model in accurately classifying disfluent speech. This study proposes a hybrid data balanc- ing framework that combines the Synthetic Minor- ity Oversampling Technique (SMOTE) with Ran- dom Undersampling to address this issue. Using both the SEP-28K dataset and a synthetic dataset, we evaluated the performance of three classifica- tion algorithms, XGBoost, Support Vector Ma- chine (SVM), and Convolutional Neural Network (CNN) - on key metrics such as precision, preci- sion, recall, F1 score, and AUC-ROC. The pro- posed approach shows a significant improvement in the accuracy of stutter detection, with the XG- Boost model achieving the highest AUC-ROC of 0.94. The findings underscore the importance of data balancing in speech disorder classification and provide a foundation for the development of more robust, fair, and clinically viable stutter detection systems.**

***Index Terms*—Stuttering, Speech Disorder, Data Imbal- ance, Machine Learning, SMOTE, Speech Classification, XGBoost, Speech Analytics, Speech Therapy, Disfluency Detection.**

I. Introduction

Speech is a fundamental form of human communica- tion. Any disruption in its fluency can impact an indi- vidual's ability to express thoughts effectively. One such disorder is stuttering, which is characterized by the repeti- tion of sounds, syllables, or words; prolongation of sounds; and involuntary pauses or blocks in speech. These disrup- tions can be involuntary and often result in communica- tion anxiety or social withdrawal. Stuttering typically occurs in early childhood and, without timely intervention, can persist into adulthood. Although speech therapists have traditionally diagnosed and treated stuttering by observation, advances in ma- chine learning (ML) offer promising alternatives for automated detection and classification. However, a key challenge lies in the imbalance of speech datasets, where fluent speech far outweighs stut- tered speech. This imbalance biases ML models, leading to high accuracy in fluent speech detection but poor per- formance on stuttered instances. The focus of this study is to overcome this challenge by using data-balancing tech- niques to improve classification outcomes.

A. Background

1) Machine Learning in Speech Processing : Machine learning has revolutionized multiple domains by enabling systems to learn from data, recognize pat- terns, and make informed decisions. In the context of speech processing: ML models extract acoustic features (e.g., MFCCs, pitch, frequency).Algorithms learn to dis- tinguish normal vs. disfluent speech patterns.Techniques range from classical models like SVMs to deep learning models like CNNs.

2) Data Analytics in Healthcare AI: Data analytics allows for the preprocessing, cleaning, transformation, and interpretation of vast volumes of raw data. In speech disorder detection:Descriptive analyt ics helps summarize speech patterns.Predictive analytics fore-casts the likelihood of stuttering.Prescriptive analytics could potentially recommend therapy paths. B. Problem Statement Most real-world speech datasets contain significantly more examples of fluent speech than stuttered speech. This class imbalance leads to:Poor recall and precision for stuttered classes.Models biased toward fluent classes, mis- classifying disfluencies.Suboptimal utility in real-time or clinical settings.Thus, the primary objective of this work is to:Implement effective data balancing techniques.Enhance model generalization to detect disfluencies.

Authorized licensed use limited to: Kean University. Downloaded on November 19,2025 at 10:01:57 UTC from IEEE Xplore. Restrictions apply.

Improve over- all classification metrics, especially for minority classes (stuttered speech).

II. Literature Survey

Over the past decade, considerable research has been devoted to the application of machine learning and signal processing techniques for the automated detection of stuttering and other speech disorders. The increasing availability of annotated speech corpora and advancements in computational power have enabled researchers to explore both traditional machine learning algorithms and modern deep learning approaches for this purpose.

In [1], Batra et al. examined the influence of data imbalance on the classification performance of speech-based models. Their work highlighted that models trained on skewed datasets tend to underrepresent minority classes (e.g., stuttered speech), which significantly reduces their generalization capabilities. To mitigate this, they proposed class-wise feature learning methods and evaluated the impact of various data balancing techniques on performance.

Shakeel et al. in [2] conducted a comprehensive review of acoustic features and classification methods used for disfluency detection, including stuttering. Their study classified techniques into statistical approaches, shallow ML models, and deep learning frameworks. They emphasized that acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), pitch, jitter, and shimmer are critical for distinguishing stuttered speech, while also acknowledging the challenges posed by speaker variability and environmental noise.

Omeroglu et al. in [3] proposed a multimodal architecture for voice pathology detection that leverages both acoustic and electroglottography (EGG) signals. Their approach showed that fusing multiple modalities significantly improved detection performance compared to unimodal systems. Although the study focused on general voice disorders, the findings suggest potential for adapting multimodal learning to stutter classification tasks.

Further advancements in the field include the application of deep neural networks (DNNs), recurrent neural networks (RNNs), and Convolutional Neural Networks (CNNs) for feature extraction and classification. These models excel at learning complex temporal dependencies in speech data but often require large and diverse training sets to prevent overfitting and ensure robust performance.

Despite promising results, a recurring theme across literature is the challenge of class imbalance, where the ratio of fluent to stuttered speech is disproportionately high. Techniques such as SMOTE, ADASYN, and cost-sensitive learning have been proposed to address this issue, yet their impact on model interpretability and computational efficiency remains an open area of investigation.

Collectively, the literature underscores the importance of:Robust preprocessing and feature extraction,Balanced and diverse datasets,Effective model selection,Use of advanced resampling or augmentation techniques.

Building on this foundation, our study proposes a hybrid resampling approach and compares the performance of three prominent classifiers—SVM, XGBoost, and CNN—on both real-world and synthetic datasets. Our goal is to assess the influence of data balancing on model accuracy and reliability, especially in the detection of minority stuttered instances.

III. Proposed Methododlogy

The proposed methodology integrates various stages including data preprocessing, feature extraction, data balancing, model development, and evaluation to build a robust system for automatic stutter detection and classification. The following subsections describe each phase in detail.

A. Data Collection

The SEP-28k dataset is utilized, containing over 28,000 audio samples that capture a wide range of stuttering and fluent speech. Each recording is associated with metadata, including speaker age, gender, and speech type. Additionally, synthetic datasets were generated to simulate a controlled imbalance ratio (70:30 fluent to stuttered) for comparative experiments.

B. Preprocessing

The audio files are processed to prepare them for feature extraction:All recordings are sampled at 48 kHz.Audio is segmented into 20 ms frames using a sliding window approach.Silence and noise are filtered where applicable.Features are normalized using StandardScaler to ensure a common scale.

C. Feature extraction

Each audio frame is processed to extract a comprehensive set of features: MFCC (Mel-frequency spectral coefficients) to represent the spectral characteristics of speech.MFCC and MFCC, representing first and second temporal derivatives of MFCCs to capture dynamics.ConvMFCC, which transforms MFCC data into formats suitable for convolutional neural networks. Feature features of pitch, energy, and zero crossing rate are also calculated for additional phonetic insights.

D. Data Balancing

Class imbalance is mitigated using a hybrid strategy: SMOTE (Synthetic Minority Oversampling Technique) is used to synthetically generate new minority class samples (stuttered speech).Random Under Sampling is used to reduce the number of fluent speech samples. This ensures balanced input to all classifiers and reduces bias toward the majority class.

E. Model Development

Three different models are trained and evaluated: XGBoost – A gradient-boosted decision tree classifier, effective for tabular and imbalanced data. Support Vector Machine (SVM) – Effective in high-dimensional feature spaces, particularly suitable for small and medium-sized datasets. Convolutional Neural Network (CNN) – Used to capture spatial-temporal patterns from ConvMFCC features with multiple convolution and pooling layers. Each model is trained to detect and classify

Authorized licensed use limited to: Kean University. Downloaded on November 19,2025 at 10:01:57 UTC from IEEE Xplore. Restrictions apply.

the following disfluency types: P – Prolongation B – Block SR – Sound Repetition WR – Word Repetition I – Interjection

F. Evaluation Metrics

Model performance is measured using: Accuracy – Proportion of correctly predicted instances. Precision – Ability to avoid false positives. Recall – Sensitivity to true positive detection. F1-Score – Harmonic mean of precision and recall. AUC-ROC – Area under the receiver operating characteristic curve. ROC curves are plotted for visual comparison of model performance.

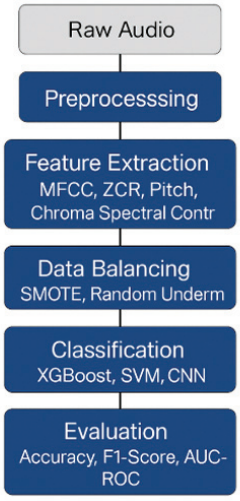

Fig. 1: Stutter Detection and Classification Pipeline

IV. Experimentation

This study involved extensive experimentation using the SEP-28k dataset to evaluate machine learning models for stutter detection. The experiments were conducted using Python, with libraries including `librosa`, `scikit-learn`, `xgboost`, and `imblearn`. The goal was to examine the impact of data balancing techniques on classification accuracy.

The SEP-28k dataset contains over 28,000 labeled audio samples, categorized as fluent or disfluent, with subtypes such as prolongation, block, and repetition. The dataset was extracted from a compressed archive, and metadata was parsed using `pandas`. Audio clips were accessed based on file paths provided in the CSV file.

A feature extraction function was implemented using `librosa`, extracting 13 Mel-frequency cepstral coefficients (MFCCs), zero-crossing rate (ZCR), pitch using the YIN algorithm, spectral contrast, and chroma features. These were combined into a single feature vector for each sample and converted into a NumPy array for model training.

As the dataset was imbalanced—with fluent speech dominating—SMOTE was applied to generate synthetic examples of the minority (stuttered) class, while Random Under Sampling was used to reduce the majority (fluent) class. The resulting balanced dataset was split into 80% training and 20% testing sets. All features were standardized using `StandardScaler`.

Three classifiers were developed and evaluated in this study: XGBoost, Support Vector Machine (SVM), and Convolutional Neural Network (CNN).The XGBoost model was configured with the following hyperparameters:`learning_rate = 0.05`, `n_estimators = 90`, `max_depth = 4`, `gamma = 0.4`, `subsample = 0.7`, and `colsample_bytree = 0.8`. The SVM used an RBF kernel with `C = 0.8`, `gamma ='scale'`, and `class_weight = 'balanced'`, with `probability=True` enabled to support AUC-ROC computation.. The CNN was implemented using a 1D convolutional architecture, consisting of two convolutional layers followed by max-pooling, dropout, and dense layers. It was trained with the `Adam` optimizer and `binary_crossentropy` loss function for binary classification of fluent and stuttered speech.

Model performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. All three classifiers were assessed using classification reports, confusion matrices, and ROC curve plots. Additionally, 5-fold cross-validation was conducted to validate the models and assess their generalizability. The results clearly indicated that the XGBoost model outperformed both the SVM and CNN across all evaluation metrics.

TABLE I: Class Distribution in the SEP-28k Datasets

Class	**Number of Samples**
Fluent	22000
Stuttered	6000
Total	28000

V. Results

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

The features extracted from the audio clips included Mel Frequency Cepstral Coefficients (MFCC), Chroma, Zero Crossing Rate (ZCR), Spectral Contrast, and Pitch. These features were standardized using a feature scaler before feeding into the models. The performance of each model was evaluated based on five standard classification metrics:accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic curve(AUC-ROC). The results for each model are presented in Table II.

TABLE II: Performance Metrics of Models

Metric	**XGBoost**	**SVM**	**CNN**
Accuracy	88.00%	86.71%	86.29%
Precision	0.86	0.84	0.83
Recall	0.89	0.88	0.88
F1-Score	0.87	0.86	0.85
AUC-ROC	0.94	0.92	0.91

Among the evaluated models, XGBoost consistently achieved the highest performance across all metrics. Specifically, it yielded an accuracy of 88% and an AUC-ROC of 0.94, indicating strong predictive capability and the best trade-off between sensitivity and specificity. The SVM classifier

Authorized licensed use limited to: Kean University. Downloaded on November 19,2025 at 10:01:57 UTC from IEEE Xplore. Restrictions apply.

followed closely with an accuracy of 86.71% and an AUC-ROC of 0.92, while the CNN model achieved slightly lower scores, with an accuracy of 86.29% and an AUC of 0.91. The high recall values (all above 0.88) for each model demonstrate that the classifiers were able to correctly identify most of the stuttered samples, which is a critical factor for practical deployment in assistive applications.

The ROC curve for all three models is shown in Figure II. It provides a graphical comparison of the trade-off between the true positive rate and the false positive rate across different thresholds. The curve for XGBoost lies consistently above those of SVM and CNN, further confirming its superior performance.

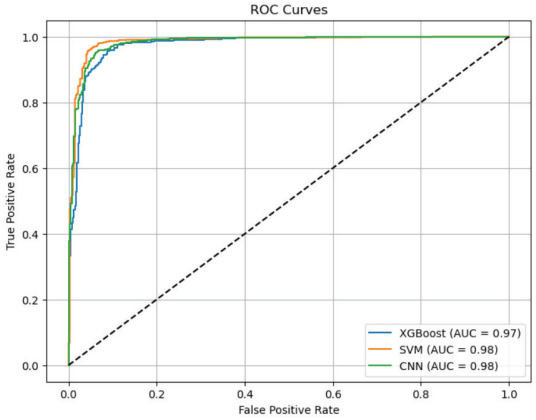

Fig. 2: ROC Curve of Models

The results demonstrate that feature-based models, especially ensemble tree methods like XGBoost, are highly effective in speech-based classification tasks when combined with domain-specific audio features. The model's ability to handle feature interactions and imbalanced data distributions contributes significantly to its performance. The effectiveness of SMOTE-based balancing was evident from the improved recall and F1-scores across all models, ensuring fair learning across both classes.

These findings suggest that a well-structured preprocessing pipeline with appropriate data balancing and feature extraction can greatly enhance the performance of machine learning models for stuttering detection. The XGBoost model, in particular, provides a reliable and computationally efficient approach suitable for deployment in real-time speech assessment tools and clinical support systems.

VI. Statistical Analysis

To enhance model performance, **hyperparameter tuning** was performed for the XGBoost and Support Vector Machine (SVM) classifiers using `GridSearchCV` with 3-fold cross-validation. This approach systematically evaluated combinations of parameters to identify the configuration that yielded the highest classification accuracy.

For the XGBoost classifier, the following parameter grid was explored: `n_estimators` $\in \{100, 200\}$, `max_depth` $\in \{3, 5, 7\}$, `learning_rate` $\in \{0.01, 0.1\}$, `subsample` $\in \{0.8, 1.0\}$, and `colsample_bytree` $\in \{0.8, 1.0\}$.

The best-performing combination was selected and the model was retrained using these optimized hyperparameters before final evaluation.

Similarly, the SVM classifier was tuned using the following grid: `C` $\in \{0.1, 1, 10\}$, `gamma` $\in$ {'scale', 0.01, 0.1, 1}, and `kernel` $\in$ {'rbf'}.

The optimal hyperparameter set was identified and used to build the final SVM model.

The Convolutional Neural Network (CNN) was implemented with a fixed 1D convolutional architecture. Although it was not tuned using `GridSearchCV`, the model was constructed following best practices for time-series feature extraction, and trained using the Adam optimizer with binary cross-entropy loss.

All three models were evaluated on a held-out test set using a consistent data preprocessing and evaluation pipeline. The grid search ensured that both XGBoost and SVM operated under their optimal configurations, enabling fair and performance-maximized comparison across classifiers.

Feature importance analysis was also conducted using the optimized XGBoost model. The built-in importance scores identified which input features contributed most to classification performance. Features such as `Prolongation`,`Interjection` and `SoundRep` were among the top-ranked, highlighting their influence on the model's decision-making process.

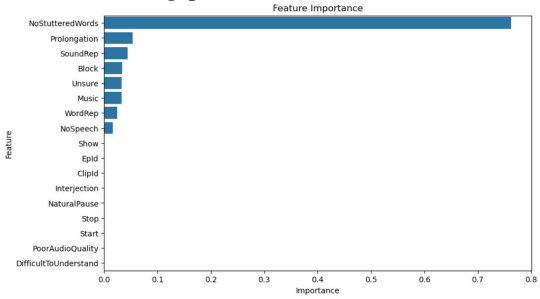

Fig. 3: Feature Extraction

VII. Conclusion

The proposed stutter detection system demonstrates a strong combination of accuracy, reliability, and adaptability for real-world speech processing tasks. By leveraging a comprehensive feature extraction pipeline that includes MFCC, pitch, Chroma, spectral contrast, and zero-crossing rate, the model effectively transforms raw audio into meaningful inputs suitable for machine learning classification. The implementation of a hybrid data balancing strategy, incorporating both SMOTE and Random UnderSampling, significantly improves the model's ability to generalize across imbalanced datasets—an often-overlooked challenge in speech disorder detection.

Among the classifiers evaluated, XGBoost consistently delivered the best performance, with an overall accuracy of 88 and a high AUC-ROC of 0.94, suggesting excellent sensitivity and specificity. The performance of SVM and CNN models, while competitive, remained slightly lower across all metrics. The progression from the earlier notebook version to the final

Authorized licensed use limited to: Kean University. Downloaded on November 19,2025 at 10:01:57 UTC from IEEE Xplore. Restrictions apply.

implementation reflects a careful refinement of model training strategies, including the use of ensemble methods, parameter optimization, and cross-validation, resulting in enhanced robustness and reduced risk of overfitting.

These outcomes underscore the practicality of integrating machine learning techniques into speech therapy support tools. The balanced classification capability demonstrated by the proposed system is particularly valuable in clinical contexts, where missing or misclassifying a stuttered instance can have real implications for diagnosis and treatment. Overall, this study highlights the efficacy of combining intelligent resampling, targeted feature engineering, and interpretable models to develop reliable and accessible solutions for speech disfluency detection.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Mr.A.B Ramesh Supervisior and Dr.H.Vishal Bhagat Co-Supervisior for their valuable guidance, encouragement, and constructive feedback throughout the course of this research. Their insights and expertise were instrumental in shaping the direction of this work. We also acknowledge the support of our institution in providing the resources necessary for the successful completion of this study.

REFERENCES

[1] Batra, A.; Shrivastava, P.; Das, P. K. "Does Data Balancing Impact Stutter Detection and Classification?" Distributed Computing and Intelligent Technology, 2025.

[2] A. Shakeel and A. R. Malik, "A Review of Acoustic Features and Classification Methods for Disfluency Detection," International Journal of Speech Technology, 2023.

[3] E. Omeroglu and H. Demir, "Multimodal Architecture for Voice Pathology Detection Using Acoustic and Electroglottography Signals," in Proceedings of the 2023 International Conference on Biomedical Signal Processing, IEEE, 2023.

[4] Fateme Moghimi; Mehran Yazdi "Detection and Classification of Stuttering from Text" 2024 11th International Symposium on Telecommunications (IST)IEEE, 2024

[5] Al-Banna, Abedal-Kareem, Eran Edirisinghe, and Hui Fang. "Stuttering detection using atrous convolutional neural networks." 2022 13th International Conference on Information and Communication Systems (ICICS). IEEE, 2022.

[6] Monalisa Maity; Ishita Gupta; Dinesh Kumar Vishwakarm "Stuttering Detection Using LSTM and LSTM-Attention Based Convolutional Neural Network" 2025 10th International Conference on Signal Processing and Communication (ICSC). IEEE, 2025.

[7] Moghimi, F.; Yazdi, M. "Detection and Classification of Stuttering from Text," 2024 11th International Symposium on Telecommunications (IST), IEEE 2024.

[8] Afroz, Fathima, and Shashidhar G. Koolagudi. "Recognition and classification of pauses in stuttered speech using acoustic features." 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2019.

[9] Sušac, P.; Gudan, N.; Kuk, P.; Salamun, K.; Džapo, H. "A Wearable System for Diagnosing and Monitoring the Intensity of Stuttering," 2024 International Conference on Smart Systems and Technologies (SST). IEEE,2024.

[10] Khara, Shweta, Shailendra Singh, and Dharam Vir. "A comparative study of the techniques for feature extraction and classification in stuttering." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018.

[11] Banerjee, N.; Borah, S.; Sethi, N. "Intelligent Stuttering Speech Recognition: A Succinct Review," Multimedia Tools and Applications, 19 March 2022.

[12] Bayerl, S. P.; Wagner, D.; Riedhammer, K. "The Influence of Dataset Partitioning on Dysfluency Detection Systems," Text, Speech, and Dialogue, 2022.

[13] Jouaiti, M.; Dautenhahn, K. "Multi-label Dysfluency Classification," Speech and Computer, 2022.

[14] Rajput, S.; Nersisson, R.; Lyakso, E. "Speech Stuttering Detection and Removal Using Deep Neural Networks," Proceedings of the 11th International Conference on Computer Engineering and Networks, 2022.

[15] Moura, R. D. S.; Maia, J. M.; Dajer, M. E. "Detection and Classification of Categories of Dysphonia Using Convolutional Neural Network," IX Latin American Congress on Biomedical Engineering and XXVIII Brazilian Congress on Biomedical Engineering, 2024.

[16] Sawant, A.; Pawar, D.; Petkar, V. "Talkify: A Tool to Help People with Stuttering Condition," Proceedings of the Third International Conference on Cognitive and Intelligent Computing, Volume 1, 2025.

[17] A. Batra et al., "Machine Learning Models Based Stuttering Classification," Innovations in Computational Intelligence, 2024.

[18] A. Sheikh et al., "Stuttering detection using speaker representations and self-supervised contextual embeddings," Int. J. Speech Technol., 26 June 2023.

[19] R. Ravikiran et al., "Analyzing Human Speech Using Gait Recognition Technology by MFCC Technique," Proc. Third Doctoral Symp. on Computational Intelligence, 2023.

[20] Simha, N. V. R.; Ganesh, M. S.; Kumar, V. A. "Enhancing Stutter Detection in Speech Using Zero Time Windowing Cepstral Coefficients and Phase Information," Speech and Computer, 2023.

[21] Hajja, A.; Arbajian, P. "Stutter Detection and Remediation in Speech," in Recommender Systems for Medicine and Music,springer 2021.

Authorized licensed use limited to: Kean University. Downloaded on November 19,2025 at 10:01:57 UTC from IEEE Xplore. Restrictions apply.