SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

An Optimized Deep Learning Model For Automatic Detection Of Stuttering Disfluencies In Speech Signals

Mrs. Rajeswary Nair[1] , K.S Kannan[2]

[1]Research Scholar, Kalasalingam Academy of Research and Education, Tamil Nadu rajeswarynrphd@gmail.com
[2]Associate Professor, Kalasalingam Academy of Research and Education, Tamil Nadu saikannan2012@gmail.com
Corresponding author mail id: rajeswarynrphd@gmail.com

Abstract
Stuttering disfluencies, marked by disruptions such as repetitions, prolongations, and speech blocks, significantly impact an individual's communication fluency and social interaction. However, current research predominantly focuses on limited feature sets and conventional deep learning models, which may not fully capture the complex temporal variations inherent in stuttered speech. To address this gap, an optimized deep learning framework is developed for effectively identifying stuttering disfluencies with improved accuracy and adaptability. The UCLASS dataset, which consists of a variety of stuttered and fluent speech samples, was utilized to ensure diversity in speech patterns. Data augmentation techniques, including time stretching and pitch shifting, were applied to balance the dataset and enrich its variability. Further, noise reduction was conducted using spectral gating to enhance signal clarity. For feature extraction, Gammatone Frequency Cepstral Coefficients (GFCC) were employed to capture robust auditory-inspired features that better represent the perceptual characteristics of speech. The proposed Adaptive Sheep Flock-driven Gate customized Long Short-Term Memory Network (ASF-GLSTM-NET) integrates adaptive gate mechanisms inspired by sheep flock optimization, enabling the model to dynamically focus on critical temporal speech patterns while reducing irrelevant noise. This architecture efficiently learns both local and long-range dependencies for precise classification of stuttering and non-stuttering events. Implemented in Python, the findings show that the ASF-GLSTM-NET approach performs better than multimodal baseline architectures, achieving superior results, with accuracy, F1-score, recall, and precision ranging from 95% to 99%. These findings support the clinical relevance of the proposed system for future diagnostic and assistive applications.

Keywords: Stuttering Disfluencies, Speech Signal Processing, Adaptive Sheep Flock-driven Gate customized Long Short-Term Memory Network (ASF-GLSTM-NET), Automatic Speech Classification, Speech Disorder Analysis.

1. Introduction

Stuttering was a multifactor speech disorder with frequent and severe disturbance in the fluency of speech production, including repetitive sounds, vowels, words, word prolongation, and high rates of interruptions or dysfluencies during the process of speech production [1]. Stuttering impacts up to 1% of the world population with far-reaching psychological, social, and educational outcomes, including the loss of self-confidence, social anxiety, and the inability to avoid speaking [2]. Stuttering disfluencies were detected and rated based on the manual assessment by qualified speech-language pathologists (SLPs) by carefully reading the audio recording of the speech sample and listening to the recording [3]. Applications of the DL model included automatic speech

recognition (ASR), speaker identification, emotion recognition, and pathological speech [4]. Strong systems based on a transformer have been substantially modulated in the automatic diagnosis of speech disorders [5]. Automated stuttering disfluency detection has several challenges in robust and versatile modeling techniques [6]. Speech was stationary since it depends on a variety of factors such as the speaker, the speed of speaking, emotional condition, and noise in the environment [7]. The DL model is used to recognize disfluency patterns across heterogeneous data, especially when using large-scale annotated speech quantities [8]. Audio, combined with other modalities, such as the analysis of facial motion or articulatory features by video, is included in the multimodal systems to enhance the accuracy of the recognition [9]. Credible automatic stuttering detection offers real-time monitoring and feedback system possibilities, which would, under certain conditions, become a component of interventions of speech therapy [10]. Speech signals from the real world frequently contain background noise. Without noise treatment, the system's performance might deteriorate in uncontrolled acoustic conditions.

The objective of this research is to create a revolutionary Adaptive Sheep Flock-driven Gate customized Long Short-Term Memory Network (ASF-GLSTM-NET) approach to enhance complex temporal variations in speech. The suggested approach is used to facilitate early diagnosis and personalized intervention for individuals with stuttering.

The remaining research consists of a literature review on stuttering disfluency detection, methodology (data pre-processing, feature extraction, and ASF-GLSTM-NET model development), and experimental results highlighting superior metrics. The conclusion emphasizes diverse speech patterns and practical applicability for early diagnosis and personalized intervention.

2. Related Works

Communication can be occasionally impeded by disfluencies, including blockages in words or phrases, interruptions, extensions, and repeats [9]. The automated approach for disfluency detection uses convolutional long short-term memory (ConvLSTM) and convolutional neural networks (CNN). The performance of ConvLSTM and CNN models has low detection accuracy. Speech therapists might use different categorizations of stuttering and its subtypes to assess the degree of stuttering, provide initial diagnoses for individuals, and facilitate communication with the voice-based commands [10]. To examine the impact of distinct signal characteristics on the classification outcomes, including Mel-Frequency Cepstral Coefficients (MFCCs), signal pitch-determining variables, and different speech representations. The Residual Network (ResNet) 18 model was used to identify stuttering with an F1-score of 0.93. A summary of related works on the detection of stuttering disfluencies in speech signals is illustrated in Table 1.

Table 1: Summary of Literature Review onDetection of Stuttering Disfluencies in Speech Signals

Ref	Technology Used	Objective	Result	Challenges
[11]	Multi-branching (MB) scheme, weighted loss function,	Address class imbalance and data scarcity for stuttering detection (SD)	Huge improvement over baseline StutterNet, 4.48% of multi-contextual StutterNet	Limited and imbalanced stuttered speech data
[12]	CNN, entropy-invariance in attention	Detect disfluency, scalable across variable-length speech,	Outperforms baselines in English; demonstrates universality and scalability	High cost of annotated data, variable-length disfluent speech

[13]	LSTM for context and multi-task learning	To improve stuttering detection performance	Extensive experiments withan F1 score of 39.8%	Uncertain relationship between detection tasks
[14]	Hyper-Heuristic Search Algorithm,	Develop automatic stutter speech recognition and audio classification	High classification accuracy for stutter audio	Complexity of stuttering pathology
[15]	DisfluencyNet, contextual embeddings, class balancing	Detect disfluency robustly with minimal training data	Disfluency types of except blocks;	Limited disfluency data, low performance
[16]	Word-level stuttered speech dataset	Enable word-level stutter detection	Surpasses previous word-level detection approaches,	Lacks comfortable stuttering, andsound-level detection

2.1 Problem Statement

The LSTM networks were computationally heavy, which can restrict utilization in low-resource applications that demand real-time performance [13]. The multi-task learning strategy was effective and which can lead to task interference. The functional application of these models within clinical practice requires additional exploration for speech-language pathologists. Hyper-Heuristic Search Algorithm can raise the computational burden and make the algorithm difficult to apply in a real-time application [14]. Heuristics can restrict the flexibility in regard to alternative stuttering patterns or other speech dissimilarities that were unpredictable. The automatic recognition of the speech component was effective in handling irregular stuttered speech with caution.

To overcome these issues, the suggested approach, ASF-GLSTM-NET used to enhance stuttering detection by accurately identifying speech disfluencies in real-time. By integrating adaptive gate mechanisms inspired by sheep flock optimization, the system achieves high precision, robustness, and scalability, supporting early diagnosis and personalized intervention for diverse speech patterns.

3. Proposed Methodology

The UCLASS dataset was used to ensure diversity in speech patterns. To balance and enhance the dataset's variability, data augmentation techniques such as pitch shifting and temporal stretching were used. Additionally, to improve signal clarity, spectral gating was used for noise reduction. To extract robust auditory-inspired features, GFCC was used. The ASF-GLSTM-NET model is used to dynamically concentrate on temporal speech patterns. Figure 1 illustrates the overall process of stuttering disfluencies in speech signals.

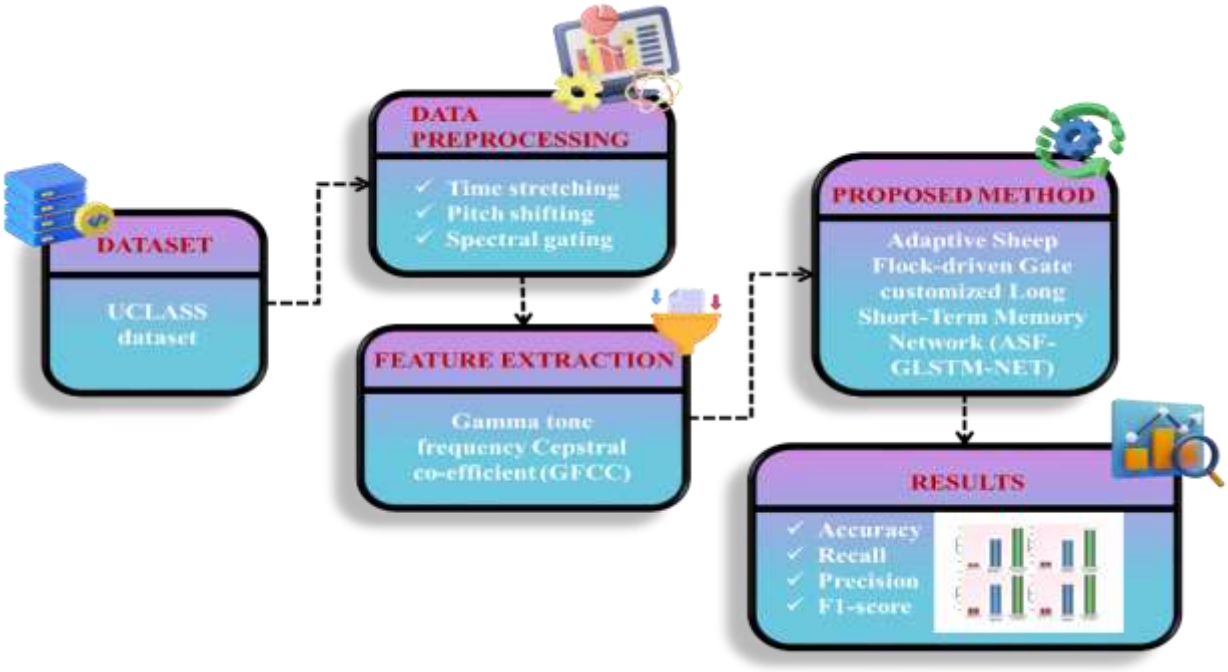

Figure 1: Overall Flow of Stuttering Disfluencies in Speech Signals.

3.1 Dataset

The UCLASS dataset was developed by the Department of Psychology and Language Sciences. It includes two key proclamations, Release One and Release Two, both of which are comprised of audio records of people stuttering, such as monologues, conversations, or even a certain task of reading. The collection is made up of many releases that include audio recordings, transcription files, and comprehensive information about speakers. This dataset helps to evaluate stuttering patterns by analyzing the sound files using synchronized phonetic and orthographic transcription data. The database makes it possible to construct automatic stuttering detection systems and enhance speech treatment. The original audio waveform was illustrated in Figure 2.
Source: https://www.uclass.psychol.ucl.ac.uk/

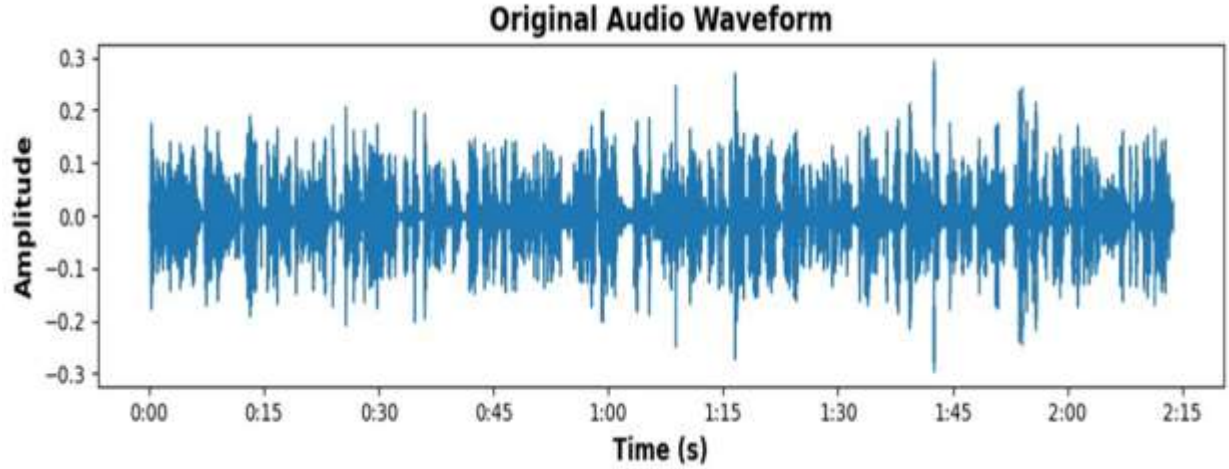

Figure 2: Visualization of the original audio waveform

3.2 Data augmentation using Time Stretching

Time stretching is used to equalize the duration of all audio streams. This operation assists the detection model to generalize more on different speaking speeds and natural speech variation that exists in real-life circumstances. The Short-Time Fourier Transform (STFT) offers frequency information with time-localized,

while the latter offers frequency information with the entire signal period. A time stretching factor was determined for every signal, and it was depicted in Figure 3. Applying the STFT to the input audio signal causes a change in the time domain and frequency domain, yielding the STFT.

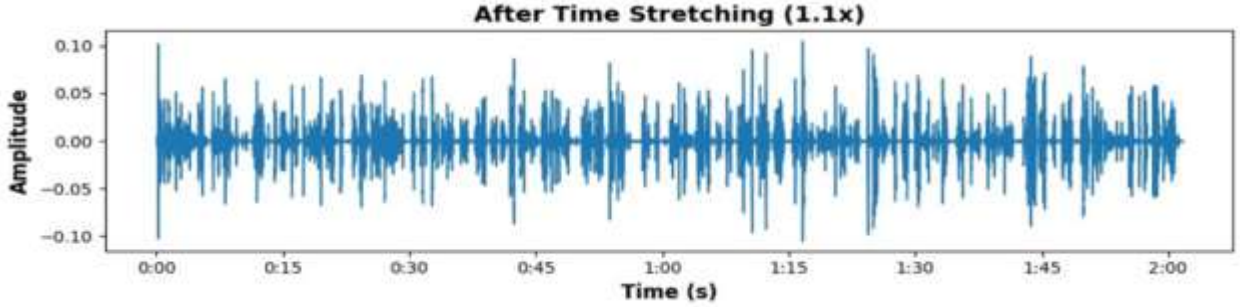

Figure 3: Time-Stretched Audio Waveform

3.2.1 Data augmentation using Pitch Shifting

Pitch shifting adjusts the original pitch of audio samples without changing the length, providing a wide range of variations to balance the data. This method augments acoustic variability, and the speaking model can be quantified better across various individuals and communication circumstances. To expand the vibration analysis training dataset, a pitch-shifting technique was suggested. This method can maintain the semantic meaning even after the signal has been altered; it might thus address the problem of signal restriction. The shift pitching technique creates a signal variation at the signal spectrum's frequency angle, as illustrated in Figure 4.

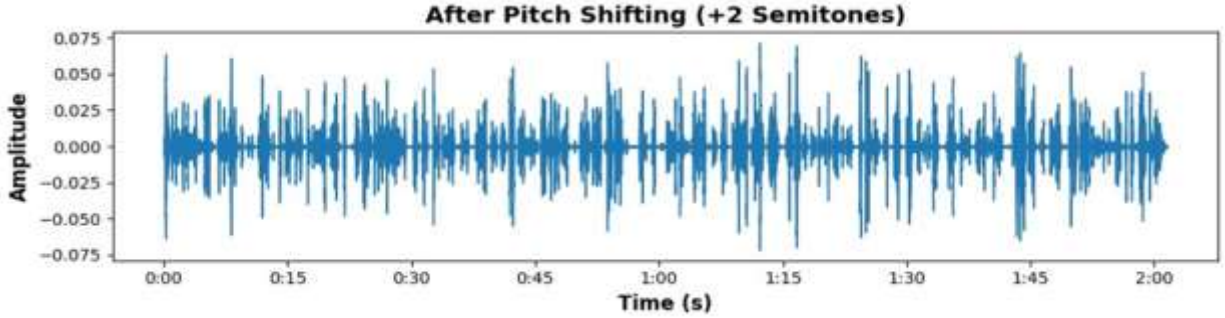

Figure 4: Pitch-shifting audio waveform

3.3 Noise Reduction using Spectral Gating

Spectral Gating used to reduce noise levels in speech signals to produce better clarity of speech sounds without distorting desirable speech information, was depicted in Figure 5. A typical spectrum noise gate method is used to attenuate the signal according to a threshold, and it is frequently used in sonorous mixing and modification. It has been applied to the general noise reduction process with some changes. Initially, it uses the previously acquired loud sound to identify the noise spectrum or auditory signature. The sound spectrum undergoes certain attenuation processes.

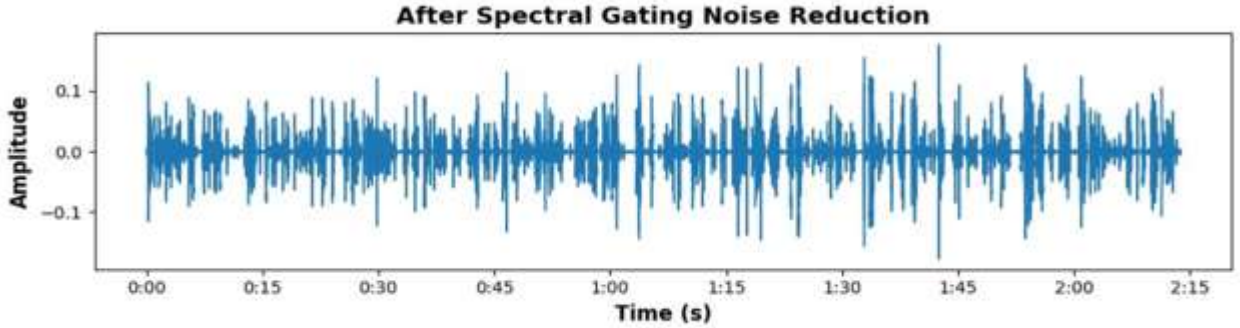

Figure 5: Waveform after Spectral Gating Noise Reduction

3.4 Feature extraction using GFCC

GFCC offers perceptually meaningful and noise-robust characteristics that capture minute speech fluctuations in the context of automatically detecting stuttering disfluencies. The gammatone filter synthesizes the nerve cells' impulse response as an auditory fiber, simulating the listening of speech messages. The Gammatone filter bank was used to form GFCCs for the frequency selectivity of the human cochlea compared to the Mel scale. GFCCs reflect the spectral envelope of the speech signals, which can be applied in a number of applications, such as voice recognition and speech emotion. The modified version of the cell function was expressed in Equation (1).

$$h(s) = bs^{m-1}f^{-2\pi a(e_d)s}\,cos(2\pi e_d s + \emptyset) \quad (1)$$

The high peak value was indicated as$b$, and the onset time as s^{m-1}, bandwidth as$f$, decay function as e_d, and the beginning phase as $\emptyset$. The GFCC coefficients were produced for speech emotion identification by using the GFCC filter, as depicted in Figure 6.

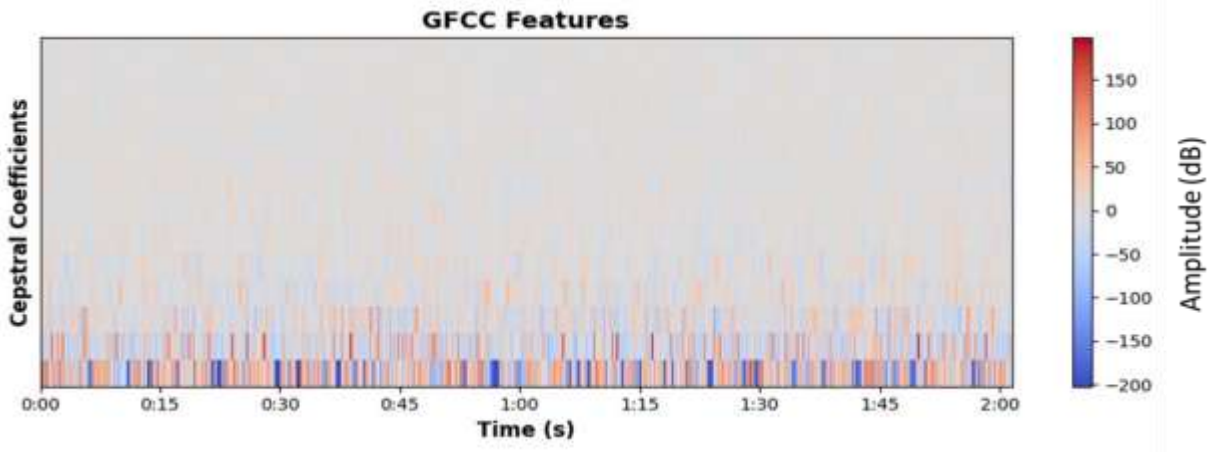

Figure 6: Speech Pattern Characterization by GFCC Coefficients

3.5 Adaptive Sheep Flock-driven Gate customized Long Short-Term Memory Network (ASF-GLSTM-NET)

The integration of ASF and GLSTM-NET presents a hybrid approach for enhancing the automatic detection of stuttering disfluencies in the speech signal. The ASF algorithm is inspired by the swarm grazing and self-adapting flocking behavior of sheep. It was specifically optimized towards dynamically exploring the solution space by the capability to manage exploration and exploitation through leader-following and random wandering techniques. Adding the gates selectively to control the GLSTM-NET architecture improves the standard LSTM structure by reducing memory loss and the vanishing gradient phenomenon. The hybrid model, ASF-GLSTM-NET, reflects the synergistic capability of evolutionary optimization, capable of discovering adaptive neural network architecture to address the complexity of speech disfluency with user-friendly speech therapy technologies. The ASF-GLSTM-NET hybrid model used to enhance the automatic detection of stuttering disfluencies in speech signal, as expressed in equation (2).

$$arg_\theta\, min\, K\,(z, \hat{z}) = arg_\theta\, min \frac{1}{M}\sum_{j=1}^{M} \mathcal{K}(z_1, GLSTM - NET\left(W_{J.}ASF\big((-)\big)\right) \quad (2)$$

Where θ represents the weighted parameters, and $\mathcal{K}$ indicates the Loss function.

3.5.1 Gate customized Long Short-Term Memory Network (GLSTM-NET)

GLSTM-NET was a complex sequence modeling through updating the gating mechanisms to model complex temporal dependencies like blocks, prolongations, and repeats. In contrast to conventional LSTMs with fixed input, forget, and output gates, GLSTM-NET employs adaptive or learnable gating functions, and their performance depends on the context of the underlying input. GLSTM-NET architecture incorporates the input, output, and forget gates. GLSTM-NETmodels were used for interacting with temporal patterns, including voice signal processing or sequential data analysis.

- **Input Gate Layer**

The Input Gate Layer regulates the flow of pertinent auditory information into the neural network, which is essential for the automated identification of stuttering disfluencies in speech signals. The current input and the preceding hidden layer provide information to the input gate layer. The input gate secures to extract long-term

patterns without being disturbed at once by unwanted data. After that, it calculates the data to provide the following output, which is expressed in equation (3).

$$j_s = \sigma\ (X_j \times \left[d_{s-1,} g_{s-1,} W_s\right] + a_j$$
$$\widetilde{D}_s = tan(X_d \times \left[d_{s-1,} g_{s-1,} W_s\right] + a_d \quad (3)$$

Where W_s and g_{s-1} represent the output and input of the preceding hidden layer,$a_j$and$a_d$illustrate the bias of the input gate,σ depicts a triggering function. The soft function was expressed in equation (4) as follows;

$$\sigma_{softsign}(w) = \frac{w}{1+|w|} \quad (4)$$

- **Forget Gate Layer**

The Forget Gate Layer, which selectively filters redundant or unnecessary information during sequence processing, was essential for improving the automated identification of stuttering disfluencies in speech signals. This gate keeps the memory of the network balanced and focused on relevant patterns by constantly refreshing. Equation (5) shows the varying weights $X_e$and bias a_e, which illustrate how the forget gate's output was calculated by using a formula similar to the input gate.

$$e_s = \sigma(X_e \times \left[d_{s-1,} g_{s-1,} W_s\right] + a_e \quad (5)$$

- **Cell State Update**

In automatic stuttering disfluency detection, the cell state update serves to detect minor temporal variation and anomalies in fluency like repetitions, prolongation, or blocks. Cell State Update was the crucial operation that helped in propagating valuable information across long sequences to distinguish subtle temporal patterns like repetitions, prolongations, and blocks. It operates by perfectly mixing the impact of the forget gate with the input gate. The forget gate determines the aspects of the cell state. The combination of these two operations of forgetting and adding was completed by element-wise multiplication and addition, resulting in the new cell state. Equation (6) illustrates the process of updating from the prior cell state.

$$d_f = e_s \times d_{s-1} + j_s \times \widetilde{D}_s \quad (6)$$

- **Output Gate Layer**

In automatic recognition of stuttering disfluent of the speech signal, the output gate layer played the role of optimizing the neural network recognition power with respect to disfluent speech patterns by guarding salient time and acoustic features. The Output Gate Layer decides the information in the current cell state revealed as the next input to the hidden layer, and the next input to the network strengthens the reliability and efficiency of stuttering detection frameworks. The output gate then multiplies this transfigured cell state with its sigmoid output, selecting to forward only the most vital components. The current input, output, and forget gate of the preceding hidden layer determine the outcome of equation (7).

$$\mathrm{P_s} = \sigma(\mathrm{X_P} \times \left[\mathrm{d_{s-1,} g_{s-1,} W_s}\right] + \mathrm{a_p})$$
$$\mathrm{g_s} = \mathrm{P_s} \times \tanh(\mathrm{d_s}) \quad (7)$$

Here $\mathrm{P_s}$, $\mathrm{g_{s,}}$and $\mathrm{d_s}$illustrate the output of the gate.

3.5.2 Adaptive Sheep Flock (ASF) Optimization

- **SF**

SF Optimization uses a sheep flock's cooperative behavior as a model to find the best feature subsets in complicated speech data. SF was a metaheuristic algorithm inspired by nature that simulates the group movement and decision-making processes of a flock of sheep. It was intended for optimization issues requiring a skillful balancing act between exploration and exploitation. Sheep herd intelligently in the wild, rearranging their locations in response to interactions with the flock leaders.

- **ASF**

ASF Optimization augmented with automatic identification of stuttering disfluencies to connect the speech signals can provide a highly durable and smart framework in the analysis of speech disorders. ASF the individual sheep navigate within the search space confirmed by the social interactions through predators, and ensuring the right distances with their neighbors. To avoid local optima, sheep interact with rules that resemble grazing and migration to discover higher quality solutions. The step sizes were gradually decreased after initially permitting the extensive investigation with large step sizes. Equation (8) represents the adaptive step size formula.

$$\propto_{\mathrm{S}} = \propto_{\max} - \left(\frac{\mathrm{S}}{\mathrm{S}}\right)(\propto_{\max} - \propto_{\min}) \quad (8)$$

Sheep walk with static step sizes in classical SFO, which might cause delayed exploration or early convergence. The step size $\propto_{\mathrm{S}}$ dynamically changes the number of iterations to enable large exploration steps in the beginning and fine-tuning close to convergence. The scattering behaviour was corrected in the SF. Narrow

scattering subsequently refines solutions in potential places, whereas wide scattering initially encourages exploration, as expressed in equations (9 and 10).

$$w_{new} = w_{current} + \propto_S (w_{best} - w_{current}) + \beta Q. \quad (9)$$

$$C_{scatter} = C_{max} - (\frac{S}{S})(\propto_{max} - \propto_{min}) \quad (10)$$

By using a random fluctuation term to direct the sheep's location toward the most well-known position, w_{best}relieson random movement.

The hybrid model, ASF-GLSTM-NET, represents a significant advancement in automated speech disfluency detection, offering a promising avenue to enhance diagnostic precision and facilitate early intervention tactics for stuttering individuals.

4. Results and discussions

Experimental data indicate that the suggested model, ASF-GLSTM-NET, improves stuttering disfluency detection performance in speech analysis tasks. The experimental system was executed using Python 3.10.1 to realize the proposed method. This Python version was selected for its compatibility and improved performance over previous versions.

4.1 Performance assessment of the suggested method

A classification model's performance was evaluated by using a confusion matrix. The overall effectiveness of the model was provided by this matrix. It displays both the True and Predicted labels for several categories. It assists in figuring out the system's operational efficiency and prospective extents for improvement. Figure 7 represents the confusion matrix of stuttering disfluencies in speech signals.

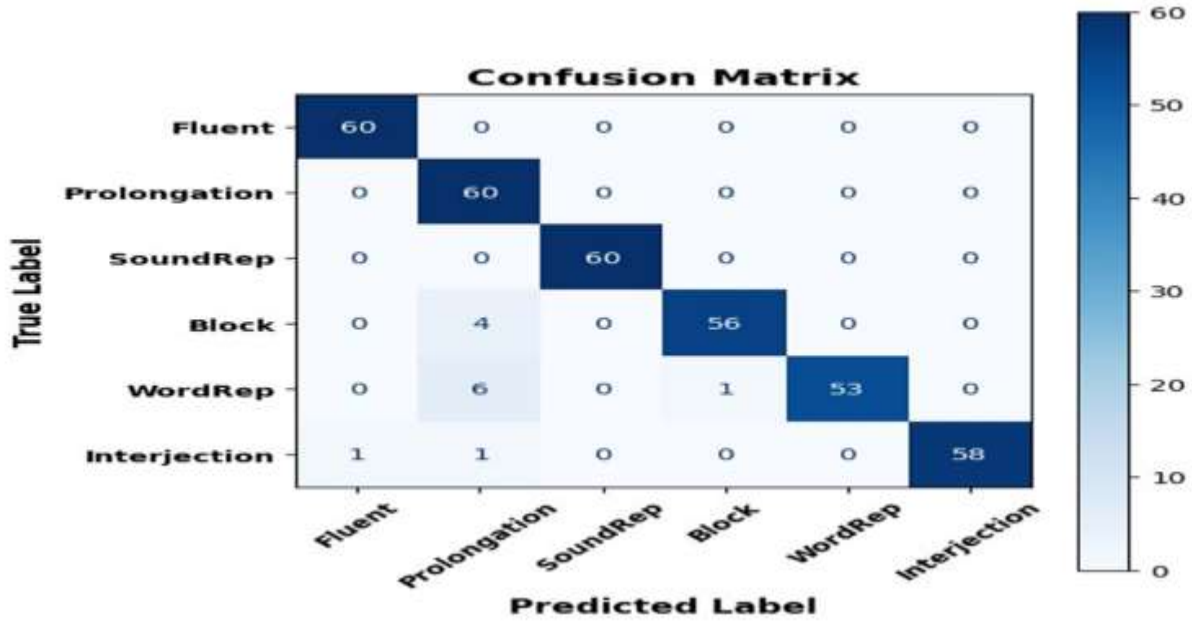

Figure 7: Confusion Matrix of stuttering disfluencies in speech signals.

Detection of fluency and prolongation in speech signals was illustrated in Figure 8. Where x- plane indicates the time in (sec) ranging from 0 to 2, and y- plane indicates the cepstral coefficients, each capturing different spectral characteristics of the speech. This segmental contrast aids in locating articulatory and rhythmic abnormalities in speech. This type of analysis has the potential to assist in the therapeutic monitoring of speech disorders like stuttering.

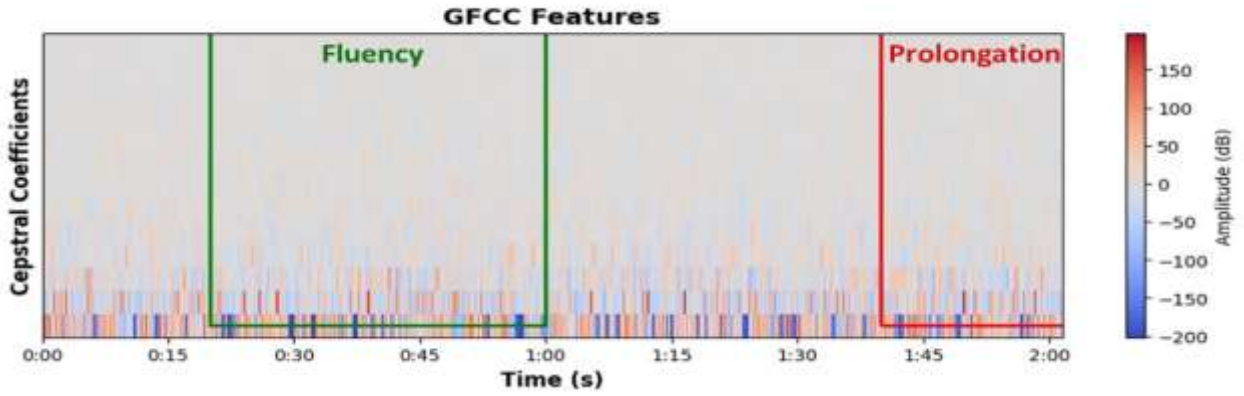

Figure 8: Detection of Fluent and Prolongation in Speech Signals

4.2 Metrics for evaluating the effectiveness of the proposed model

- **Accuracy:** Measures the overall rate of accurate predictions. This measure shows the level of differentiation between stuttered and fluent portions of speech.
- **Precision:** This metric shows the number of real cases of disfluency in the total number of predicted disfluencies.
- **Recall:** Measures the model's capacity to detect all actual instances of a class. It helps to identify all the true cases of stuttering disfluencies.
- **F1-Score:** Balances precision and recall using their harmonic mean. This metric shows the stability and trust of the model to identify stuttering disfluencies in distinct speech patterns.

The proposed ASF-GLSTM-NET approach is used for detecting stuttering disfluencies in speech patterns. The system achieves an accuracy of 98%, demonstrating its high reliability in stuttered and fluent speech. With a precision of 97.65% and a recall of 97.50%, it effectively captures true disfluency instances while minimizing misclassifications. Furthermore, the F1-score of 97.15% validates the model. These metrics validate that the ASF-GLSTM-NET approach provides an effective, adaptive solution for stuttering disfluency detection.

4.3 Comparative Result Analysis

Stuttering causes significant psychological, social, and educational effects, such as the depletion of self-confidence, shyness in society, and the inability to suppress speaking. The StutterNet-based hybrid lightweight (SNet-BHL) model, which was deployed on resource-limited devices based on StutterNet, has reduced the ability to reproduce multi-acoustic features compared with the networks [17]. Ethical issues based on the privacy and manipulation of automatic speech monitoring are concerned with sensitive situations. SNet-BHL model has performance metrics accuracy of (93.84%), lacks the task of reliably determining subtle instances of disfluency. The Bi-Directional Long-Short Term Memory (Bi-LSTM) demonstrates strong specifications in relation to aligning the contextual phenomena in the speech signal, for the automatic identification of stuttering moments of speech has some significant restrictions. Bi-LSTM models perform strongly on the amount and diversity of annotated data required to generalize effectively over speakers, dialects, and spontaneous speech setups [18]. Bi-LSTM models have a performance metric accuracy of 96.67%, and are computationally more complex, which might not be feasible in terms of real-time implementation on low-resource devices.

The performance results of the proposed method are discussed in this section. The outcomes are contrasted with the other approaches, like the SNet-BHL model [17] and Bi-LSTM [18]. The Comparative performance of the ASF-GLSTM-NET models, precision, F1-score, accuracy, and recall, was illustrated in Table 2 and Figure 9.

Table 2: Comparative performance of ASF-GLSTM-NET models, precision, F1-score, accuracy, and recall.

Methods	F1-Score (%)	Precision (%)	Accuracy (%)	Recall (%)
SNet-BHL [17]	93.19	93.11	93.84	93.30
Bi-LSTM [18]	96.01	96.18	96.67	96.31
ASF-GLSTM-NET[proposed]	**97.15**	**97.65**	**98**	**97.50**

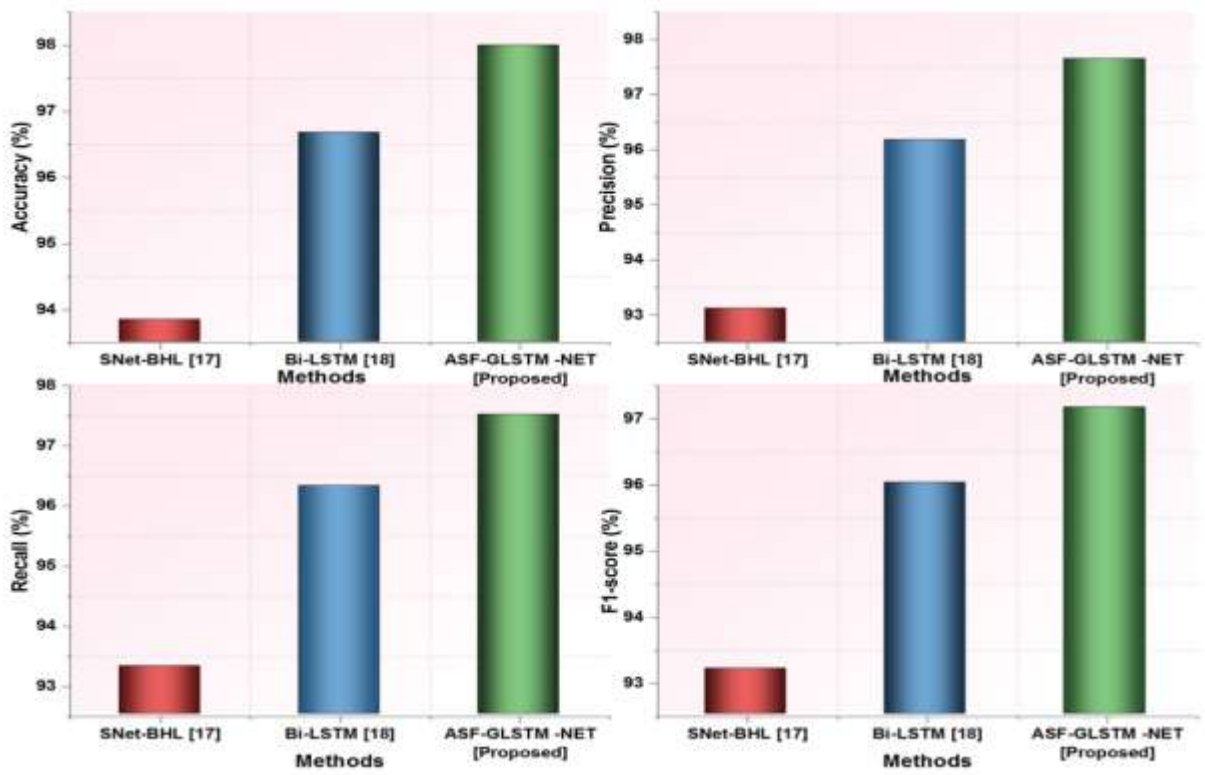

Figure 9: Evaluation of ASF-GLSTM-NET models, precision, accuracy, recall, and F1-score for enhancing stuttering disfluencies in speech signals

To overcome these limitations, the suggested ASF-GLSTM-NET model enhances stuttering disfluency detection. The integration of ASF Optimization dynamically refines temporal focus, while the customized GLSTM-NET architecture efficiently captures both short- and long-term speech dependencies. This combined approach delivers precise, adaptable classification of stuttered and fluent speech patterns. Overall, the ASF-GLSTM-NET approach provides a robust, efficient, and scalable solution for next-generation speech disorder diagnostics.

5. Conclusions

Stuttering disfluencies influenced the communication fluency and social interaction of an individual to a considerable degree and were manifested by interruptions of speech components like speech blocks, prolongations, and repeats. Speech pattern diversity was provided using the UCLASS dataset. Two algorithms of data augmentation were performed to balance and enhance the variability of the dataset, which included pitch shifting and temporal stretching. Spectral gating was used to remove noise and maximize signal clarity. GFCC was used to good auditory-inspired features. The ASF-GLSTM-NET paradigm dynamically concentrated on temporal speech patterns. Experimental testing revealed a better result in all the above accuracies (98%), F1-score (97.15%), recall (97.50%), and precision (97.65%), showing that the model has the possibility of enhancing automation in the process of stuttering detection.

Limitations and Future Scope

Stuttering disfluencies were typically made up of the total duration of speech. This imbalance might result in a biased detection where fluent speech prevails in more predictions with lower recall of disfluent events. Future studies might investigate customized or speaker-adaptive models that adapt to each user's unique stuttering patterns to increase accuracy. Integrating automatic detection with interactive feedback mechanisms to assist speech therapy, track progress, and provide actionable insights for professionals.

Funding

The authors did not receive any specific funding for this research.

References

1. 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.https://doi.org/10.1109/TASLP.2021.3110146
2. 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. https://doi.org/10.1142/S0219649222500204
3. Murugan, K., Cherukuri, N.K. and Donthu, S.S., 2022, June. Efficient recognition and classification of stuttered words from the speech signal using a deep learning technique. In 2022 IEEE World Conference on Applied Intelligence and Computing (AIC) (pp. 774-781). IEEE.https://doi.org/10.1109/AIC55036.2022.9848868
4. Prabhu, Y. and Seliya, N., 2022, December. A CNN-based automated stuttering identification system. In 2022, 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1601-1605). IEEE.https://doi.org/10.1109/ICMLA55696.2022.00247
5. Shen, J. and Zhang, X., 2025. Individual-independent and cross-language detection of speech disfluencies in stuttering based on multi-adversarial tasks and self-training. Biomedical Signal Processing and Control, 100, p.107051.https://doi.org/10.1016/j.bspc.2024.107051
6. Manjula, G., Shivakumar, M., Latha, M. and Keerthi Kumar, M., 2023. Development of an adaptive optimization ANN model for the automatic identification and classification of stuttering in speech. SN Computer Science, 4(4), p.361.https://doi.org/10.1007/s42979-023-01793-2
7. Basak, K., Mishra, N. and Chang, H.T., 2023. TranStutter: A convolution-free transformer-based deep learning method to classify stuttered speech using 2D mel-spectrogram visualization and attention-based feature representation. Sensors, 23(19), p.8033.https://doi.org/10.3390/s23198033
8. Nirosha, K. and Venkatesh, N., 2024, December. Automated Dynamic Stuttered Speech Recognition System and Conversion System using Mel Filter and Enhanced Logistic Regression Model. In 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1254-1261). IEEE.https://doi.org/10.1109/ICACRS62842.2024.10841611
9. Alhakbani, N., Alnashwan, R., Al-Nafjan, A. and Almudhi, A., 2025. Automated Stuttering Detection Using Deep Learning Techniques. Journal of Clinical Medicine, 14(10), p.3552.https://doi.org/10.3390/jcm14103552
10. 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.https://doi.org/10.3390/app13106192
11. Sheikh, S.A., Sahidullah, M., Hirsch, F. and Ouni, S., 2023. Advancing stuttering detection via data augmentation, class-balanced loss, and multi-contextual deep learning. IEEE Journal of Biomedical and Health Informatics, 27(5), pp.2553-2564.https://doi.org/10.1109/JBHI.2023.3248281
12. Liu, J., Wumaier, A., Wei, D. and Guo, S., 2023. Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths. Applied Sciences, 13(13), p.7579.https://doi.org/10.3390/app13137579
13. Liu, X., Xu, C., Yang, Y., Wang, L. and Yan, N., 2024. An End-To-End Stuttering Detection Method Based On Conformer And BILSTM. arXiv preprint arXiv:2411.09479.https://doi.org/10.48550/arXiv.2411.09479
14. Grandhi, S.H. and Kamruzzaman, M.M., 2024. Automatic Stutter Speech Recognition and Classification Using a Hyper-Heuristic Search Algorithm. International Journal of Automation and Smart Technology, 14(1).https://doi.org/10.5875/dm761m36
15. Mohapatra, P., Pandey, A., Islam, B., and Zhu, Q., 2022, July. Speech disfluency detection with contextual representation and data distillation. In Proceedings of the 1st ACM international workshop on intelligent acoustic systems and applications (pp. 19-24).https://doi.org/10.1145/3539490.3539601
16. Shih, Y.J., Gkalitsiou, Z., Dimakis, A.G. and Harwath, D., 2024, December. Self-Supervised Speech Models For Word-Level Stuttered Speech Detection. In 2024 IEEE Spoken Language Technology Workshop (SLT) (pp. 937-944). IEEE.https://doi.org/10.1109/SLT61566.2024.10832220
17. Abubakar, M., Mujahid, M., Kanwal, K., Iqbal, S., Asghar, N. and Alaulamie, A., 2024. StutterNet: stuttering disfluencies detection in synthetic speech signals via mel frequency cepstral coefficients features using deep learning. IEEE Access.https://doi.org/10.1109/ACCESS.2024.3429343
18. Gupta, S., Shukla, R.S., Shukla, R.K. and Verma, R., 2020. Deep learning bidirectional LSTM-based detection of prolongation and repetition in stuttered speech using weighted MFCC. International Journal of Advanced Computer Science and Applications, 11(9).http://dx.doi.org/10.14569/IJACSA.2020.0110941