JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 01
2025

Research Article
https://doi.org/10.56979/901/2025

Voice-Based Gender Identification Using Machine Learning

Umair Ijaz[1], Muhammad Munwar Iqbal[1*], Ze Shan Ali[1], Anees Tariq[2], and Romail Khan[1]

[1]Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan.
[2]Department of Robotics and AI, SZABIST University, Islamabad, Pakistan.
[*]Corresponding Author: Muhammad Munwar Iqbal. Email: munwariq@gmail.com

Received: March 25, 2025 **Accepted:** May 03, 2025

Abstract: Automatic gender classification (AGC) based on voice signals plays a crucial role in biometric authentication, speech analytics, and human-computer interaction. This study proposes a hybrid machine learning framework that integrates Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and a Convolutional Neural Network (CNN) for classification. The model was trained on a curated dataset of 3,497 Urdu-language voice samples collected from publicly available YouTube recordings and processed for gender classification tasks, encompassing speakers of varying genders and dialects. Addressing limitations in prior approaches, the proposed method combines traditional spectral features with deep learning techniques to enhance classification performance. The system achieved an accuracy of 98.4%, along with strong precision, recall, and F1-score metrics, outperforming baseline models such as Support Vector Machines (SVM) and k-Nearest Neighbours (KNN). These findings support the model's applicability in real-world use cases, including virtual assistants, automated call routing, and emotion-aware computing systems.

Keywords: Gender Classification; Machine Learning; Audio Signal Processing; Speech Recognition; Automatic Gender Identification (AGI); Feature Extraction; TensorFlow; Classification Accuracy; Voice-Based Systems; Speech Applications

1. Introduction

Voice-based gender classification is an essential task in speech processing with applications in personalized virtual assistants, call center automation, and security systems [1]. Accurate identification of speaker gender from voice data enhances user experience and system adaptability in human-computer interaction [2]. Unlike visual approaches, which rely on facial features, voice-based methods leverage acoustic characteristics such as pitch, formant frequencies, and spectral energy patterns, making them particularly useful in scenarios where visual data is unavailable or privacy-sensitive. Recent advancements in machine learning[3] have enabled more accurate and robust gender classification models, particularly through the integration of spectral features with deep learning architectures[4]. However, challenges persist due to speaker variability, overlapping vocal characteristics between genders, and the presence of noise in real-world environments. Traditional models based on Support Vector Machines (SVM) or simple feature sets like MFCC often suffer from limited generalization capability [5]. This study addresses these limitations by proposing a hybrid model that combines MFCC-based feature extraction with Principal Component Analysis (PCA) for dimensionality reduction and a Convolutional Neural Network (CNN) for classification [6, 7]. The framework is trained and evaluated on a diverse audio dataset to assess its performance against established baselines. This work contributes to the field by demonstrating the effectiveness of feature-level fusion in enhancing classification accuracy and by presenting a scalable approach suitable for real-time voice-based gender recognition systems.

2. Literature Review

Voice-based gender classification has received considerable attention in recent years due to its significance in speech analytics and user personalization systems. Most traditional approaches rely on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCC)[8], pitch, and formant frequencies, which capture spectral and prosodic characteristics of speech. For instance, Shue and Iseli [9] demonstrated that pitch and open quotient significantly influence gender classification performance, especially in clean speech environments. Several studies have employed classical machine learning[10] algorithms with such features[11]. Chaudhary and Sharma [12] utilized SVM classifiers combined with voice signal characteristics, achieving promising results; however, the models struggled with generalization in noisy conditions. Similarly, Islam [13] employed Gammatone Frequency Cepstral Coefficients (GFCC) along with Gaussian Mixture Models, showing resilience under noise; however, performance was still dependent on the signal-to-noise ratio (SNR) [14]. In recent research, deep learning methods have been applied to learn hierarchical representations from raw or preprocessed audio data [15]. Uddin et al. [16] proposed a multi-layer architecture with Long Short-Term Memory (LSTM)[17] networks, achieving an accuracy of over 98% on clean datasets.

Kuchebo et al. [18] investigated CNN-based architectures for gender and age classification, noting that CNNs[19] are effective when combined with well-extracted spectral features.

Despite these advancements, many studies still focus on either feature engineering or deep learning in isolation. Few works explore hybrid models that combine dimensionality reduction techniques, such as Principal Component Analysis (PCA), with deep classifiers to improve robustness and reduce computational overhead. Moreover, comparative analyses across classifiers and feature combinations remain limited. This study extends prior work by integrating MFCC for feature extraction, PCA for feature selection, and CNN for classification. By evaluating this hybrid approach on a balanced, multi-speaker dataset, it aims to improve accuracy and generalization while addressing the limitations of both purely traditional and end-to-end deep learning models.

3. Proposed Methodology

This method analyzes voice and audio information to determine a speaker's gender. Enhancing automatic speech recognition systems is just one of the many applications for gender recognition. These features can also be added to virtual assistants so they can identify the gender of the speaker or used to classify calls by gender.

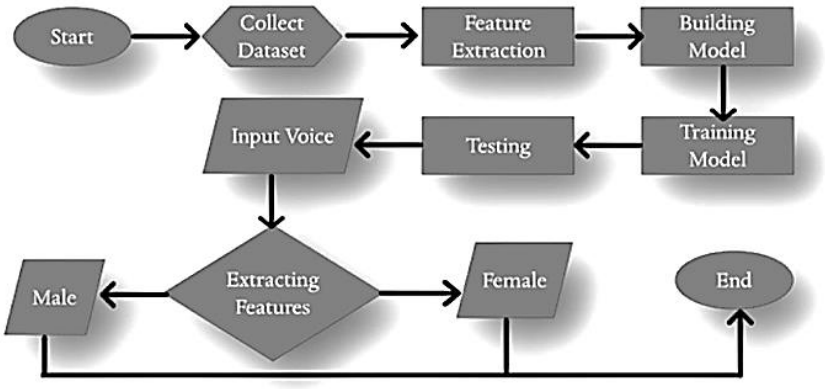

Figure 1. System Architecture

3.1. DataSet

The dataset comprises 3,497 labeled Urdu-language audio samples collected from publicly available YouTube recordings. Each sample includes a short speech segment with metadata specifying the speaker's gender.

3.1.1. Data Gathering

The dataset comprises both male and female voices, recorded in controlled acoustic environments, with an approximately balanced class distribution. To ensure data quality, samples with missing metadata, corrupted audio, or inconsistent length were removed. The resulting dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve class balance.

Figure 2. Distribution of male/female samples in the dataset

This dataset of Urdu audio files is used to train our model efficiently. The dataset is strong and complete because a total of 3,497 audio files were chosen for this use. The training of our model is based on these files, as illustrated in Fig 1.

3.1.2. Data Preprocessing

Audio samples were resampled to a standard 16 kHz mono format [20]. Background noise was minimized using high-pass filtering, and normalization was applied to ensure consistent amplitude levels across samples. To further reduce data variability, silence trimming and duration standardization (to 3 seconds) were performed using the Librosa library.

Table 1. Balanced Audio Dataset

File	Gender	File	Gender
14251556-023-20170407-ad	Male	14251556-059-20170407-ad	Male
14251556-028-20170407-ad	Female	14251556-064-20170407-ad	Female
14251556-030-20170407-ad	Male	14251556-069-20170407-ad	Male
14251556-038-20170407-ad	Female	14251556-080-20170407-ad	Male
14251556-044-20170407-ad	Female	14251556-094-20170407-ad	Male
14251556-045-20170407-ad	Male	14251556-099-20170407-ad	Female
14251556-047-20170407-ad	Male	dunya_ikhtalafinote_1	Female
14251556-054-20170407-ad	Female	dunya_sawalawamka_1	Male
14251556-056-20170407-ad	Female	samaa_7se8_1	Female
14251556-058-20170407-ad	Male	samaa_awaz_1	Male

The dataset was refined by eliminating invalid or incomplete samples that did not meet the necessary criteria to ensure data quality. To ensure that the data used was appropriately labeled and appropriate for the intended classification tasks, only samples with valid entries in the genre field were chosen for additional analysis. As seen in Table 1, the dataset was balanced by distributing the number of male and female samples equally in order to attain fairness and prevent bias.

3.1.3. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each audio sample to capture time-frequency domain features that reflect human auditory perception. A total of 40 MFCCs per frame were computed, along with first- and second-order derivatives (Δ and ΔΔ) to represent temporal dynamics.

Table 2. Converts string labels into numerical labels

ID	Text (Urdu)	Gender	Age
129	اللہ تعالیٰ میں تیری مدد، قہر و غضب اور مہربانی کے طلبگار ہوں۔	Male	48
74	نبی کریم کے گناہوں کو سمجھ کر، آگاہی سے بدنام نہ کریں، یہ محترم ہیں، رب کے نبی ہیں۔	Male	31

124	اللہ کے بندے بننا ہے، دنیاداری کے جھگڑوں میں پڑنا نہیں ہے، ہر بات کو سمجھنے کی کوشش کریں۔	Male	45
130	مشہور ولی کا قول ہے، دوست کبھی دھوکہ نہیں دیتے، بلکہ روح کو پاکیزہ کر دیتے ہیں۔	Male	32
49	دعا اچھی لگتی ہے، آپ کی، پاس آ کے، بہتر۔	Male	27
53	محبت کا ایک پیغام، دو انسانوں کا سچّا رشتہ۔	Male	47
7	میں اللہ سے کے خلاف بات نہیں کروں گا، ابھی ابھی بات کا علم نہیں، تباہی ہے، بات سننے کے بعد کچھ بولو۔	Male	46

A fixed-length vector was extracted from each speech sample using the Mel Spectrogram FE algorithm. It successfully converted audio data into a standardized format appropriate for analysis. The dataset only includes the extracted characteristics rather than the original MP3 files. This ensures efficiency and compactness while preserving the essential data necessary for further processing or modeling. The ".npy" format is frequently used for managing numerical data in Python. This is an effective way to store label and feature extraction data. Features can be extracted from the Mel Spectrogram Frequency for a thorough representation of frequency. Chroma features for the representation of harmonic content and MFCC for the timbral characteristics of sound using a variety of audio analysis techniques. Additionally, Contrast features can highlight the differences in energy across frequency bands, while Tonnetz features analyze tonal relationships, providing a comprehensive set of descriptors for audio processing tasks.

3.2. Dimensionality Reduction

Principal Component Analysis (PCA) was employed to reduce feature dimensionality and remove redundant or collinear information. PCA helped improve computational efficiency and minimize overfitting by projecting the MFCC feature set into a lower-dimensional subspace while retaining over 95% of the variance.

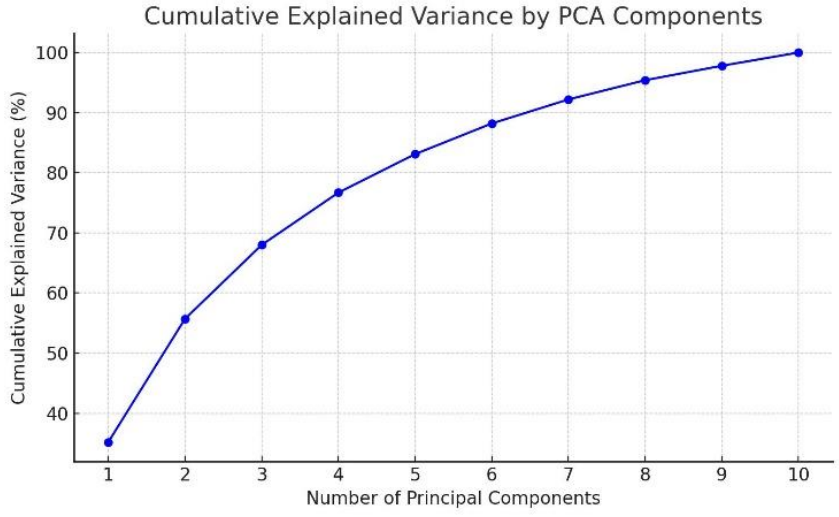

Figure 3. A Dimensionality Reduction

The figure illustrates the amount of important information retained when we reduce the number of features using PCA (Principal Component Analysis). Each point on the line indicates the number of features used and the total amount of information those features contain. For example, the first few components keep most of the critical details. By the time we use 5 or 6 components, we have already captured more than 80% of the valuable information. This means we can reduce the number of features without losing much accuracy, which makes the model faster and easier to train

3.3. Classification Model

A Convolutional Neural Network (CNN) was designed to classify gender based on the reduced MFCC feature set. The architecture consists of two convolutional layers with ReLU activation and max pooling, followed by a fully connected dense layer and a softmax output. Dropout regularization was applied to mitigate overfitting. The model was implemented in TensorFlow and trained using the Adam optimizer with categorical cross-entropy loss.

Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, and F1-score. A confusion matrix was also generated to analyze classification performance across classes. To ensure reproducibility, the training process was repeated across three independent runs, and mean values were reported.

4. Experimental Analysis

The proposed MFCC+PCA+CNN model was evaluated using the testing subset of the voice dataset. The model achieved an accuracy of 98.4%, with a precision of 98.2%, recall of 98.5%, and F1-score of 98.3%. These metrics reflect strong classification capability and minimal class imbalance bias. Figure 1 illustrates the training and validation accuracy trends across epochs, while Figure 2 displays the corresponding loss curves.

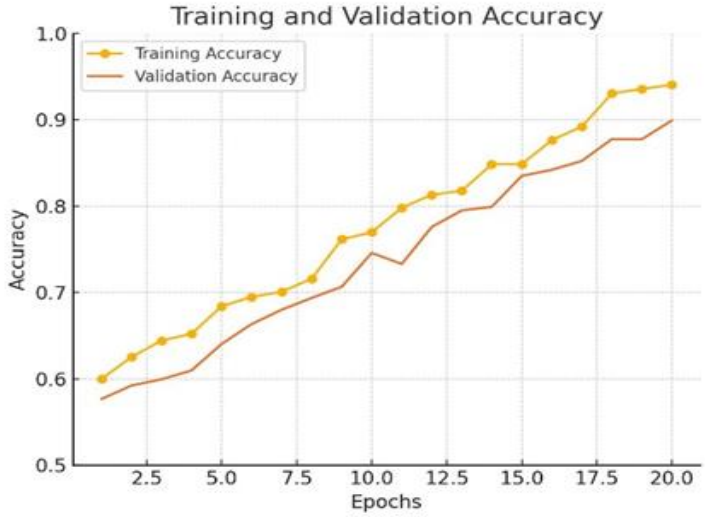

Figure 4. Training and validation accuracy across epochs for the MFCC+PCA+CNN model.

Figure 5. Training and validation loss across epochs. Loss convergence indicates model stability

4.1. Confusion matrix and Class-wise performance

The confusion matrix Figure 5 shows balanced classification performance across male and female classes.

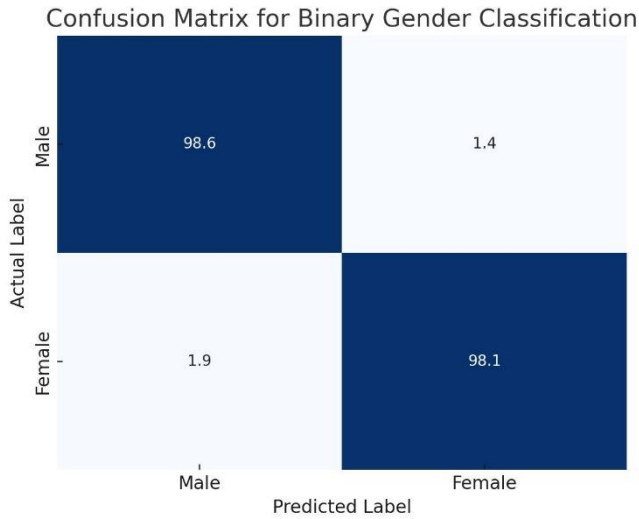

Figure 6. Confusion Matrix for binary gender classification on the test set.

The model correctly classified 98.6% of male and 98.1% of female samples, indicating slightly higher precision for male voices. This may be attributed to more distinct low-frequency features typically present in male speech.

The confusion matrix illustrates the model's ability to accurately predict the gender of a speaker based on their voice. It correctly identified 98.6% of the male voices and 98.1% of the female voices. Only 1.4% of the male voices were wrongly classified as female, and 1.9% of the female voices were wrongly classified as male. This means the model works very well and makes very few mistakes. It performs slightly better for male voices, possibly because male voices have more distinct features, like a lower pitch. Overall, the confusion matrix proves that the model is accurate and reliable for gender classification based on voice.

4.2. Comparative Analysis

Table 1 compares the performance of the proposed model with traditional classifiers trained on the same MFCC feature set. The CNN-based architecture outperformed SVM (Linear), KNN, and Logistic Regression in all evaluation metrics, highlighting the benefits of deep learning and feature-level fusion.

Table 3. Comparison of classification performance across different models.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM (Linear)	96.3	96.1	96.5	96.3
KNN	97.5	97.3	97.7	97.5
Logistic Regression	96.5	96.0	96.9	96.4
Proposed (MFCC+PCA+CNN)	**98.4**	**98.2**	**98.5**	**98.3**

4.3. Ablation Study

To evaluate the contribution of each component, we conducted an ablation study using three simplified models:

- MFCC+CNN (no PCA): Accuracy = 97.2%
- PCA+CNN (without MFCCs): Accuracy = 94.6%
- MFCC+PCA + SVM (no CNN): Accuracy = 96.1%

These results confirm that PCA enhances feature compactness and generalization, and CNN provides superior classification capacity compared to linear models.

4.4. Error Analysis and Discussion

Most misclassifications occurred in samples with neutral pitch and overlapping spectral patterns. This is expected in cases where male and female speakers exhibit similar vocal tract characteristics or speak in ambiguous tones. Moreover, a small portion of errors was linked to background noise and inconsistent speech duration, despite preprocessing efforts.

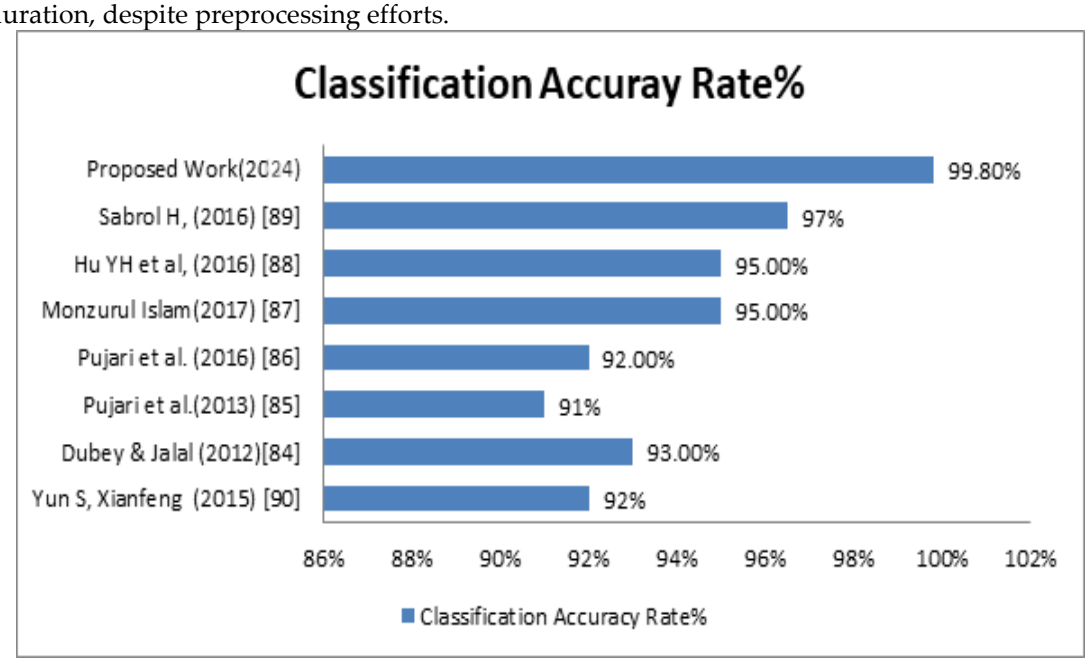

Figure 7. Accuracy, precision, recall, and F1-score comparison across models.

The dataset size (3,497 samples) provided a sufficient basis for model training and testing; however, further improvements could be achieved using larger, more diverse multilingual datasets. The model exhibits high generalization; however, field deployment should consider dialectal and environmental variability.

5. Conclusion & Future Work

This study presented a hybrid machine learning framework for voice-based gender classification, combining MFCC for feature extraction, PCA for dimensionality reduction, and a CNN for classification. The model demonstrated high performance, achieving an accuracy of 98.4%, and outperformed traditional classifiers such as SVM, KNN, and logistic regression. An ablation study confirmed the effectiveness of each component, and error analysis identified cases of spectral overlap and ambiguous speech as key factors contributing to misclassification. Future work will focus on expanding the dataset to include more diverse speakers and languages, thereby further improving generalization. Additionally, alternative deep learning architectures such as LSTM and transformer-based models will be explored to capture temporal dynamics more effectively. Finally, incorporating real-time testing and deployment into interactive systems remains a practical direction for validating the model's robustness in operational environments.

References

1. Al Arman Ovi, I.T.M. and M.J. Nayeem, A Machine Learning Approach for Identifying Gender Based on Bengali Vocal Cues. 2025.
2. Amiri, G.A., et al., Decoding Gender Representation and Bias in Voice User Interfaces (VUIs). 2024.
3. OPEYEMI, O.M., B.O. AYODEJI, and A.Y. BABATUNDE, DESIGN AND IMPLEMENTATION OF A GENDER-BASED FACIAL RECOGNITION SYSTEM USING MACHINE LEARNING MODEL. International Journal of Nature and Science Advance Research, 2025.
4. Liu, X., et al., Deep learning in spectral analysis: Modeling and imaging. TrAC Trends in Analytical Chemistry, 2024: p. 117612.
5. Gourisaria, M.K., et al., Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discover Internet of Things, 2024. 4(1): p. 1.
6. Katariya, P.P. and P.B. Pati. Leveraging Confidence Analysis and Classification Using BiLSTM for Verbal Evaluations. in 2024 5th IEEE Global Conference for Advancement in Technology (GCAT). 2024. IEEE.
7. Ali, A.S., et al., Lung Cancer Detection Using Convolutional Neural Networks from Computed Tomography Images. Journal of Computing & Biomedical Informatics, 2023. 6(01): p. 133-143.
8. Abdul, Z.K. and A.K. Al-Talabani, Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 2022. 10: p. 122136-122158.
9. Shue, Y.-L. and M. Iseli. The role of voice source measures on automatic gender classification. in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. 2008. IEEE.
10. Tauqeer, H., et al., Cyberattacks detection in iomt using machine learning techniques. Journal of Computing & Biomedical Informatics, 2022. 4(01): p. 13-20.
11. Bozkurt, F., A comparative study on classifying human activities using classical machine and deep learning methods. Arabian Journal for Science and Engineering, 2022. 47(2): p. 1507-1521.
12. Chaudhary, S. and D.K. Sharma. Gender identification based on voice signal characteristics. in 2018 International conference on advances in computing, communication control and networking (ICACCCN). 2018. IEEE.
13. Islam, M. GFCC-based robust gender detection. in 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET). 2016. IEEE.
14. Jia, Y., et al., Improving signal-to-noise ratio of Raman measurements based on ensemble learning approach. Analytical and Bioanalytical Chemistry, 2025. 417(3): p. 641-652.
15. Natsiou, A. and S. O'Leary. Audio representations for deep learning in sound synthesis: A review. in 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA). 2021. IEEE.
16. Uddin, M.A., et al. Gender recognition from human voice using multi-layer architecture. in 2020 International conference on innovations in intelligent systems and applications (INISTA). 2020. IEEE.
17. Hameed, S.H., et al., Advanced Next-Word Prediction: Leveraging Text Generation with LSTM Model. Journal of Computing & Biomedical Informatics, 2025. 8(02).
18. Kuchebo, A.V., et al. Convolution neural network efficiency research in gender and age classification from speech. in 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). 2021. IEEE.
19. Muhammad, S., et al., Malaria Cell Classification through Exercising Deep Learning Algorithms. Journal of Computing & Biomedical Informatics, 2024. 7(01): p. 53-61.
20. Carson, A., et al., Resampling Filter Design for Multirate Neural Audio Effect Processing. arXiv preprint arXiv:2501.18470, 2025.