Automatic SATB Singing Voice Classification Using Statistical Pseudo-Labelling and CNN-LSTM

Suwarno[1], Leon Agustian Salim[2], Tony Wibowo[3]

[1-3]Universitas Internasional Batam, Indonesia
✉ Correspondence e-mail: suwarno.liang@uib.ac.id

History of Article
Submitted : June 30, 2026
Revised : July 22, 2026
Accepted : August 10, 2026
Published : September 22, 2026
DOI : https://doi.org//10.37253/conescintech.v6i1.12799

Copyright© 2026 by Author(s). This work is licensed under a Creative Commons Attribution-Non Commercial-Share Alike 4.0 International License. All writings published in this journal are personal views of the authors and do not represent the views of this journal and the author's affiliated institutions.

Abstract

Accurate classification of singers into Soprano, Alto, Tenor, and Bass (SATB) voice categories is essential for choir performance and vocal health. However, existing automatic SATB classification methods are constrained by the limited availability of labeled datasets. This study proposes a statistically guided pseudo-labeling framework to expand labeled singing voice data and improve automatic SATB classification. Fundamental frequency (F0) was extracted from professionally labeled recordings in the Choral Singing Dataset (CSD) and ESMUC Choir Dataset using the YIN algorithm. Ten F0-derived statistical features were evaluated through statistical significance testing, effect size analysis, overlap analysis, and single-feature classification to identify the most discriminative feature for threshold-based pseudo-label generation. The selected feature was then used to assign SATB pseudo-labels to 3,093 VocalSet recordings, producing a combined dataset of 3,188 labeled recordings. The expanded dataset was segmented into four-second audio chunks, converted into Mel-spectrograms, and used to train and compare Convolutional Neural Network (CNN) and CNN–Long Short-Term Memory (CNN-LSTM) models. Experimental results identified the 75th percentile (P75) of F0 as the most discriminative statistic, achieving 98.95% single-feature classification accuracy. On the internal test set, the CNN-LSTM model outperformed the baseline CNN, achieving 94.6% accuracy and a macro F1-score of 0.92, compared with 92.6% accuracy and a macro F1-score of 0.90 for CNN. External evaluation on previously unseen YouTube recordings further demonstrated superior generalization, with CNN-LSTM achieving 65.4% accuracy versus 42.5% for CNN. These findings demonstrate that statistically guided pseudo-labeling effectively expands labeled training data and that integrating temporal modeling significantly improves automatic SATB singing voice classification.

Keywords: SATB classification; pseudo-labelling; fundamental frequency; Mel-spectrogram; CNN-LSTM.

Introduction

In Western choral singing, singers are classified into four voice categories, Soprano, Alto, Tenor, and Bass (SATB) each occupying a distinct frequency range that shapes a choir's harmonic balance (Chandna et al., 2022; Pratama et al., 2021). Accurate classification keeps singers within their physiological vocal capacity (Lycke et al., 2013); misclassification, by contrast, forces singers outside their natural range, raising the risk of vocal fatigue and long-term vocal disorders (Lycke et al., 2013; Meerschman et al., 2025). SATB classification therefore matters for both musical performance and vocal health.

Despite this importance, SATB classification is still performed manually by conductors or vocal instructors, relying on auditory perception and subjective experience rather than standardized acoustic measurement. Because evaluators differ in training, cultural background, and tonal perception, classifications are often inconsistent (Frizzell & Windsor, 2021), and no universally accepted objective method yet exists (Lycke et al., 2013), pointing to the need for more reliable, data-driven approaches.

Fundamental frequency (F0) is a key acoustic cue for distinguishing SATB categories. Among pitch-estimation methods, the YIN algorithm offers high accuracy with substantially lower error rates than conventional approaches, making it well suited to singing voice analysis (de Cheveigné & Kawahara, 2002). Deep learning further enables automatic extraction of discriminative acoustic features: Convolutional Neural Networks (CNNs) learn spectral representations from Mel spectrograms, while CNN-LSTM architectures additionally capture temporal dependencies, improving performance on audio classification tasks (Cuesta & Gómez, 2022b; Zaman et al., 2023)

Deep learning's performance, however, depends heavily on large, accurately labeled datasets. Existing choir datasets such as CSD and the ESMUC Choir Dataset are expert-labeled but small, while VocalSet offers many more solo recordings but lacks SATB annotations making labeled-data scarcity a major obstacle. Threshold-based pseudo-labeling has emerged as an effective way to expand labeled data while reducing manual annotation effort (Vishwakarma et al., 2024), but prior work has focused mainly on the labeling process itself, with little attention to which F0 statistic best supports accurate pseudo-label generation.

This study therefore proposes a statistically guided pseudo-labeling framework for automatic SATB classification. F0 is first extracted from expert-labeled recordings using YIN, and several F0-derived statistics are compared to identify the most discriminative one; this statistic is then used to pseudo-label additional recordings, producing a substantially larger training set. CNN and CNN-LSTM models are subsequently trained and compared on this expanded dataset. The contributions are threefold: (1) identifying the most discriminative F0 statistic for SATB pseudo-labeling, (2) proposing a statistically guided pseudo-labeling framework to expand SATB-labeled data, and (3) comparatively evaluating CNN and CNN-LSTM for automatic SATB classification on the expanded dataset.

Research Method

Research Design

This study uses a quantitative experimental design to develop and evaluate an automatic SATB classification framework based on deep learning. The framework has two stages: the first constructs a larger labelled dataset through objective pseudo-label generation from fundamental frequency (F0) analysis; the second trains and compares two architectures CNN and CNN-Long Short-Term Memory (CNN-LSTM) for SATB classification. The workflow is shown in Figure 1.

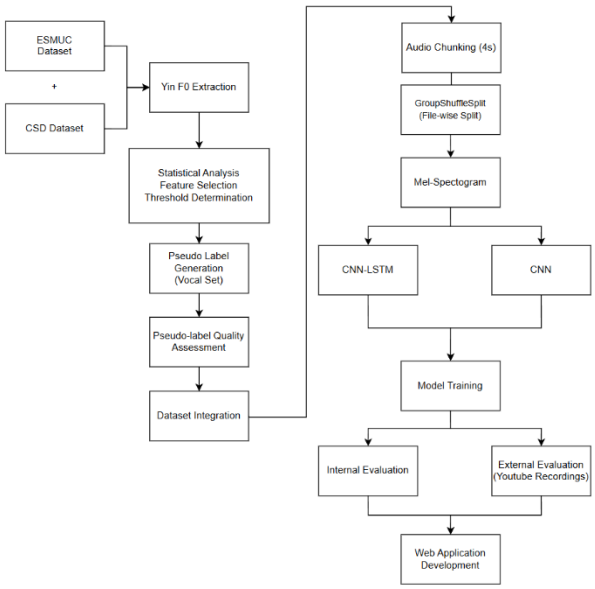


Figure 1. *Research Workflow*

Dataset

This study uses three public singing voice datasets, each with a distinct role. The Choral Singing Dataset (CSD) and ESMUC Choir Dataset (ECD) serve as reference datasets, providing professionally recorded SATB performances with verified voice-part annotations (Cuesta et al., 2019; Cuesta & Gómez, 2022a; Wilkins et al., 2018).

Only isolated voice recordings were used, ensuring each extracted F0 represents a single singer without interference from other choir parts. Although the full ESMUC dataset has some microphone leakage between singers, its isolated tracks remain reliable reference material while preserving realistic choir recording conditions (Chandna et al., 2022; Cuesta & Gómez, 2022a).

Dataset	Role	Total Recordings	SATB Labels
CSD (Choral Singing Dataset)	Reference (threshold derivation)	47	Verified (ground truth)
ESMUC Choir Dataset	Reference (threshold derivation)	48	Verified (ground truth)
Vocal Set	Target (pseudo-labelling)	3,093	Pseudo-labelled

Tabel 1. *Dataset Summary*

The verified SATB labels provided in the CSD and ESMUC Choir Dataset are used exclusively for statistical analysis and threshold estimation during pseudo-label generation. In contrast, VocalSet recordings are automatically assigned SATB labels using the proposed framework and subsequently incorporated into the model training process.

Proposed Framework

The proposed framework combines statistical acoustic analysis with deep learning-based classification. As shown in Figure 1, the workflow begins by extracting F0 from the reference datasets using YIN, then analyzing several F0 statistics to find the one that best separates SATB categories.

The selected statistic is then used to pseudo-label VocalSet via threshold-based classification. Before integration, the generated labels are checked for class distribution, ambiguous samples, and potential pitch-extraction errors, then merged with the original SATB datasets to form a larger training corpus.

Each recording is then segmented into fixed-duration chunks and converted into Mel-spectrogram representations, split into training, validation, and testing subsets using a group-based strategy. CNN and CNN-LSTM models are trained and evaluated under identical settings to compare their effectiveness for SATB classification.

Pseudo-Label Generation

1. Fundamental Frequency Extraction

   F0 was extracted from every recording using the YIN algorithm in Librosa, chosen for its robust pitch estimation and substantially lower octave-error rate than conventional autocorrelation-based methods (de Cheveigné & Kawahara, 2002) and its wide adoption in singing voice analysis due to computational efficiency and accuracy (Chandna et al., 2022).

   Pitch estimation used each recording's native sampling rate without resampling, with detectable frequencies bounded between C2 (65.41 Hz) and C7 (2093 Hz), and Librosa's default frame length (2048 samples) and hop length (512 samples).
2. Statistical Feature Selection

After extracting frame-level F0 values, ten statistics were computed per recording: minimum, maximum, mean, median, standard deviation, and the 10th, 25th, 50th, 75th, and 90th percentiles.

Each statistic's discriminative power was assessed through descriptive visualization, distribution comparison, one-way ANOVA, Kruskal-Wallis testing, effect size analysis, overlap analysis between adjacent voice categories, and single-feature classification jointly capturing both statistical separability and practical discriminative value.

3. Threshold Determination

Threshold values were estimated from the P75 distributions of the CSD and ESMUC reference datasets. Because male and female voices occupy substantially different F0 ranges, thresholds were determined separately for each gender rather than globally (Chandna et al., 2022; de Cheveigné & Kawahara, 2002). This strategy reduces overlap between adjacent SATB categories and provides a more reliable basis for threshold-based classification. Since VocalSet already provides gender information through its directory structure, the corresponding gender-specific threshold was directly applied during the pseudo-labeling process. The numerical threshold values obtained from this analysis are presented in the Results and Discussion section.

4. Automatic Pseudo-Label Generation

Using the derived thresholds, every VocalSet recording was automatically assigned to one of the four SATB categories. Recordings unsuitable for reliable F0 estimation, such as spoken voice, lip trills, vocal fry, whistle tones, growl phonation, and inhaled phonation, were excluded beforehand, as they do not represent sustained singing.

These labels are pseudo-labels rather than manually verified ground truth; however, prior work shows deep learning models can learn effectively despite moderate label noise (Cochoy, 2023).

5. Pseudo-Label Quality Assessment

Before incorporating the pseudo-labeled VocalSet recordings into the training dataset, their quality was assessed through three complementary analyses. First, the class distribution was examined to identify potential class imbalance. Second, the distance between each recording's selected F0 statistic and its corresponding decision threshold was analyzed as an indicator of labeling confidence. Third, recordings with extreme F0 values were inspected to identify possible pitch estimation errors or non-vocal artifacts. The outcomes of these analyses are reported in the Results and Discussion section.

Feature Extraction and Preprocessing

After pseudo-label generation, recordings from VocalSet, CSD, and ESMUC were merged into a single dataset and segmented into non-overlapping four-second chunks, with each chunk inheriting the SATB label of its source recording. The chunks were

converted into log-scaled, normalized Mel-spectrograms using 128 Mel filter banks, an FFT size of 2048 samples, and a hop length of 512 samples. Mel-spectrograms were selected because they closely represent human auditory perception and have demonstrated strong performance as input features for CNN-based audio classification (Bashir et al., 2025; Zaman et al., 2023). To prevent data leakage, the dataset was partitioned using GroupShuffleSplit based on the original recordings, ensuring that chunks from the same recording were never assigned to different subsets.

CNN and CNN-LSTM Architecture

The first model is a conventional CNN with three convolutional blocks, each containing a convolution layer followed by batch normalization, ReLU activation, max-pooling, and dropout. Feature channels increase progressively from 32 to 64 to 128 before flattening into fully connected layers for four-class SATB classification. CNNs have demonstrated strong capability in learning discriminative spectral representations directly from Mel-spectrograms (Pratama et al., 2021; Zaman et al., 2023).

The second model extends the CNN encoder with a Long Short-Term Memory (LSTM) layer after convolutional feature extraction. The convolutional feature map is reshaped into a temporal sequence and processed by a single-layer LSTM with 128 hidden units, whose final hidden state is passed to the classifier. This hybrid architecture models local spectral and temporal information jointly within singing voice signals (Cuesta & Gómez, 2022b).

Experimental Setup

The dataset was split into training (70%), validation (15%), and testing (15%) subsets using a group-based strategy. All experiments employed the Adam optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}, while weighted cross-entropy loss, computed from the training set's class distribution, was used to mitigate class imbalance (Zaman et al., 2023). Training was conducted for up to 100 epochs with a batch size of 32. To improve generalization, early stopping with a patience of 10 epochs was applied to prevent overfitting, and the ReduceLROnPlateau scheduler automatically reduced the learning rate when the validation Macro F1-score ceased to improve. Finally, the model achieving the highest validation Macro F1-score was retained for evaluation on the test set.

Performance Evaluation

Performance was evaluated using accuracy, precision, recall, Macro F1-score, confusion matrices, and class-wise classification reports, with Macro F1-score serving as the primary evaluation metric because it assigns equal importance to each voice category despite class imbalance (Zaman et al., 2023). To further assess the model's generalization capability, external validation was performed using previously unseen YouTube singing recordings. These recordings were segmented into four-second chunks, with silent

segments removed through RMS-based filtering. Prediction probabilities from all chunks were then averaged to generate a final recording-level SATB prediction, enabling evaluation of the model's robustness under real-world recording conditions.

Web Application Development

To demonstrate practical applicability, the best-performing model was deployed as a FastAPI web application. It accepts user-uploaded WAV recordings, automatically performs chunking and Mel-spectrogram extraction, and predicts the SATB category through a graphical interface enabling rapid, accessible SATB classification for choir practitioners and music educators.

Results and Discussion

Selection of the Most Discriminative F0 Statistic

The first objective of this study was to determine which F0-derived statistic provides the strongest discrimination among SATB voice categories before pseudo-label generation. Ten statistics were initially computed. Based on F0 statistic analysis, four candidate statistics (mean, median, P25, and P75) showing the highest discriminative potential were selected for detailed evaluation and evaluated using one-way ANOVA, Kruskal–Wallis testing, effect size analysis, overlap analysis, and single-feature classification.

Feature	ANOVA F	η^2	Mean Overlap	Classification Accuracy
Mean	271.91	0.9	0.194	88.42%
Median	321.16	0.914	0.139	90.53%
P25	129.55	0.81	0.095	96.84%
P75	442.99	0.936	0.064	98.95%

Table 2. *Summarizes F0 the comparative results.*

Among all evaluated statistics, the 75th percentile (P75) consistently achieved the best overall performance. It produced the highest ANOVA F-statistic (442.99), indicating the greatest between-class variation relative to within-class variation. Similarly, P75 obtained the largest effect size ($\eta^2 = 0.936$), suggesting that almost all variance in the feature can be explained by differences between SATB categories. In addition, P75 yielded the smallest average overlap coefficient (0.064) between adjacent voice categories and the highest single-feature classification accuracy (98.95%).

Pseudo-Label Result

Applying the selected P75 thresholds to VocalSet produced pseudo-labels for 3,093 recordings. The resulting distribution consisted of 465 Bass recordings (15.03%), 1,242 Tenor recordings (40.16%), 356 Alto recordings (11.51%), and 1,030 Soprano recordings (33.30%).

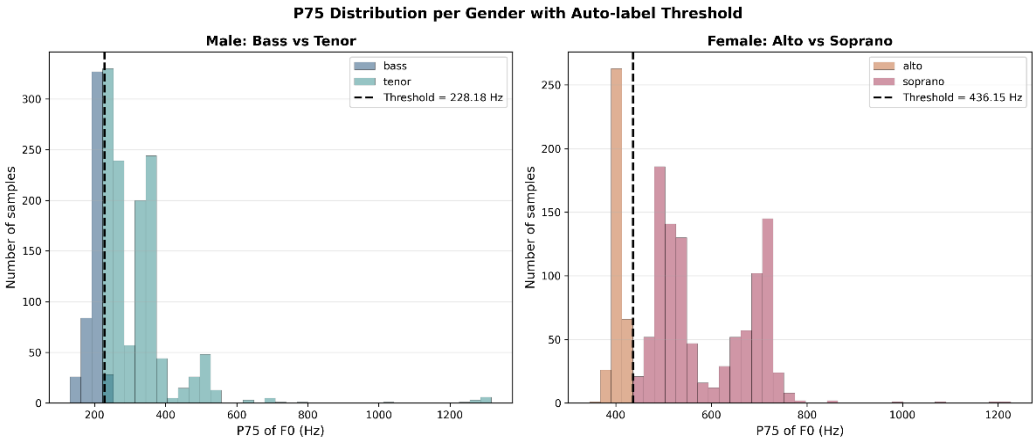


Figure 2. *P75 Distribution Per Gender Pseudo-Labels*

Although the dataset exhibits moderate class imbalance, with an imbalance ratio of approximately 3.49 between the largest and smallest classes, weighted cross-entropy loss was employed during model training to reduce the influence of this imbalance.
Only 2.0% of the recordings were located within ±5 Hz of the decision thresholds, indicating that the vast majority of samples were assigned with relatively high confidence. The average distance from the thresholds was 98.71 Hz, while the median distance was 59.98 Hz.

Outlier inspection also revealed only a limited number of recordings with unusually high or low P75 values. These samples were manually reviewed to identify possible octave estimation errors or recordings containing non-standard vocal techniques. Overall, the quality assessment suggests that the generated pseudo-labels provide a sufficiently reliable training corpus for supervised deep learning.

Performance of CNN and CNN-LSTM

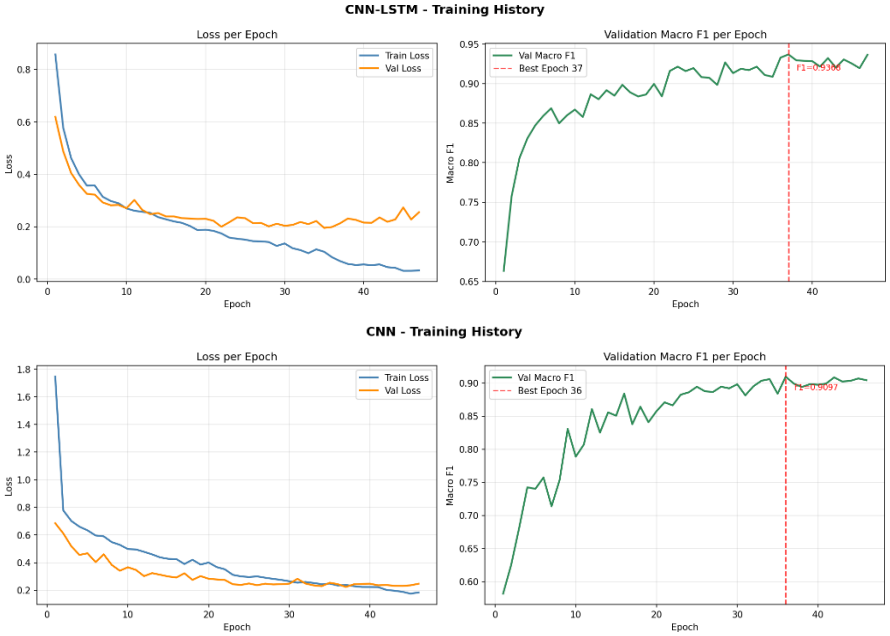


Figure 3. *Training History CNN and CNN-LSTM*

Figure 3 presents the training history of both architectures. Both models converged stably without severe overfitting, as indicated by the decreasing training loss and relatively stable validation loss. The validation Macro F1-score increased rapidly during the early epochs before reaching convergence, demonstrating that Mel-spectrogram representations effectively captured discriminative spectral features for SATB classification. Similar behavior has been reported in previous singing voice classification studies (Bashir et al., 2025; Pratama et al., 2021).

The CNN model reached its best validation Macro F1-score of 0.9097 at epoch 36, while CNN-LSTM achieved 0.9368 at epoch 37. Although CNN converged slightly faster, CNN-LSTM consistently produced better overall performance by incorporating temporal information.

Metric	CNN	CNN-LSTM
Accuracy	92.60%	94.60%
Macro Precision	0.89	0.93
Macro Recall	0.91	0.92
Macro F1-score	0.9	0.92

Table 3. *Comparison between CNN and CNN-LSTM*

As shown in Table 3, CNN-LSTM outperformed CNN across all evaluation metrics, improving the overall accuracy by 2.0% and the Macro F1-score from 0.90 to 0.92. The improvement suggests that temporal modeling complements the spectral features extracted by CNN, consistent with previous hybrid CNN-LSTM studies (Van et al., 2019; Zaman et al., 2023).

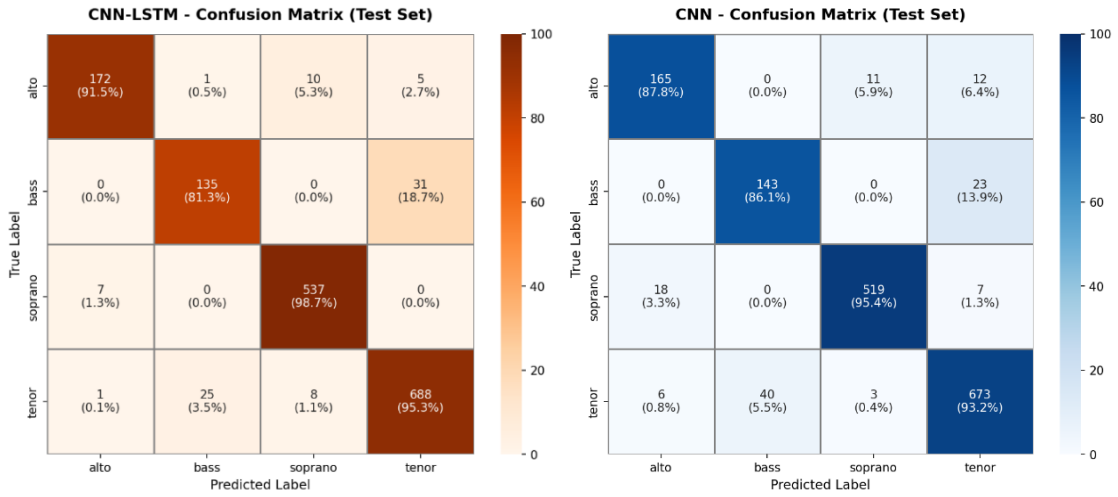


Figure 4. *Confusion Matrix CNN vs CNN-LSTM*

The confusion matrices indicate that Bass–Tenor confusion remained the dominant source of error in both models, while CNN-LSTM reduced misclassification involving the Alto class. This result suggests that temporal information improves discrimination between neighboring voice categories.

On the external YouTube dataset, both models experienced performance degradation due to domain differences between the controlled training recordings and real-world audio. Most YouTube recordings were captured using a wide variety of recording equipment with inconsistent audio quality and had often undergone post-processing, such as equalization, compression, noise reduction, and other editing techniques. These factors altered the original spectral and temporal characteristics of the singing voices, making them more challenging to classify accurately. Despite this domain shift, the CNN-LSTM model generalized substantially better, achieving 65.4% accuracy compared with 42.5% for the CNN model. This result suggests that incorporating temporal modeling enables the CNN-LSTM architecture to better handle recording variability and preserve discriminative voice characteristics than relying on spectral features alone (Zaman et al., 2023).

Discussion

The comparison between the two architectures indicates that incorporating temporal modeling through an LSTM layer provides a measurable benefit over relying on convolutional features alone, with accuracy improving from 92.6% to 94.6% and macro F1-score from 0.90 to 0.92 on the internal test set. This improvement is consistent with the hypothesis that sequential dependencies across frames, such as pitch contour shifts and vibrato, carry information relevant to voice type that a purely spatial convolutional representation does not fully capture. This finding aligns with prior work demonstrating that recurrent architectures can extract additional discriminative value from temporal vocal patterns (Van et al., 2019) and reinforces the broader pattern in audio classification literature where convolutional and recurrent components are often combined to capture complementary local and sequential structure (Zaman et al., 2023).
Despite this improvement, Bass remains the most difficult class to classify correctly across both architectures, with the lowest F1-score in each case (0.82 for CNN, 0.83 for CNN-LSTM). Two compounding factors likely contribute to this result. First, Bass is the second-smallest class in the combined dataset (483 samples, compared to 1,275 for Tenor), giving the model comparatively fewer examples from which to learn its characteristic acoustic patterns despite class weighting during training. Second, the confusion matrices for both models show that the majority of Bass misclassifications involve Tenor specifically, consistent with the acoustic proximity between Bass and Tenor ranges noted in prior SATB classification research (Cuesta & Gómez, 2022b).

The external evaluation results reveal a particularly important finding regarding generalization. The baseline CNN shows severe performance degradation, collapsing to predict almost exclusively Bass on external recordings (overall accuracy 42.5%). This behavior is consistent with majority-class bias: without the regularizing effect of temporal context, the CNN relies on low-level spectral features that are less stable across different recording environments. The CNN-LSTM, by contrast, maintains substantially better performance (65.4%) by leveraging temporal patterns that appear more robust to acoustic

variation introduced by different microphones, room acoustics, and audio compression typical of YouTube recordings.

The reduction in Alto-related misclassification under CNN-LSTM (from a combined 23 errors in the CNN model to 15 in CNN-LSTM) is also notable, given that Alto is the smallest class overall (377 samples). This suggests that the performance gain from temporal modeling is not uniform across classes but appears to disproportionately benefit classes with fewer training examples, possibly because sequential patterns provide a more sample-efficient signal than spatial spectral patterns alone.

When compared to prior SATB classification work, the results obtained here are favorable. Pratama et al. (2021) reported effective SATB separation using convolutional architectures on time-frequency representations, consistent with the performance achieved by the baseline CNN in this study. The CNN-LSTM result of 94.6% accuracy is also comparable to the 92.84% accuracy reported by Van et al. (2019) for Bidirectional LSTM-based singer identification, despite the fact that SATB classification inherently involves greater inter-class acoustic overlap.

These results should nonetheless be interpreted with several limitations in mind. The pseudo-labeling pipeline relies exclusively on a single F0 statistic (P75) to assign voice type, and while this statistic was selected through a comprehensive statistical comparison, it does not capture timbral or formant-related characteristics known to contribute to voice classification. Prior work has shown that the spectral energy distribution within the singer's formant region differs systematically across voice types independently of fundamental frequency (Müller et al., 2022), suggesting that incorporating such features could further reduce Bass-Tenor confusion. Additionally, because VocalSet pseudo-labels were not validated against expert annotation, some proportion of label noise particularly near the classification threshold is expected to persist in the training data (Vishwakarma et al., 2024).

Conclusion

This study developed a SATB singing voice classification system that addresses the limited availability of labeled choir data through a statistically grounded pseudo-labeling pipeline. By extracting fundamental frequency from labeled choir recordings using the YIN algorithm and systematically comparing ten F0-derived statistics, the 75th percentile (P75) was identified as the most discriminative feature for separating Bass, Tenor, Alto, and Soprano categories. This feature was used to derive labeling thresholds applied to pseudo-label VocalSet, which was then combined and merge with the original choir recordings to produce a training corpus of 3,188 labeled recordings.

Two CNN-based architectures were trained and compared. The baseline CNN achieved 92.6% accuracy and macro F1 of 0.90 on the internal test set; the CNN-LSTM variant achieved 94.6% accuracy and macro F1 of 0.92. External evaluation on YouTube-sourced recordings further differentiated the architectures: CNN collapsed to majority-class bias (42.5% accuracy), while CNN-LSTM maintained substantially better generalization (65.4% accuracy), demonstrating that temporal modeling captures features

more invariant to recording conditions than spectral features alone. Across both architectures, Bass consistently emerged as the most difficult class to classify, with the majority of errors attributable to confusion with the acoustically adjacent Tenor category. These findings indicate that a statistically validated, threshold-based pseudo-labeling strategy can meaningfully expand training data for SATB classification without requiring manual annotation or an auxiliary labeling model, and that combining convolutional and recurrent components yields tangible performance gains across both internal and external evaluation conditions.

Several directions remain open for future work. First, incorporating timbral or formant-based acoustic features alongside F0 may help resolve the persistent Bass-Tenor confusion, as these characteristics vary systematically across voice types independently of pitch (Müller et al., 2022). Second, validating a subset of VocalSet pseudo-labels against expert annotation would allow actual label noise to be quantified directly. Third, evaluating the system on a larger and more diverse external dataset would provide a stronger indication of real-world generalization. Finally, further architectural comparisons, including attention-based temporal models, may help clarify which aspects of sequential information are most relevant to SATB classification.

Reference

Bashir, R. N., Shahid, M. A., Rashid, T., Faheem, M., Saidani, T., Saidani, O., & Khan, A. R. (2025). Voice pathology identification using mel spectrogram features and deep learning. *Signal, Image and Video Processing*, *19*(11), 909. https://doi.org/10.1007/s11760-025-04527-4

Chandna, P., Cuesta, H., Petermann, D., & Gómez, E. (2022). A Deep-Learning Based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles. *Frontiers in Signal Processing*, *2*(April), 1–15. https://doi.org/10.3389/frsip.2022.808594

Cochoy, J. (2023). *Human Voice Pitch Estimation: A Convolutional Network with Auto-Labeled and Synthetic Data*. 1–11. http://arxiv.org/abs/2308.07170

Cuesta, H., & Gómez, E. (2022a). *ESMUC Choir Dataset*. Zenodo. https://doi.org/10.5281/zenodo.5848990

Cuesta, H., & Gómez, E. (2022b). Voice Assignment in Vocal Quartets Using Deep Learning Models Based on Pitch Salience. *Transactions of the International Society for Music Information Retrieval*, *5*(1), 99. https://doi.org/10.5334/tismir.121

Cuesta, H., Gómez, E., Martorell, A., & Loáiciga, F. (2019). *Choral Singing Dataset*. Zenodo. https://doi.org/10.5281/zenodo.2649950

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930. https://doi.org/10.1121/1.1458024

Frizzell, E. Y., & Windsor, L. C. (2021). Effects of teaching experience and culture on choral directors' descriptions of choral tone. *PLOS ONE*, *16*(12), e0256587. https://doi.org/10.1371/journal.pone.0256587

Lycke, H., De Jong, F., Decoster, W., & Van Hulle, M. (2013). *Identification of three natural voice Groups By Phonetography - A data driven approach.* https://www.researchgate.net/profile/Hugo-Lycke/publication/364029030_Voice_Classification/links/63369eaf769781354ea83bc2/Voice-Classification.pdf

Meerschman, I., D'haeseleer, E., Cammu, H., Kissel, I., Papeleu, T., Leyns, C., Daelman, J., Dannhauer, J., Vanden Abeele, L., Konings, V., Demarbaix, E., & Van Lierde, kristiane. (2025). Voice Quality of Choir Singers and the Effect of a Performance on the Voice. *Journal of Voice*, *39*(2), 561.e1-561.e10. https://doi.org/10.1016/j.jvoice.2022.08.017

Müller, M., Wang, Z., Caffier, F., & Caffier, P. P. (2022). New objective timbre parameters for classification of voice type and fach in professional opera singers. *Scientific Reports*, *12*(1), 17921. https://doi.org/10.1038/s41598-022-22821-w

Pratama, K. B., Suyanto, S., & Rachmawati, E. (2021). Human Vocal Type Classification using MFCC and Convolutional Neural Network. *2021 International Conference on Communication & Information Technology (ICICT)*, 43–48. https://doi.org/10.1109/ICICT52195.2021.9568474

Van, T. P., Quang, N. T. N., & Thanh, T. M. (2019). Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music. *Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019*, 255–260. https://doi.org/10.1145/3368926.3369700

Vishwakarma, H., Lin, H., Sala, F., & Vinayak, R. K. (2024). *Promises and Pitfalls of Threshold-based Auto-labeling*. http://arxiv.org/abs/2211.12620

Wilkins, J., Wahl, A., Pardo, B., & Science, C. (2018). *VocalSet: a singing voice dataset*. 468–474. https://doi.org/10.5281/zenodo.1193957

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE Access*, *11*(September), 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Generative AI Statement

During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.5) and Google Gemini 2.5 Flash to assist with language refinement, improvement of academic writing, grammar correction, manuscript organization, and drafting of non-analytical text. These tools were also used to summarize relevant literature to facilitate the writing process. All research design, statistical analyses, experimental implementation, data processing, interpretation of results, and final conclusions were performed and verified by the authors. The outputs generated by the AI tools were critically reviewed, revised, and validated to ensure accuracy, consistency, and scientific

integrity. The authors take full responsibility for the final content of this manuscript. AI tools were used solely as writing support and are not credited with authorship.

E-ISSN: 2808-5485
Available online at website: https://journal.uib.ac.id/index.php/conescintech/index