Speech-Based Automatic Chronic Kidney Disease Diagnosis via Transformer Fusion of Glottal and Spectrogram Features

Jihyun Mun[1], Sunhee Kim[2], Minhwa Chung[1]

[1]Department of Linguistics, Seoul National University, Republic of Korea
[2]Department of French Language Education, Seoul National University, Republic of Korea
{jhhh_1202, sunhkim, mchung}@snu.ac.kr

Abstract

Chronic kidney disease (CKD) is a global health concern characterized by a gradual and irreversible decline in kidney function. Early diagnosis and timely intervention are crucial, yet current methods rely primarily on invasive blood and urine tests. Since CKD affects the respiratory system and alters speech production, vocal characteristics may serve as biomarkers for disease detection. This study proposes a deep learning-based approach that integrates spectrogram and glottal features for CKD diagnosis. Spectrograms capture broad acoustic characteristics, whereas glottal features, known to be influenced by CKD, provide complementary phonatory information. To effectively fuse these features, we employ a transformer-like architecture. The proposed method achieves an accuracy and a macro F1 score of 0.96, demonstrating its potential as an objective, non-invasive diagnostic tool. In addition, we analyze attention weights and gradient-based saliency maps to enhance model interpretability.

Index Terms: chronic kidney disease, automatic diagnosis, spectrogram, glottal characteristic, voice pathology

1. Introduction

Chronic kidney disease (CKD) is a significant global health concern, with rising morbidity and mortality rates worldwide [1]. It is characterized by a gradual and irreversible decline in kidney function, often accompanied by structural damage. Although CKD progresses over months or years, its early stages are frequently asymptomatic, which makes timely diagnosis and intervention challenging [2]. One major challenge in CKD management is its frequent underdiagnosis and inadequate treatment. Early detection and proper management of comorbidities not only reduce overall and cardiovascular mortality rates but also improve patients' quality of life.

Currently, CKD diagnosis primarily relies on detecting albuminuria and estimating the glomerular filtration rate (eGFR) through serum creatinine levels, both of which require invasive blood and urine tests [3]. However, the increasing prevalence of CKD is outpacing the capacity of available nephrologists and specialized healthcare facilities, highlighting the need for scalable, automated systems to support early detection. Although machine learning and deep learning approaches have been explored for CKD diagnosis and prediction, most models still depend on medical data obtained through blood tests and other clinical evaluations, necessitating hospital visits and regular checkups [4, 5, 6].

Recent studies suggest that CKD affects multiple physiological systems, including the respiratory system, which significantly influences speech production [7]. CKD patients exhibit reduced respiratory muscle strength and endurance compared to healthy individuals, as well as impaired laryngeal function [7, 8]. In end-stage renal disease (ESRD), uremic toxin accumulation, acid-base imbalances, and fluid overload can exacerbate vocal fold edema and lung dysfunction, resulting in noticeable changes in voice quality [9]. Given that respiration plays a fundamental role in speech production [10], the vocal characteristics of CKD patients may serve as indicators of disease presence and progression.

Several studies employ speech analysis to automatically detect or stage CKD using machine learning models [11, 12]. These methods typically rely on manually extracted acoustic features considered relevant to CKD. Although such approaches are interpretable, they require extensive feature engineering that is time-consuming and depends on domain expertise [13]. In contrast, deep learning models can learn meaningful representations directly from raw data, circumventing the constraints of manual feature extraction.

In this study, we propose a method that combines the strengths of both conventional acoustic feature extraction and deep learning-based approaches to enable automatic CKD diagnosis using speech data. Building on recent advancements in spectrogram-based disease detection [14, 15, 16], we utilize spectrogram representations of speech as an input. Furthermore, as previous work demonstrates that CKD affects glottal function [12], we explicitly incorporate glottal features that are not fully captured in spectrograms. To effectively integrate these two feature sets, we introduce a transformer-like architecture that leverages self-attention to model their interactions. In addition, we employ a seed ensemble strategy to reduce performance variability arising from random initialization and to improve overall reliability. Finally, we analyze the model's decision-making process by examining key features that drive predictions and their relevance to CKD characteristics.

The remainder of this paper is structured as follows: Section 2 details the methodologies employed in this study, Section 3 describes the experimental setup and presents the results, Section 4 discusses the findings, and Section 5 concludes the paper.

2. Methods

Figure 1 illustrates the overall workflow of the proposed method, consisting of two main components: (i) feature extraction and (ii) a Transformer-based classification model.

2.1. Feature Extraction

2.1.1. Spectrogram Features

A spectrogram provides a time-frequency representation of an audio signal, capturing both temporal and spectral characteristics of speech. In this study, we compute Mel spectrograms using the short-time Fourier transform (STFT). Each audio sig-
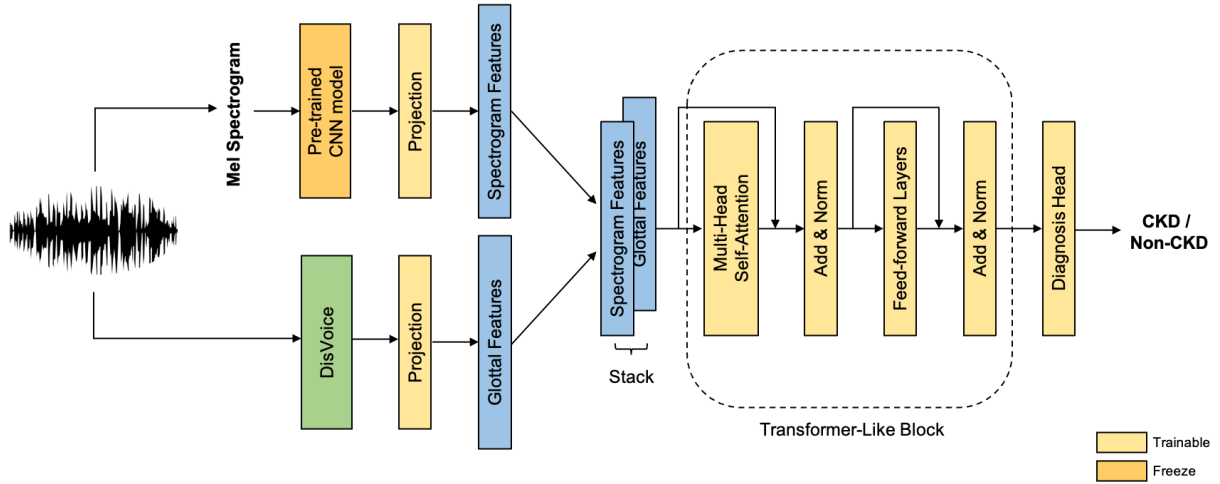
Figure 1: *Overview of the proposed method for automatic CKD diagnosis*

nal is segmented into overlapping windows, transformed into frequency components, and then mapped onto the mel scale to better align with human auditory perception. We zero-pad each spectrogram based on the maximum signal duration in the dataset. After normalization, we process the spectrograms using a pretrained ResNet-18 model [17], originally trained on large-scale image datasets. ResNet-18 has demonstrated strong performance in previous speech pathology detection tasks [18, 19]. To fully utilize the inherent knowledge of the pre-trained model, all weights are frozen, requiring no additional training.

2.1.2. Glottal Features

In addition to spectrogram-based features, we extract nine glottal parameters to capture phonatory characteristics that may not be fully represented in spectrograms. Since CKD affects both respiratory and laryngeal function, variations in these glottal parameters may indicate disease status [12]. The Disvoice toolkit [20] is employed to extract the following nine glottal parameters:

- Variability of time between consecutive glottal closure instants (GCI) (*var GCI*)
- Average and variability of opening quotient (QOQ) for consecutive glottal cycles: (opening phase duration) / (glottal cycle duration) (*avg QOQ, std QOQ*)
- Average and variability of normalized amplitude quotient (NAQ) for consecutive glottal cycles: ratio of the amplitude quotient and the duration of the glottal cycle (*avg NAQ, std NAQ*)
- Average and variability of H1H2: difference between the first two harmonics of the glottal flow signal (*avg H1H2, std H1H2*)
- Average and variability of harmonic richness factor (HRF): ratio of the sum of the harmonics amplitude to the amplitude of the fundamental frequency (*avg HRF, std HRF*)

All features are normalized before being fed into the classification model.

2.2. Transformer-based Classification Model

2.2.1. Feature Fusion

We fuse spectrogram and glottal features by projecting both sets into 128-dimensional embeddings. These embeddings are then stacked to form a two-token input sequence, which we feed into a Transformer-like block. Structuring the input as a sequence allows the self-attention mechanism to learn how the two feature sets interact, thus enhancing the model's ability to identify CKD-related vocal characteristics.

2.2.2. Transformer-Like Block

The Transformer-based architecture leverages a multi-head self-attention (MHSA) mechanism to capture dependencies between feature tokens [21]. We implement:

1. Multi-Head Self-Attention (MHSA): Allows each feature token to attend to information in the other token, thus modeling relationships between spectrogram and glottal features.
2. Feed-Forward Layers: Process the output of MHSA through residual connections, layer normalization, and ReLU activation to stabilize training and refine feature representations.

This Transformer-based approach enables effective integration of spectrogram and glottal embeddings, improving CKD classification robustness.

2.2.3. Diagnosis Head

The output of the Transformer block is passed through a fully connected layer that generates a single logit. This logit is then transformed into a probability score indicating the presence or absence of CKD. By leveraging self-attention in Transformer-like block, the model effectively combines acoustic and phonatory features into a final diagnostic output.

3. Experiments

3.1. Data

We employ the CKD dataset introduced by [22], which is designed for pathological voice analysis, automatic disease diagnosis, and severity prediction. The dataset includes sustained vowel phonations, a sentence composed entirely of voiced

sounds, and a paragraph containing six phonetically balanced sentences of varying lengths. For this study, we focus on the sustained vowel /a/ utterances, given their stability and suitability for acoustic analysis. The dataset is divided into training, validation, and test sets, with an 8:1:1 ratio in a speaker-independent manner. Table 1 summarizes the number of samples and total duration in the training, validation, and test sets. Detailed information regarding CKD stages and participant demographics is available in [22].

Table 1: *Summary of dataset composition*

| | Non-CKD | CKD | | | | |
		1	2	3	4	5
Train	415	49	121	207	86	20
	(1h 11m 13s)	(13m 40s)	(34m 49s)	(56m 43s)	(20m 26s)	(4m 33s)
Validation	52	6	19	25	16	4
	(8m 33s)	(1m 47s)	(6m 3s)	(7m 30s)	(3m 41s)	(1m 15s)
Test	49	6	10	27	19	3
	(8m 25s)	(1m 49s)	(2m 30s)	(7m 44s)	(4m 45s)	(47s)

3.2. Experimental Details

3.2.1. Preprocessing

We use the Librosa library [23] to extract spectrogram representations from the raw audio data. The mel-spectrogram is generated with a sampling rate of 16,000 Hz, 128 mel bands, an FFT window length of 1024, a hop length of 160, a Hamming window function, and a frequency range of 0 to 8000 Hz. After extraction, we normalize the spectrograms and convert single-channel spectrograms to three-channel format for compatibility with models trained on RGB images.

3.2.2. Hyperparameter Settings

The hyperparameters used in the experiments are configured as follows:

- Model initialization: He initialization
- Optimizer: AdamW with an initial learning rate of 1e-4 and weight decay of 1e-2
- Loss function: Cross-entropy loss with class weighting
- Batch size: 16
- Number of attention heads: 4
- Dropout ratio: 0.1
- Projection layer: Maps extracted spectrogram and glottal features to 128-dimensional feature vectors
- Fully-connected layers in Transformer-like block
  - First layer: 128 input units → 256 output units (ReLU activation)
  - Second layer: 256 input units → 128 output units
- Number of epochs: 200 (Early stopping if validation loss does not improve for 20 consecutive epochs)
- Number of ensemble models: 5

3.2.3. Ensemble of Models

To enhance the stability and reliability of predictions, we train five independent models using different random initializations. We then average their predictions during inference. This ensemble strategy reduces variance, mitigates the effect of any single model's miscalibration, and improves overall classification performance.

3.2.4. Evaluation Metrics

We employ accuracy, macro F1-score, macro precision, and macro recall to evaluate the performance of our binary CKD classification (non-CKD vs. CKD). These metrics provide a balanced assessment across classes, which is critical for the imbalanced dataset.

3.3. Results

We conduct a binary classification experiment to detect CKD presence. Table 2 presents the classification results obtained under various configurations of input features and model components, as part of an ablation study designed to assess the contribution of each element. In the case of single feature type conditions, the baseline refers to a model configuration in which the Transformer block of the proposed architecture is removed.

Table 2: *CKD classification results across different settings*

Feature Type	Fusion Strategy	Accuracy	F1-score	Precision	Recall
Spectrogram	Baseline	0.9561	0.9556	0.9535	0.9590
	Transformer Block	0.9474	0.9463	0.9463	0.9463
Glottal	Baseline	0.8596	0.8586	0.8576	0.8644
	Transformer Block	0.8509	0.8495	0.8542	0.8481
Spectrogram + Glottal	Concatenation	0.9561	0.9554	0.9554	0.9565
	Transformer Block (Proposed)	**0.9649**	**0.9642**	**0.9642**	**0.9642**

Our proposed method, which integrates spectrogram and glottal features using a Transformer-based architecture, outperforms all other settings. Using only spectrogram features achieves high classification performance, confirming their strong effectiveness in detecting CKD. However, introducing a Transformer block to spectrogram-only features slightly decreases performance. Using only glottal features is considerably less effective, and applying a Transformer block to glottal-only features does not improve performance. When spectrogram and glottal features are simply concatenated, the performance remains similar to that of spectrogram-only features. However, integrating spectrogram and glottal features with the Transformer-based architecture improves classification accuracy, demonstrating the effectiveness of self-attention in capturing interactions between these feature sets. These findings suggest that the proposed feature-fusion strategy enhances CKD detection by leveraging complementary information from both spectral and glottal features.

3.4. Model Interpretation

In this section, we analyze the model's interpretability by examining attention weight distributions and performing gradient-based saliency analysis to identify the key features and relationships that drive the model's predictions.

3.4.1. Attention Weight Analysis

To understand how the model utilizes different features, we compute the attention weights of the final ensemble model by averaging the attention distributions across individual models. Figure 2 shows the attention weight distributions for both non-CKD and CKD data.

From Figure 2, we observe that for non-CKD speech data, the spectrogram component as the query references its own information 59% of the time and the glottal information 41%.
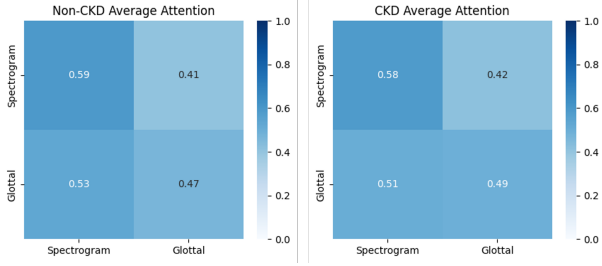
Figure 2: *Attention weight distribution of the model*

Similarly, the glottal component as the query references spectrogram 53% of the time and glottal 47%. These results suggest both spectrogram and glottal features contribute significantly to the classification of non-CKD cases, with a slightly stronger reliance on spectrogram features. A similar trend is observed for CKD data, but with a slight increase in attention toward glottal information, indicating that the model places slightly higher emphasis on glottal features when distinguishing CKD cases.

3.4.2. Gradient-Based Saliency Analysis

To further investigate the specific glottal features that influence the model's decisions, we conduct gradient-based saliency analysis [24]. This method involves computing the gradient of the loss function for each class with respect to the input glottal features, allowing us to identify the most influential factors driving classification.
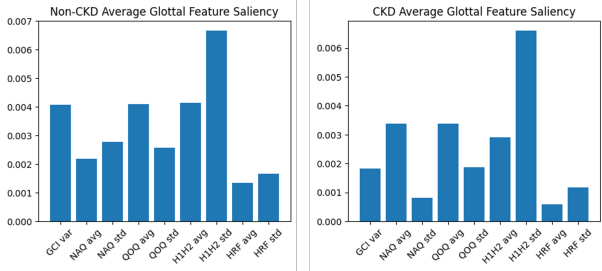

Figure 3: *Gradient-based saliency of glottal features*

Figure 3 presents the average gradient-based saliency values for the nine glottal features, comparing their impact on non-CKD (left) and CKD (right) predictions. In both cases, the standard deviation of H1H2 emerges as the most salient feature, indicating that fluctuations in H1H2 strongly influence classification decisions. For non-CKD classification, the variation of GCI, average QOQ, and average H1H2 serve as key predictive features. For CKD classification, the average NAQ, QOQ, and H1H2 are identified as the most influential features.

4. Discussions

We propose a method that integrates spectrogram and glottal features using a Transformer-like block and validate its effectiveness through a CKD classification task.

Our experiments reveal several noteworthy insights into how each feature set contributes to model performance. Specifically, spectrogram feature alone already achieve high classification performance in CKD detection, whereas glottal features alone yield comparatively lower performance; neverthe-

less, glottal features remain valuable and produce outcomes similar to certain machine learning-based approaches [12] despite using fewer features (17 vs. 9). Simply concatenating spectrogram and glottal features does not significantly enhance performance, indicating that a more sophisticated strategy for feature interaction is necessary.

Introducing a Transformer block to single-feature inputs also results in a performance decline, likely because the self-attention mechanism receives only one token and thus reverts to the original input without providing additional benefits. Moreover, this added complexity increases the number of model parameters and ultimately degrades performance. In contrast, integrating the Transformer block with both spectrogram and glottal features substantially improves classification results. This boost arises from the self-attention mechanism, which effectively fuses the two feature sets by modeling their interdependence. Thus, when glottal features are incorporated with spectrograms via self-attention, they provide complementary information that spectrograms alone do not capture, yielding performance gains beyond those achieved through simple concatenation.

Further, in clinical domain, model interpretability is of critical importance, as understanding the reasoning behind model predictions is essential for reliability and applicability [25]. To gain insights into the key features and relationships that contribute to the model's decision-making process, we conduct a detailed model prediction analysis. The attention weight analysis reveals that spectrogram features contribute the most to the model's predictions overall, while glottal features also play a substantial role, demonstrating the effectiveness of the proposed approach, which incorporates both features for robust classification.

The gradient-based saliency analysis further highlights that the standard deviation of H1H2 is the most influential feature for both CKD and non-CKD classifications. H1H2 measures the amplitude difference between the first and second harmonics of the voice signal, and its standard deviation indicates how much this difference varies over the utterance, reflecting voice instability. Thus, these findings suggest that voice instability during phonation serves as a key distinguishing factor between CKD and non-CKD speech. Given that individuals with CKD often experience difficulty controlling the muscles required for phonation and may struggle to maintain stable phonation due to compromised respiration-related muscles [12], the gradient saliency results suggest that the model successfully leverages these physiological characteristics for CKD detection.

5. Conclusions

This study proposes a method that integrates spectrogram and glottal features using a Transformer-like block for CKD diagnosis. With an accuracy and a macro F1-score of 0.96, our experimental results confirm the efficacy of combining spectrogram and glottal features via self-attention. Key contributions of this study include: 1) Demonstrating the effectiveness of combining spectral and glottal features for CKD diagnosis, 2) Explaining the model's decision-making process by analyzing attention weights and gradient saliency, identifying the most influential features and linking them to CKD-specific voice characteristics.

Future work aims to investigate which regions of the spectrogram inputs contribute most to the model's predictions and to explore how these regions correlate with CKD symptoms. We also plan to extend our approach to predict CKD severity, which would further enhance its clinical utility.

6. References

[1] J. Yang and W. He, *Chronic kidney disease: Diagnosis and treatment*. Springer, 2019.

[2] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *The lancet*, vol. 389, no. 10075, pp. 1238–1252, 2017.

[3] S. H. Kwon and D. C. Han, "Diagnosis and screening of chronic kidney disease," *The Korean Journal of Medicine*, vol. 76, no. 5, pp. 515–520, 2009.

[4] A. Ogunleye and Q.-G. Wang, "Enhanced xgboost-based automatic diagnosis system for chronic kidney disease," in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, 2018, pp. 805–810.

[5] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020.

[6] R. A. Alassaf, K. A. Alsulaim, N. Y. Alroomi, N. S. Alsharif, M. F. Aljubeir, S. O. Olatunji, A. Y. Alahmadi, M. Imran, R. A. Alzahrani, and N. S. Alturayeif, "Preemptive diagnosis of chronic kidney disease using machine learning techniques," in *2018 international conference on innovations in information technology (IIT)*. IEEE, 2018, pp. 99–104.

[7] E. S. Hassan, "Effect of chronic renal failure on voice: an acoustic and aerodynamic analysis," *The Egyptian Journal of Otolaryngology*, vol. 30, pp. 53–57, 2014.

[8] F. M. Abd El-gaber, Y. Sallam, and H. Mohammed Eid El Sayed, "Acoustic characteristics of voice in patients with chronic kidney disease," *International Journal of General Medicine*, pp. 2465–2473, 2021.

[9] S. Y. Jung, J.-H. Ryu, H. S. Park, S. M. Chung, D.-R. Ryu, and H. S. Kim, "Voice change in end-stage renal disease patients after hemodialysis: correlation of subjective hoarseness and objective acoustic parameters," *Journal of Voice*, vol. 28, no. 2, pp. 226–230, 2014.

[10] R. B. Kumar and J. S. Bhat, "Voice in chronic renal failure," *Journal of Voice*, vol. 24, no. 6, pp. 690–693, 2010.

[11] J. Mun, S. Kim, M. J. Kim, J. Ryu, S. Kim, M. Chung *et al.*, "Automatic detection and severity prediction of chronic kidney disease using machine learning classifiers," *Phonetics and Speech Sciences*, vol. 14, no. 4, pp. 45–56, 2022.

[12] J. Mun, S. Kim, M. J. Kim, J. Ryu, S. Kim, and M. Chung, "An analysis of glottal features of chronic kidney disease speech and its application to ckd detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 1573–1577.

[13] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial," *Machine learning*, vol. 113, no. 7, pp. 3917–3928, 2024.

[14] L. Zahid, M. Maqsood, M. Y. Durrani, M. Bakhtyar, J. Baber, H. Jamal, I. Mehmood, and O.-Y. Song, "A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease," *IEEE Access*, vol. 8, pp. 35 482–35 495, 2020.

[15] T. Zhang, Y. Zhang, Y. Cao, L. Li, and L. Hao, "Diagnosing parkinson's disease with speech signal based on convolutional neural network," *International Journal of Computer Applications in Technology*, vol. 63, no. 4, pp. 348–353, 2020.

[16] M. Chaiani, S. A. Selouani, M. Boudraa, and M. S. Yakoub, "Voice disorder classification using speech enhancement and deep learning models," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 2, pp. 463–480, 2022.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] A. Elfaki, A. L. Asnawi, A. Z. Jusoh, A. F. Ismail, S. N. Ibrahim, N. F. M. Azmin, and N. N. W. B. N. Hashim, "Using the short-time fourier transform and resnet to diagnose depression from speech data," in *2021 IEEE International Conference on Computing (ICOCO)*. IEEE, 2021, pp. 372–376.

[19] A. K. Dutta and A. R. Wahab Sait, "A speech disorder detection model using ensemble learning approach," *Journal of Disability Research*, vol. 3, no. 3, p. 20240026, 2024.

[20] E. A. Belalcázar-Bolanos, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Glottal flow patterns analyses for parkinson's disease detection: Acoustic and nonlinear approaches," in *International Conference on Text, Speech, and Dialogue*. Springer, 2016, pp. 400–407.

[21] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[22] J. Mun, S. Kim, M. J. Kim, J. Ryu, S. Kim, and M. Chung, "A speech corpus for chronic kidney disease," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2022, pp. 1–6.

[23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[25] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural computing and applications*, vol. 32, no. 24, pp. 18 069–18 083, 2020.