МРНТИ 20.01.07
УДК 378.091.12:004.451

10.51889/2959-5894.2023.82.2.022

DETECTION OF PARKINSON'S DISEASE PATIENTS BASED ON VOICE RECORDING USING CONVOLUTION NEURAL NETWORK

Hashim S.M. [1], *Kutucu H.* [1,2*], *Assanova B.*[2], *Shazhdekeyeva N.*[2], *Taishiyeva A.*[2]

[1] *Karabuk University, Karabuk, Turkey*
[2] *Atyrau State University, Atyrau, Kazakhstan*
**e-mail: hakankutucu@karabuk.edu.tr*

Abstract

Parkinson's disease is a commonly observed neurological disorder that affects the nervous system and hinders, people's essential functions. The primary goal of this study is to identify the presence of Parkinson's disease by utilizing spectrogram images from voice recordings through the implementation of Convolutional Neural Networks (CNN). We conducted our research using a dataset from the Argentina. Our research made a significant contribution by performing various audio preprocessing operations. We split the audio samples into multiple segments of the same duration (2 seconds) and then implement audio augmentation techniques to increase the dataset. Finally, we converted these audio samples into spectrogram images to train our model. K-fold cross-validation method was used, set by (k=10) for further analysis. The model underwent 150 epochs of training, resulting in an Average Training Accuracy of 99.3% and an Average Testing Accuracy of 97.9%. The effectiveness of the proposed model is compared with five state-of-art models (AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet) and the local binary pattern descriptors which were applied to the same dataset. As a result, the proposed model was found to be superior.

Keywords: Spectrogram, Parkinson's disease, voice analysis, CNN, k-fold cross-validation.

Аңдатпа

С. М. Хашим[1], *Х. Кутучу*[1,2], *Б.У. Асанова*[2], *Н.К. Шаждекеева* [2] , *А.Г. Тайшиева* [2]
[1] *Карабүк университеті, Карабүк, Түркия*
[2] *Атырау мемлекеттік университеті, Атырау қ., Қазақстан*

ПАРКИНСОН АУРУЫНА ҚҰРЫЛҒАН ПАЦИАЛАРДЫ КОНВОЛЮЦИЯЛЫҚ НЕЙРЛІК ЖЕЛІЛІК АРҚЫЛЫ ДАУЫС ЖАЗУ НЕГІЗІНДЕ АНЫҚТАУ

Паркинсон ауруы - бұл жүйке жүйесіне әсер ететін және адамдардың маңызды функцияларын бұзатын жиі байқалатын неврологиялық ауру. Бұл зерттеудің негізгі мақсаты конволюционды нейрондық желілерді (CNN) енгізу арқылы дауыстық жазбалардан спектрограммалық кескіндерді пайдалану арқылы Паркинсон ауруының болуын анықтау болып табылады. Біз зерттеуімізді Аргентинадан алынған деректер жиынтығы арқылы жүргіздік. Біздің зерттеуіміз дыбысты алдын ала өңдеудің әртүрлі операцияларын орындау арқылы айтарлықтай үлес қосты. Біз дыбыс үлгілерін бірдей ұзақтықтағы бірнеше сегменттерге (2 секунд) бөлеміз, содан кейін деректер жинағын ұлғайту үшін дыбысты кеңейту әдістерін енгіземіз. Соңында, біз модельімізді үйрету үшін осы аудио үлгілерді спектрограмма кескіндеріне айналдырдық. Әрі қарай талдау үшін (k=10) белгілеген K-есептік кросс-валидация әдісі қолданылды. Модель оқытудың 150 дәуірінен өтті, нәтижесінде жаттығудың орташа дәлдігі 99,3% және орташа тестілеу дәлдігі 97,9% болды. Ұсынылған модельдің тиімділігі бес заманауи үлгімен (AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet) және бір деректер жиынына қолданылған жергілікті екілік үлгі дескрипторларымен салыстырылады. Нәтижесінде ұсынылған модельдің артықшылығы анықталды.

Түйін сөздер: спектрограмма, болезнь Паркинсона, голосовой анализ, CNN, k-кратная перекрестная проверка.

Аннотация

С. М. Хашим[1], *Х. Кутучу*[1,2], *Б.У. Асанова*[2], *Н.К. Шаждекеева* [2] , *А.Г. Тайшиева* [2]
[1] *Университет Карабук, г. Карабук, Турция*
[2] *Атырауский государственный университет, г. Атырау, Казахстан*

ВЫЯВЛЕНИЕ ПАЦИЕНТОВ С БОЛЕЗНЬЮ ПАРКИНСОНА НА ОСНОВЕ ЗАПИСИ ГОЛОСА С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Болезнь Паркинсона – это часто наблюдаемое неврологическое расстройство, которое поражает нервную систему и нарушает основные функции человека. Основная цель этого исследования – выявить наличие болезни Паркинсона путем использования изображений спектрограмм из записей голоса посредством внедрения сверточных нейронных сетей (CNN). Мы провели исследование, используя набор данных из Аргентины. Наши исследования внесли значительный вклад, выполнив различные операции предварительной обработки звука. Мы разделяем аудиосэмплы на несколько сегментов одинаковой продолжительности (2 секунды), а затем реализуем методы улучшения звука, чтобы увеличить набор данных. Наконец, мы преобразовали эти аудиосэмплы в изображения спектрограмм для обучения нашей модели. Для дальнейшего анализа использовался метод K-кратной перекрестной проверки, заданный значением (k=10). Модель прошла 150 эпох обучения, в результате чего средняя точность обучения составила 99,3%, а средняя точность тестирования – 97,9%. Эффективность предложенной модели сравнивается с пятью современными моделями (AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet) и дескрипторами локальных двоичных шаблонов, которые применялись к тому же набору данных. В результате предложенная модель оказалась более эффективной.

Ключевые слова: спектрограмма, Паркинсон ауруы, дауысты талдау, CNN, k-fold cross-validation.

1. Introduction

Millions of older people worldwide are affected by Parkinson's disease (PD), which is the progressive degeneration of the nervous system that becomes more widespread with age [1]. PD is believed that approximately 1% of individuals over 60 years of age are affected by this condition [2]. PD exhibits two categories of symptoms: motor symptoms and non-motor symptoms. Motor symptoms affect movement and are more conspicuous. These can include slow movements. Non-motor symptoms are less noticeable and can contain problems with sleep, speaking, swallowing, and chewing. PD can also affect speech by changing the rhythm, pronunciation and voice [3]. Dopamine is a critical neurotransmitter that has a significant impact on assisting the movements of the body [4]. The production of dopamine-producing neurons results in a decrease in neuron function and impacts the communication mechanisms within the brain. It more commonly appears in men than in women [5]. Despite advances in medical technology, current methods for diagnosing PD can be invasive and not always reliable. For example, current diagnostic methods rely on subjective clinical assessments, which can be influenced by factors such as the skill and experience of the clinician, as well as by the patient's willingness to disclose symptoms. In addition, these methods may also require the use of expensive and complex equipment, which may not be accessible in all parts of the world. These difficulties motivate us to utilize deep learning techniques that aggregate time, effort and cost. One of the foremost critical hurdles encountered by individuals with Parkinson's disease is changes in speech or difficulty speaking. Making early detection of speech disorders in Parkinson's patients can be crucial in relieving the potential impacts of the disease. The speech signals of Parkinson's disease patients differ significantly from those of healthy individuals, highlighting the importance of a timely and accurate diagnosis.

As technology advances, there has been a substantial increase in the volume of medical data, demonstrating significant growth and creating a demand for innovative methods to extract valuable insights from this information. Data mining and machine learning techniques provide an ideal solution for uncovering new knowledge hidden within this vast dataset [6]. This study employs deep learning techniques that can accurately detect Parkinson's disease based on pronouncing the letter" a". The audio recording is transformed into a spectrogram image, which is input for a Convolutional Neural Network (CNN) model. Regrettably, multiple obstacles have contributed to the significant challenges encountered in addressing this task. The most formidable aspect is the limited availability of a dataset that remains unaffected by noise, stretching, resizing, or speed variation methods, which were previously commonplace.

The research makes the following contributions.

1. The researchers have devised a system for detecting Parkinson's disease that achieves a classification accuracy of 97.9% by utilizing recorded speech samples from individuals with Parkinson's disease and by employing a convolutional neural network (CNN) for this purpose.

2. Implemented a preprocessing technique that involved splitting the audio into 2-second segments, ensuring equal segment lengths. This process facilitated the creation of equivalent record files and separated the audio recording into harmonic distortion components.

3. The researchers applied augmentation techniques such as oversampling, pitch shifting, and adding Gaussian noise to address the challenge of limited dataset size, enhancing and increasing the dataset.

4. We assessed the performance of our model using a confusion matrix, which allowed us to calculate various metrics such as precision, accuracy, AUC, recall, and F1-Score. The evaluation results indicated that our model demonstrated the highest level of performance in terms of accuracy.

The article is referred to as structured as follows: Section 2 presents the related work, while Section 3 introduces the Parkinson's dataset, spectrogram representation, and proposed CNN architecture. The results are elaborated in Section 4, and Section 5 concludes the paper.

2. Related Work

Guatelli R. et al. [7] proposed a method for data augmentation through the creation of spectrograms from voice signals using various color palettes. A total of 13 color palettes from the Matlab colormap tool were utilized, and popular CNN models such as AlexNet, VGG 16, ResNet 50, Inception v3, and Squeezenet were employed to assess the effectiveness of the approach. The evaluation results revealed that the VGG16 network exhibited the highest performance metrics, achieving an average success rate of 95.98%.

Gelvez-Almeida E. et al. [8] aimed to distinguish between individuals with Parkinson's disease and those who are healthy using extreme learning machine neural networks. The study utilized the local binary pattern (LBP) algorithm to preprocess a database consisting of 58 Parkinson's disease patients and 77 healthy individuals. A comparison was made between the performance of single hidden layer and multilayer extreme learning machine networks. The findings indicated that the multilayer extreme hierarchical extreme learning machine networks attained identical accuracy rates of 90.12 % for the training set, 92.59% for the validation set, and 81.48% for the test set.

Wang X. et al. [9] presented a novel auxiliary diagnosis algorithm for Parkinson's disease, utilizing deep learning and hyperparameter optimization. The system incorporates ResNet50 for feature extraction and classification, comprising speech signal processing, algorithm refinement using hyperparameter optimization and Artificial Bee Colony of ResNet50. The proposed algorithm, GDABC, demonstrates significant accuracy improvement, achieving a diagnosis system accuracy of 96%.

Reddy MK. et al. [10] explored the potential of utilizing an exemplar-based sparse representation (SR) classification method to identify Parkinson's disease (PD) through speech analysis. To assess the efficiency of the suggested technique, experiments were conducted on the DDK and sentence reading tasks of the PC-GITA database, utilizing both the IS10 and combined feature sets. The results show that the Proposed-NSRC method achieved the highest accuracy of 82.50% for the DDK task.

Mamun M. et al. [11] utilized vocal features to detect Parkinson's disease (PD) and evaluated the performance of ten machine learning algorithms. The algorithms employed for PD detection were Decision Tree, Random Forest, Bagging, LightGBM, AdaBoost, XGBoost, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Naive Bayes classifiers. The dataset comprised 195 vocal records from patients obtained from the UCI Repository dataset, and the results revealed that LightGBM exhibited the highest accuracy of 95%.

Govindu A. et al. [12] focused on the utilization of machine learning methodologies within telemedicine for the early identification of Parkinson's disease. To achieve this goal, researchers investigated the MDVP audio data from 30 Parkinson's patients and healthy individuals, using four distinct machine learning models during training. The Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) models were compared to determine their classification results. Among the various machine learning techniques tested for detecting Parkinson's disease, the Random Forest classifier exhibited the highest level of effectiveness. The model attained a detection accuracy of 91.83% and a sensitivity of 0.95%.

Gaur S et al. [13] used the K-nearest neighbor method for identifying healthy and fatigued voices.. They evaluated the potential of the harmonic-to-noise ratio (HNR) as a speech biomarker for distinguishing between normal and fatigued voices. A total of 32 healthy young male volunteers aged 20 to 40 years were recorded for sustained vowel /a/ and visual reaction time following one night of sleep deprivation. The KNN classifier was employed to investigate the effectiveness of speech HNR as a biomarker for detecting healthy and fatigued voices. The HNR feature was extracted from an acoustic sample three times, and at 3 AM, the HNR demonstrated a significant change ($p=0.05$). The KNN classification's best performing k-neighbors value for visual reaction time was determined, with a validation accuracy of 56% and a test accuracy of 78%.

3. Methodology

The technique was partitioned into separate procedures in practical implementation. To accomplish the task, the project utilized the Python programming language along with the LIBROSA library. Due to our limited dataset size, we implemented augmentation techniques to expand it. We followed these steps to perform augmentation in a systematic manner as shown in Figure 1.

3.1 Dataset details

Recently, researchers have explored the possibility of using voice analysis for assessing Parkinson's patients without invasive procedures. To this end, a team of neurologists from various disciplines has come together to construct a database of voices belonging to Parkinson's patients. In 2019, recordings of voices were taken at a sound laboratory in Hospital RIVADAVIA. The laboratory was set up by Universidad Nacional de La

MATANZA UNLAM and was made ready for use by a sound technician and a speech therapist. Individuals with Parkinson's underwent a neurological evaluation (UPDRS), voice recording, and vocal cord endoscopy as part of the study. The voice database consists of recordings from 55 Parkinson patients (24 female and 31 male) and 64 non-Parkinson individuals, all of whom participated under similar conditions and followed the same protocol [14].

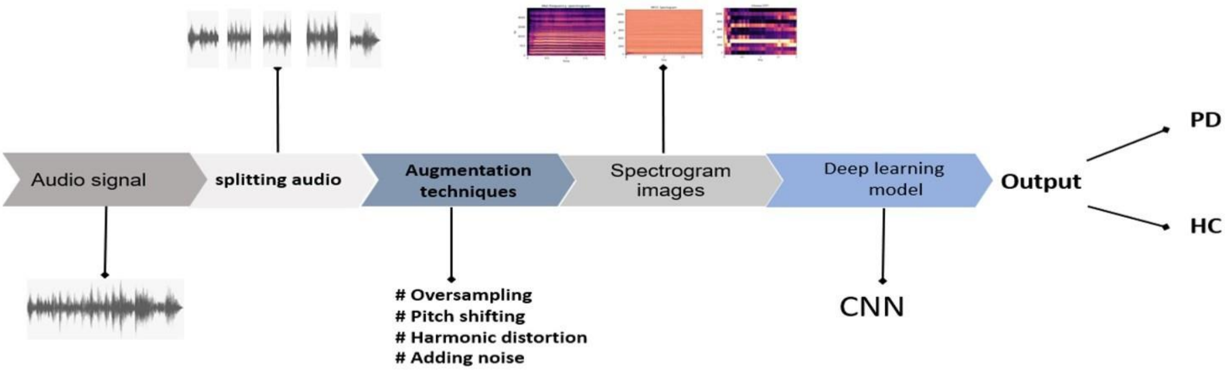


Figure 1. Workflow of building a deep learning model for Parkinson's disease detection

3.2 Data preprocessing

In deep learning, data preprocessing refers to preparing and transforming raw input data before it is fed into a neural network for training or inference. It is a critical step in deep learning because the accuracy and performance of the resulting model can be greatly influenced by it. By preparing the data correctly, deep learning algorithms can more easily identify patterns and relationships in the data, leading to better predictions and insights.

3.3 Splitting preprocessing

Audio splitting is a preprocessing step used in this study that involved dividing the long audio recording into segments of fixed duration, where each segment contained 2 seconds of audio data. To implement the audio splitting method, we used the open-source Python library LIBROSA. This allowed us to easily read the audio data and segment it into fixed-length intervals. We ensured that the segments did not overlap to avoid duplication of data. After segmenting the audio, we utilized augmentation techniques to prepare the data for training our deep learning model. This technique effectively generated the required amount of training data for our deep-learning model.

3.4 Data Augmentation Techniques

Since the number of audio files is relatively small, we used new ways to prepare data and make more of it. Data augmentation techniques are used for regularizing deep neural network inputs, which typically involves creating additional samples from the original data. There are generally two types of data augmentation methods: the first involves perturbing existing samples to create new samples, which can then be added to the fresh dataset, explicitly increasing the size of the dataset. Other methods used in this paper are oversampling, pitch Shifting, Gaussian noise, and harmonic component separation. And then trained a particular type of AI called a convolutional neural network using a large dataset, more than the original privately available, to avoid overfitting. It is worth mentioning that the initial dataset comprised 126 samples. However, with the utilization of data augmentation techniques, the dataset underwent substantial expansion, ultimately reaching a total of 1400 samples. as shown in Figure 2, Figure 3, and Figure 4 (a).

Separating audio signals into harmonic component

Separating an audio signal into its harmonic and percussive components is a common signal processing problem that can be addressed using Python and the LIBROSA and sound file libraries. By using the Librosa. Effects. Harmonic function, we can extract the harmonic components of an input audio signal and write them to a new file with a modified name using the sound file library. This technique is a powerful tool for analyzing and manipulating audio signals, as it allows for a deeper understanding of the harmonic and non-harmonic components of a sound. This can lead to new insights and applications in the field of audio processing, such as music transcription, speech analysis, and sound separation, as shown in Figure 2, Figure 3, and Figure 4 (b).

Audio data augmentation with Oversampling

Oversampling is a prevalent approach in deep learning for augmenting training datasets to enhance their size. The resampling technique is specifically employed in this research, utilizing the librosa.resample()

function. Resampling involves adjusting the sampling rate at which samples are extracted from the original signal, generating a new signal with a different sample rate. Following the resampling process, the resulting oversampled signal is saved as a new audio file, with the filename appended with "-oversampling" to signify the utilization of this resampling technique.

This technique is particularly useful in imbalanced datasets where the number of examples in one class is significantly lower than in the other classes. Evaluating the performance of the model on the original, unbalanced dataset is also crucial to ensure that the oversampling is indeed improving the model's performance, as shown in Figure 2, Figure 3, and Figure 4 (c).

Audio data augmentation with Pitch-Shifting

Pitch Shifting was the chosen method for sound recording due to its widespread usage and simplicity. We utilized an existing Python library for audio processing and analysis called LIBROSA for the implementation of this technique. The generated audio samples were created by increasing the pitch of the original samples by 0.5 steps, with each step representing a semitone. The number of steps was determined through a manual examination, wherein two independent listeners evaluated the majority of the augmented recordings. This evaluation aimed to ensure that the pitch-shifting process did not affect the vocal features. This technique is utilized to create variations of the original audio signal for training deep learning models. It can aid in increasing the size of the training dataset, improving the model's ability to generalize, and simulating different environments [15], as shown in Figure 2, Figure 3, and Figure 4 (d).

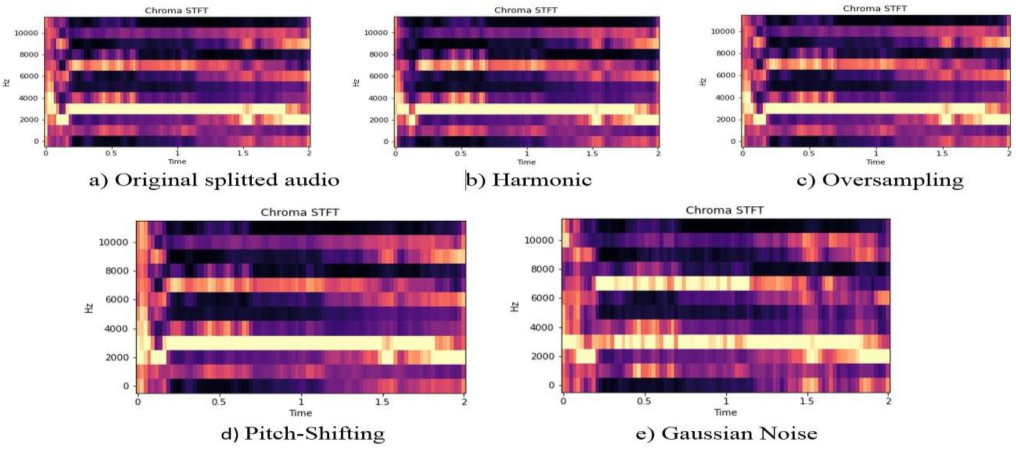
Audio data augmentation with Gaussian noise

Adding Gaussian noise can make audio smoother and easier to learn. It is possible to add noise to more than just audio, like weights and gradients. The noise's amplitude, measured by σ, cannot be so small and may not have enough effect on the system, while a value that is too large may impede the classifier's ability to learn. The acceptable range for σ is [0-0.005]. We used the mean=0 and std=0.005 The resulting sample after adding noise can be represented in Equation (1) [16], as shown in Figure 2, Figure 3, and Figure 4 (e).

$$x(t + 1) = x(t) + \sigma, \tag{1}$$

3.5 Spectrogram

A spectrogram is a visual depiction that represents a signal that shows the distribution of different frequencies over time. The frequency and time dimensions are represented on the vertical and horizontal axes of the spectrogram, respectively. This type of visualization provides more detailed information about the time-frequency characteristics of a signal than other methods. Three types of spectrogram representations were utilized, each trained for 150 epochs. These types are as follows.



a) Original splitted audio

b) Harmonic

c) Oversampling

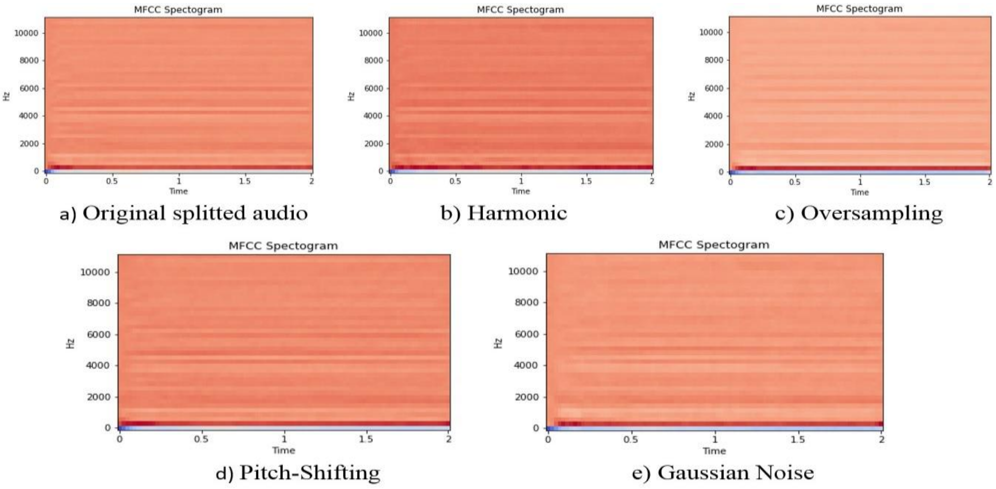
d) Pitch-Shifting

e) Gaussian Noise

Figure 2. Chroma STFT Spectrogram of the augmented sounds

Chroma STFT (Short-Time Fourier Transform)

The Chroma STFT is a signal processing technique that utilizes the audio waveform to analyze the musical content of an audio sample. By applying the Short-Time Fourier Transform (STFT) to the audio waveform and converting the resulting frequency content into 12" chroma" bins, these bins can be exhibited as a spectrum in a" chroma-gram" where the audio notes are shown on the vertical axis and time on the horizontal axis. The model contains a total of (630,528) parameters. The evaluation metrics indicate an average training accuracy of 96.3 % and average test accuracy of 92.9 %. Figure 2 displays a visual representation of the Short-Time Fourier Transform (STFT).



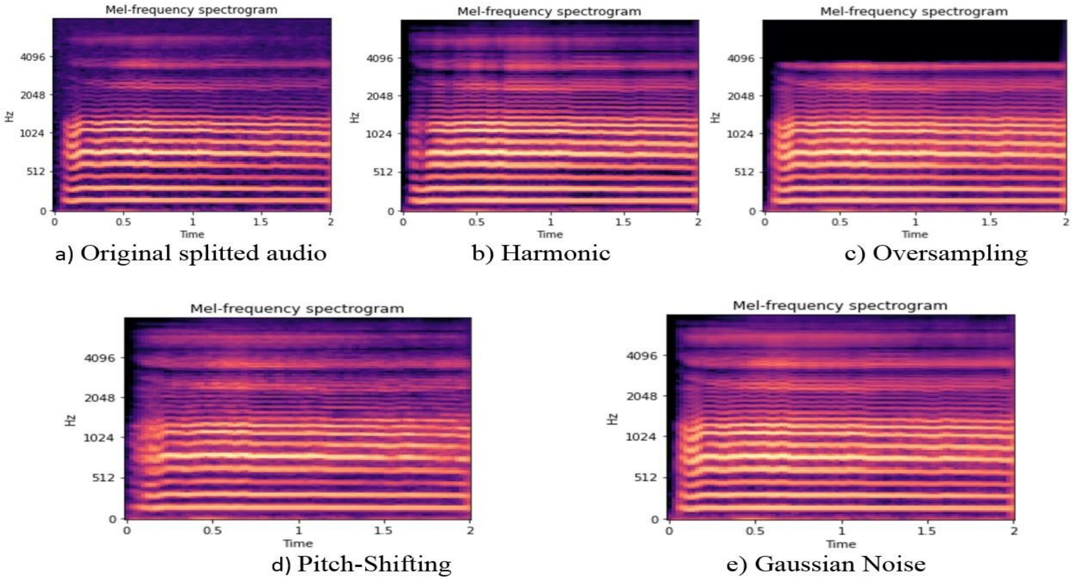
a) Original splitted audio b) Harmonic c) Oversampling

d) Pitch-Shifting e) Gaussian Noise

Figure 3. Mel-frequency cepstral coefficients (MFCCs) of the augmented sounds.

Mel- Spectrogram

It is a method for computing a spectrogram with a frequency axis that is scaled according to the Mel scale. It is a common technique for representing the time-varying frequency content of an audio signal and typically involves dividing the signal into 128 frequency bins [18]. The model contains a total of (890.624) parameters. The evaluation metrics indicate an average training accuracy of 99.3 % and average test accuracy of 97.9 %, as shown in Figure 4.



a) Original splitted audio b) Harmonic c) Oversampling

d) Pitch-Shifting e) Gaussian Noise

Figure 4. Mel Spectrogram of the augmented sounds

3.6 Network architecture

In this research, a Convolutional Neural Network (CNN) is employed due to its notable achievements in various classification assignments. The CNN draws inspiration from the interconnectivity patterns observed in neurons and comprises multiple layers, including the input layer, convolutional layers, pooling layer, one or more fully connected layers, and, ultimately, the output layer [19]. The convolution operation is a fundamental component of CNNs, serving as the primary building block. Its main function is extracting features, which requires significantly less preprocessing than traditional methods. The major advantage of CNNs is that feature extraction does not require manual extraction of matrices or formula design. Our CNN utilized 2D convolution layers with varying filter sizes, and the initial layer performed convolution on a spectrogram containing 128 features. Following the convolution and pooling layers, the fully connected layers are added. In order to transform sounds into spectrograms, various CNN architectures with various filters and layers were designed to enhance to avoid overfitting. The dropout technique was implemented with 0.4 units. The optimizer algorithm (Adam) was utilized to minimize the losses and to obtain the most precise outcomes in the current study. A loss function was employed to enable accurate predictions, precisely categorical cross-entropy, which is optimal for multi-class classification labels. To simplify the models, measures were taken to decrease their complexity. To minimize overfitting and improve efficiency, we utilized k -fold cross-validation. This approach involves partitioning the entire dataset into k equally sized subsets. Each of the k models trained uses a different subset as the hold-out validation set instead of dividing the data into separate training and testing sets by creating k subsamples that can serve as a validation dataset for testing a model [20]. We implement early stopping learning to prevent a model from overfitting on the training data. It involves stopping the training process of the model when its performance on a validation set starts to deteriorate. This helps the model to generalize better on unseen data and avoid overfitting, which can lead to poor performance on new data. This architecture consists of three 2D convolutional layers that use a (3x3) filter and have channel sizes of 64, 128, and 254, respectively. The subsequent layer is a max pooling layer with a (2x2) kernel size, followed by two Dense layers with sizes of 1024 and 2, respectively. The activation function used throughout the network is Hyperbolic Tangent (Tanh) except for the Dense layers, in which we used the ReLU activation function. We used L2 regularization (0.01). The final layer uses Softmax activation. In order to mitigate the issue of overfitting, dropout layers with a rate of 0.4 were implemented. The Adam optimizer was used with the Mean Square Error (MSE) as the loss function shown in Figure 5.

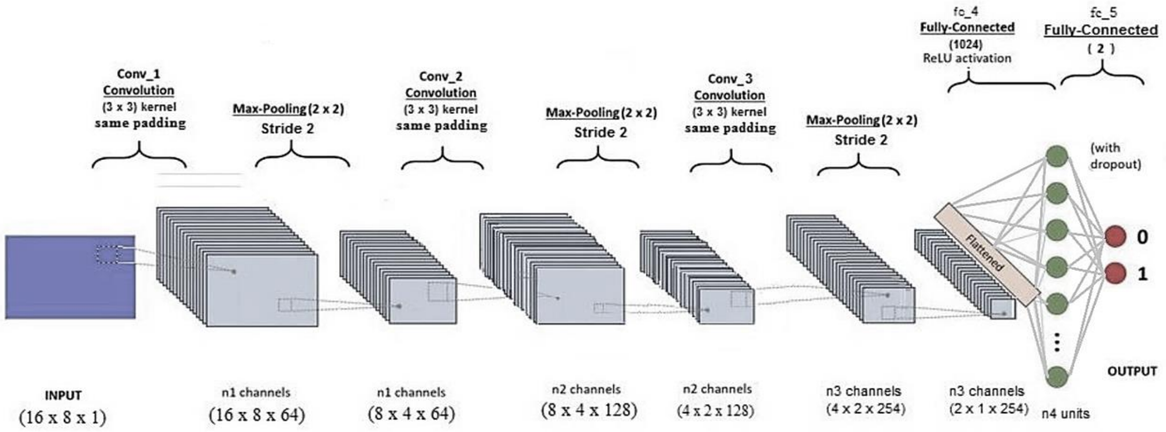


Figure 5. The proposed CNN Model

The results present the outcomes where Mel-spectrogram was utilized for sound transformation to the spectrogram, yielding a positive outcome. Notably, the third approach demonstrated the highest performance and was deemed the most effective.

4. Results

After being trained, the Chroma-STFT, Mel-Spectrogram and Mel-frequency cepstral coefficients (MFCCs), demonstrated their respective performances with 150 Epochs, which measures how many times the entire dataset was used to train the model. The accuracy and loss summaries of the models are typically presented in Table 1. In the context of the table, the accuracy summary shows the percentage of correctly classified data points in the dataset, while the loss summary shows how well the model is performing in terms of minimizing errors. We experimented with various techniques for converting voice recordings related to Parkinson's disease into spectrograms, such as Mel-frequency cepstral coefficients (MFCCs), Chroma STFT, and Mel spectrogram. After training the models for 150 epochs, we obtained different results in terms of accuracy, precision, recall, and F1-score. A summary of our findings is presented in Table 2. Once trained, Mel-frequency cepstral coefficients (MFCCs), Chroma-STFT and Mel-Spectrogram exhibited their unique performances over 150 epochs, representing the number of times the complete dataset was employed to train the models.

Table 1. Comparison of test accuracy result

METHODS	*Min. Accuracy %*	*Max. Accuracy %*	*Average Accuracy %*
Chroma STFT	*88.6*	*94.3*	*91.9*
MFCC	*83.6*	*95.0*	*90.3*
Mel Spectrogram	*95.7*	*99.3*	*97.9*
AlexNet [7]	*81.20*	*91.74*	*87.64*
VGG16 [7]	*92.88*	*98.01*	*95.98*
Inception V3 [7]	*78.92*	*86.89*	*83.13*
ResNet50 [7]	*86.89*	*90.03*	*88.09*
SqueezeNet [7]	*73.79*	*84.05*	*80.09*
H-ELM [8]	*NA*	*NA*	*81.48*
ML-ELM with 2 layers [8]	*NA*	*NA*	*81.48*
ML-ELM with 3 layers [8]	*NA*	*NA*	*81.48*
ELM [8]	*NA*	*NA*	*77.78*

The information displayed in Table 1 shows the outcomes where Mel-spectrogram was utilized for sound transformation to the spectrogram, yielding a positive outcome. Notably, the third approach demonstrated the highest performance and was deemed the most effective. The parameters we mentioned in Table 2 are commonly used in the context of training neural networks for deep learning tasks.

Table 2. Results are summarized, including loss and accuracy metrics

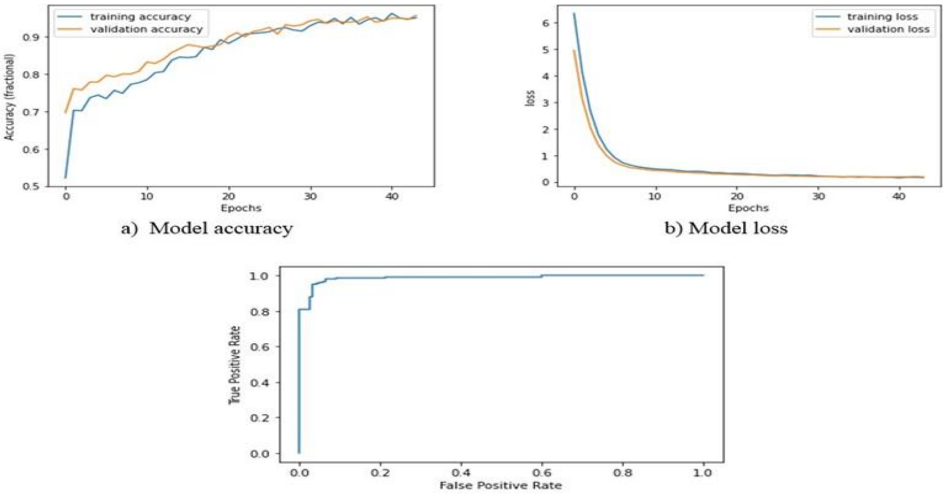
Representation of spectrum	*Train Accuracy*	*Test Accuracy*	*Precision*	*Recall*	*F1- score*
Chroma-STFT	*96.3*	*92.9*	*90*	*89.5*	*89.5*
MFCC	*94.6*	*91.3*	*91*	*91.5*	*90.5*
Mel-Spectrogram	*99.3*	*97.9*	*96*	*96.5*	*96*

We also utilize the evaluation matrix to assess the performance of our model on both the training and testing datasets. Figure 6 showcases the strength and robustness of our model, demonstrating its remarkable accuracy surpassing the most recent study with an impressive 97.9%. The parameters we mentioned in Table 3 are commonly used in the context of training neural networks for deep learning tasks.

Table 3. Control Parameters

Parameters	*Value*
Epoch	*150*
k-fold	*10*
Batch-size	*128*
Drop out	*0.4*
Learning rate	*0.001*
Optimizer	*Adam*
Activation functions	*tanh, relu, softmax*
Regularizer	*L2 with 0.01*

By implementing multiple K-fold cross-validations, we can achieve a more comprehensive assessment of the model's performance, enabling us to make informed decisions about its effectiveness and generalization capabilities. This approach enhances the evaluation process and allows us to make well-informed judgments about the model's overall performance and ability to generalize to unseen data. In Table 4, the results of the various K-fold cross-validations are presented, providing insights into the model's performance across different data splits.



a) Model accuracy

b) Model loss

c) ROC. The AUC= 0.988

Figure 6. The evaluation of the proposed CNN Model

Table 4. The accuracy of the model with different K-FOLD

No. K	*Average Train Accuracy*	*Average Train Loss*	*Average Test Accuracy*	*Average Test Loss*
K=5	*0.988*	*0.062*	*0.971*	*0.116*
K=6	*0.983*	*0.080*	*0.969*	*0.129*
K=7	*0.992*	*0.055*	*0.976*	*0.105*
K=8	*0.993*	*0.052*	*0.978*	*0.104*
K=9	*0.992*	*0.056*	*0.972*	*0.106*
K=10	*0.986*	*0.064*	*0.979*	*0.115*

5. Conclusion

The outcomes demonstrate that the precision achieved for the validation set is greater than 97.9%, which is similar to the cutting-edge performance. It is worth mentioning that the results are further enhanced, considering that only frequency-based traits obtained from spectrograms were utilized for the classification of the voice recordings. The presented approach has the additional benefit of utilizing the sustained vowel /a/. The pronunciation of this vowel is a simple task for patients to perform, as it does not necessitate lengthy instructions or prior exercises. Moreover, it is a quick task, requiring only a few seconds. As a result, the suggested technique could assist physicians in detecting Parkinson's disease during the diagnosis process.

Our plans for future work include using a Transfer learning pre-trained model on a related dataset to extract useful features and then fine-tune it on our small dataset. This way can leverage the knowledge captured by the pre-trained model and still obtain good performance on our task. In summary, this article outlines implement augmentation techniques and the application of a CNN algorithm for detecting Parkinson's disease using audio recordings of sustained vowels /a/. The dataset used for testing the algorithm consisted of 126 patients, and the audio recordings were converted into image-based representations that only described frequency features. A pre-trained network was utilized for classification, and the algorithm achieved an accuracy of over 97.9% in distinguishing between a healthy control group and a group of individuals diagnosed with Parkinson's disease. This study provides a strong foundation for future research on deep learning architectures that can automatically or semi-automatically extract features from voice recordings to diagnose Parkinson's disease.

The research developed a Parkinson's disease detection system with 97.9% accuracy using recorded speech and a convolutional neural network (CNN). Preprocessing techniques were implemented to split audio into 2-second segments, enabling the creation of equivalent record files and separating harmonic distortion components. Augmentation techniques such as oversampling, pitch shifting, and adding Gaussian noise were applied to enhance and increase the dataset. The model's performance was assessed using a confusion matrix, yielding high precision, recall, accuracy, F1-Score and AUC metrics.

6 Future work

Our future plans involve employing a transfer learning approach, where we perform a pre-trained model on a related dataset to extract relevant features and then fine-tune the model on our smaller dataset. This will enable us to demonstrate the knowledge captured by the pre-trained model and still achieve excellent performance on our specific task.

7 Acknowledgement

We sincerely thank the Universidad Nacional de La MATANZA UNLAM for establishing the sound laboratory at Hospital RIVADAVIA and the participants with Parkinson's disease who underwent neurological evaluation (UPDRS). Their contributions were invaluable to the success of our study.

References:

1 Hemmerling D. and Wojcik-Pedziwiatr M. (2022) Prediction and Estimation of Parkinson's Disease Severity Based on Voice Signal, Journal of Voice 36(3):439.e9-439.e20.

2 Reeve A., Simcox E., Turnbull D. (2014) Ageing and Parkinson's disease: why is advancing age the biggest risk factor? Ageing Res. Rev. 14:19–30.

3 Parra-Gallego L.F., Arias-Vergara T., Vásquez-Correa J.C et al. (2018) Automatic Intelligibility Assessment of Parkinson's Disease with Diadochokinetic Exercises. Communications in Computer and Information Science, Springer Verlag: 223–230.

4 Ouhmida A., Terrada O., Raihani A. et al. (2021) Voice-Based Deep Learning Medical Diagnosis System for Parkinson's Disease Prediction, International Congress of Advanced Technology and Engineering, ICOTEN 2021, Institute of Electrical and Electronics Engineers Inc.

5 Schapira A.H.V., Chaudhuri K.R., Jenner P. (2017) Non-Motor Features of Parkinson Disease. Nat. Rev. Neurosci. 18:435–450.

6 Saeed F.et al., (2022) Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods. Computers, Materials and Continua, 71(2):5639–5657.

7 Guatelli R., Aubin V., Mora M. et al. (2021) SAIV, Simposio Argentino de Imágenes y Visión Detección de Parkinson mediante Espectrogramas en Color y Redes Neuronales Convolucionales, 21-25. (in Spanish)

8 Gelvez-Almeida E, Vásquez-Coronel A., Guatelli R. et al. (2022) Classification of Parkinson's disease patients based on spectrogram using local binary pattern descriptors. Journal of Physics: Conference Series, 2153(1).

9 Wang X. et al. (2023) A Parkinson's Auxiliary Diagnosis Algorithm Based on a Hyperparameter Optimization Method of Deep Learning. IEEE/ACM Trans Comput Biol Bioinform.

10 Reddy M.K. and Alku P. (2023) Exemplar-based Sparse Representations for Detection of Parkinson's Disease from Speech. IEEE/ACM Trans Audio Speech Lang Process:1-11.

11 Mamun M., Mahmud M.I., Hossain M.I. et al. (2022) Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms. 2022 IEEE 13th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2022: 566–572.

12 Govindu A. and Palwe S. (2023) Early detection of Parkinson's disease using machine learning. Procedia Comput Sci, 218: 249–261.

13 Gaur S., Kalani P. and Mohan M. (2023) Harmonic-to-noise ratio as speech biomarker for fatigue: K-nearest neighbour machine learning algorithm. Med J Armed Forces India.

14 Giuliano M., Perez S.N., Maldonado M. et al. (2021) Construction of a Parkinson's voice database. International Conference on Industrial Engineering and Operations Management (Sao Paulo: IEOM Society International) pp:940

15 Hamdi S., Oussalah M., Moussaoui A., et al. (2022) Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound. J Intell Inf Syst, 59(2):367-389.

16 Wei S., Zou S., Liao F. et al. (2020) A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. Journal of Physics: Conference Series, 1453(1).

17 Rajesh S. and Nalini N.J. (2020) Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science, Elsevier BV, pp:16–25.

18 Alkhawaldeh R.S., (2019) DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network. Sci Program, vol.2019.

19 Syed S.A., Rashid M., Hussain S. et al. (2021) Comparative Analysis of CNN and RNN for Voice Pathology Detection. Biomed Res Int, vol.2021.

20 Wong T.T. and Yeh P.Y. (2020) Reliable Accuracy Estimates from k-Fold Cross Validation. IEEE Trans Knowl Data Eng, 32(8):1586-1594.