PARKINSON'S DISEASE DETECTION USING DEEP LEARNING BASED ON VOICE RECORDING

2023
MASTER THESIS
COMPUTER ENGINEERING

Saja Murtadha HASHIM

Thesis Advisor
Assoc. Prof. Dr. Hakan KUTUCU

PARKINSON'S DISEASE DETECTION USING DEEP LEARNING BASED ON VOICE RECORDING

Saja Murtadha HASHIM

Thesis Advisor

Assoc. Prof. Dr. Hakan KUTUCU

T.C.

Karabuk University

Institute of Graduate Programs

Department of Computer Engineering

Prepared as

Master Thesis

KARABUK

June 2023

I certify that, in my opinion, the thesis submitted by Saja Murtadha HASHIM titled "PARKINSON'S DISEASE DETECTION USING DEEP LEARNING BASED ON VOICE RECORDING " is fully adequate in scope and quality as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Hakan KUTUCU .....
Thesis Advisor, Department of Software Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. June 8, 2023

Examining Committee Members (Institutions)	Signature
Chairman : Prof. Dr. Oğuz FINDIK (KBU)	.....
Member : Prof. Dr. Emre ÇOMAK (MAKU)	.....
Member : Assoc. Prof. Dr. Hakan KUTUCU (KBU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Müslüm KUZU .....
Director of the Institute of Graduate Programs

"I hereby state that all the information incorporated in this thesis has been collected and presented in accordance with academic regulations and ethical principles. Moreover, I have conscientiously adhered to the demands specified by these regulations and principles, duly acknowledging all sources referenced in this work that are not original to it."

Saja Murtadha HASHIM

ABSTRACT

M. Sc. Thesis

PARKINSON'S DISEASE DETECTION USING DEEP LEARNING BASED ON VOICE RECORDING

Saja Murtadha HASHIM

Karabuk University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assoc. Prof. Dr. Hakan KUTUCU
June 2023, 48 pages

Parkinson's disease is a neurological disorder that hampers essential functions of the nervous system, causing difficulties in speech, writing, and balance. To automatically diagnose Parkinson's, machine learning techniques have been explored, such as analyzing acoustic signals, handwriting, and gaits. This study aims to detect Parkinson's by utilizing spectrogram images from voice recordings through Convolutional Neural Networks (CNN).

Using a private dataset from the Argentina database, consisting of recordings from 55 Parkinson's patients (24 female and 31 male) and 71 non-Parkinson individuals, this research made significant contributions. Various audio preprocessing operations were performed, including splitting the audio into 2-second segments, oversampling, adding Gaussian noise, pitch shifting, and separating harmonic components. These techniques

augmented the dataset to 1400 audio samples. The audio samples were then converted into spectrogram images for training the model.

The model underwent 150 epochs of training, resulting in an Average Training Accuracy of 99.3% and an Average Testing Accuracy of 97.9% using k-fold (k=10) cross-validation. In comparison to five state-of-the-art models (VGG16, ResNet50, Inception V3, SqueezeNet, AlexNet), as well as local binary pattern descriptors, on the same dataset, the proposed model showcased its superiority through the obtained results.

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

Science Code **:** 92432

ÖZET

Yüksek Lisans Tezi

DERİN ÖĞRENME İLE SES KAYDINA DAYALI PARKİNSON HASTALIĞI TESPİTİ

Saja Murtadha HASHIM

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Doç. Dr. Hakan KUTUCU
Haziran 2023, 48 sayfa

Parkinson hastalığı, sinir sisteminin temel işlevlerini engelleyen, konuşma, yazma ve dengede zorluklara neden olan nörolojik bir hastalıktır. Parkinson hastalığını otomatik olarak teşhis etmek için akustik sinyalleri, el yazısını ve yürüyüşleri analiz etmek gibi makine öğrenimi teknikleri araştırıldı. Bu çalışma, Konvolüsyonel Sinir Ağları (CNN) aracılığıyla ses kayıtlarından alınan spektrogram görüntülerinden yararlanarak Parkinson hastalığını tespit etmeyi amaçlamaktadır.

55 Parkinson hastası (24 kadın ve 31 erkek) ve 71 Parkinson olmayan bireyin kayıtlarından oluşan Arjantin veri tabanından özel bir veri seti kullanan bu araştırma önemli katkılar sağlamıştır. Sesi 2 saniyelik bölümlere ayırma, yüksek hızda örnekleme, Gauss gürültüsü ekleme, perde kaydırma ve harmonik bileşenleri ayırma dahil olmak üzere çeşitli ses ön işleme işlemleri gerçekleştirildi. Bu teknikler veri

setini 1400 ses örneğine genişletti. Ses örnekleri daha sonra modeli eğitmek için spektrogram görüntülerine dönüştürüldü.

Model, k-katlı (k=10) çapraz doğrulama kullanılarak %99,3 Ortalama Eğitim Doğruluğu ve %97,9 Ortalama Test Doğruluğu elde ederek 150 eğitim dönemi geçirdi. Önerilen model, beş son teknoloji model (AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet) ve aynı veri kümesi üzerindeki yerel ikili örüntü tanımlayıcıları ile karşılaştırıldı. Sonuçlar, önerilen modelin üstünlüğünü göstermiştir.

Anahtar Kelimeler **:** Spektrogram, Parkinson hastalığı, ses analizi, CNN, k-katlı çapraz doğrulama.

Bilim Kodu **:** 92432

ACKNOWLEDGMENT

I would like to express my heartfelt appreciation and deep gratitude to Allah for giving me such an opportunity to carry on pursuing my dream of becoming a master's degree student at one of the most reputable universities in Turkey. It was such a wonderful journey that will never be forgotten due to the valuable and delible feelings I had during my study for the past two years.

Allow me to convey my deepest thank to my advisor Dr Hakan Kutucu for his outstanding and valuable support and remarkable contribution to getting this thesis done.

My Mother, you were my biggest supporter. Thanks for believing in me and encouraging me to proceed further after all the chaos I had.

I dedicate this success to my mother, my country, my advisor and my friends who stood beside me and provided all the needed courage and support to complete this work.

Thanks a lot.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVIATIONS INDEX

CNN : Convolutional Neural Network

RELU : Rectified Linear Units

AI : Artificial Intelligent

Tanh : Hyperbolic Tangent

FC : Fully Connected Layer

DL : Deep learning

MFCC : Mel Frequency Cepstral Coefficients

STFT : Short-Time Fourier Transform

L1 : Regularizer 1 technique

L2 : Regularizer 2 technique

PART 1

INTRODUCTION

When a person is born, the brain has the highest count of nerve cells, which are also known as neurons [1]. Nerve cells differ from other cells in our body in that they are unable to undergo self-repair. As a consequence, as we age, these neurons progressively die and cannot be replaced [2].

The death of neurons is typically the cause of PD (Parkinson's disease); it generally occurs due to neuronal death, which leads to a reduction in the production of dopamine, a chemical substance responsible for regulating body movement. Consequently, this neurological disorder progresses gradually and affects multiple communication pathways in the brain [3].

Parkinson's disease is a progressive and enduring condition that impacts bodily movement. As the illness develops, signs show such as tremors, muscle rigidity, bradykinesia, dementia arise and orthostatic hypotension, and other signs [4].

Individuals with Parkinson's disease may experience additional complications, including issues with balance, changes in vocal and facial expressions, loss of sense of smell, and sleep disturbances [5]. A notable observation is that Parkinson's disease (PD) is often diagnosed in individuals aged 50 or above.

Diagnosing Parkinson's Disease (PD) during the initial phase and initiating treatment can potentially impede the rate of progression of this degenerative disorder. Currently, PD diagnosis relies on a patient’s medical history, signs, and symptoms since no definitive biochemical test is available.

The fundamental criteria used for diagnosis typically involve clinical neurological examinations and brain scans. However, these methods can be costly and necessitate specialized expertise [6].

Speech malfunctions are a significant indication of Parkinson's disease, that was the reason behind using voice recordings of patients who are suspected of having Parkinson's disease can serve as a useful diagnostic tool to aid in early detection. This method is affordable and can be combined with advanced deep learning techniques to improve the accuracy of predictions [7,8].

The aim of this study is to utilize a Convolutional Neural Network (CNN) for the recognition and categorization of Parkinson's disease.

As deep neural networks continue to advance, numerous novel architectural models are emerging, making it crucial to evaluate which ones offer optimal performance while minimizing time consumption. Although exceptionally deep neural network systems may be excessive for certain tasks, simpler strategies can often achieve comparable results and conserve resources. Consequently, addressing each -issue requires thorough and extensive research [9].

1.1. MOTIVATIONS

Current methods for diagnosing Parkinson's disease are invasive and unreliable, relying on subjective clinical assessments that may be influenced by various factors. These methods also require expensive equipment that is not easily accessible, leading to delays in diagnosis and treatment that affect patients' quality of life. Deep learning techniques offer a potential solution by providing accurate and non-invasive diagnostic tools that can reduce time, effort, and costs. They can identify patterns and correlations in large datasets, providing a more reliable diagnosis regardless of the clinician's experience or patient's willingness to disclose symptoms. Additionally, these techniques could potentially reduce the need for expensive equipment, making Parkinson's disease diagnosis more accessible globally. The use of deep learning

techniques in Parkinson's disease diagnosis could significantly improve patient's quality of life worldwide.

1.2. THESIS PROBLEM

Parkinson's disease is a multifaceted neurodegenerative condition that poses several challenges to its detection. The decision to utilize voice recordings as a means of detecting Parkinson's disease was grounded on the recognition that changes in vocal tone are amongst the earliest manifestations of the disease. Given the challenging nature of this endeavor, our successful ability to obtain promising results speaks to the efficiency of our approach.

One of these challenges is the limited availability of datasets, which can impede the development and evaluation of accurate diagnostic tools. In response to this challenge, the study aims to explore and provide answers to a set of questions that pertain to the acquisition of data for Parkinson's disease discovering and the application of Convolutional Neural Networks (CNNs). These questions could include issues related to the collection, curation, and augmentation of datasets, as well as the application of CNNs for feature extraction and classification.

By addressing these questions, the study seeks to enhance our understanding of the challenges associated with detecting Parkinson's disease and provide insights into potential solutions. Ultimately,s to assist in the development of improved, precise, and dependable diagnostic tools for detecting Parkinson’s disease, which could improve patients' quality of life and support early interventions.

Can the utilization of spectrogram analysis and convolutional neural networks (CNNs) be deemed a suitable approach for detecting Parkinson's disease in audio samples?

Can the classifier's performance be affected by utilizing audio features in the form of an image spectrogram?

What is the most suitable library to enhance Python's performance?

How can we create a pipeline for converting audio into spectrogram?

1.3. DATA PREPARATION

In this study, we conducted data preparation to transform and organize raw audio data into a suitable format for analysis by deep learning algorithms. Initially, we performed a splitting process to divide the audio file into multiple segments of equal duration, namely 2 seconds, thereby increasing the size of our dataset. Additionally, we employed various data augmentation techniques such as oversampling, Gaussian filtering, pitch shifting, and harmonic distortion to further augment the dataset since it was relatively small.

Our results show that effective data preparation is crucial for deep learning models to learn patterns and relationships accurately and generate meaningful outcomes.

1.4. AIMS OF THE STUDY

The following accomplishments are part of this research:

- Giving an overview of prior literature.
- Achieving high accuracy and good performance in the detection of Parkinson's disease through voice recording.
- One way to expand the dataset is by utilizing diverse augmentation techniques.
- Demonstrating the ability of a Convolutional Neural Network (CNN) for classifying sounds in addition to images.
- Establishing the strong capability of spectrograms in facilitating the utilization of CNN models for audio classification.

1.5. STRUCTURE OF THESIS

Section 1. Introduction for thesis

This section comprises contextual information, inspiration, and the identification of the problem. Additionally, it outlines the goals and objectives of the study.

Section 2. Related Work

We conducted a thorough review of the latest research, examined how our study fits into the broader academic landscape, and emphasized its originality, importance, and impact in this field.

Section 3. Methodology

In this part, we introduce an overview of the sequential stages involved in the classification of Parkinson's disease. We begin by preparing the dataset and then provide a detailed explanation of the entire process by employing a spectrogram representation and integrating a CNN. Throughout the implementation phase, furthermore, we conduct various studies.

Section 4. Results

Our study encompasses a comprehensive assessment of the depiction of the scope and the approaches employed in the CNN. We present and compare the results obtained from our approach with those of other techniques used in different classifications.

Section 5. Discussion

We conduct a comparative analysis between our research and existing studies. By highlighting the differences between our study and prior research, we provide an in-depth analysis of our findings.

Part 6. Conclusions

In this section, we present a brief overview of the findings obtained from our study and draw conclusions. We emphasize the contributions made by our research and provide a concise overview of potential prospects for future endeavors.

PART 2

LITERATURE REVIEW

The major goal of this thesis is to detect Parkinson's disease by leveraging its unique auditory signals using a Convolutional Neural Network. CNNs have demonstrated outstanding achievement across various domains of natural language processing. In this portion, we will highlight the most recent studies and initiatives that are relevant to our study.

Xingbo Wang et al. [10] present 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 of speech signal processing, algorithm refinement using Artificial Bee Colony (ABC), and hyperparameter optimization of ResNet50. The proposed algorithm, GDABC, demonstrates significant accuracy improvement, achieving a diagnosis system accuracy of 96%.

Mittapalle Kiran Reddy et al. [11] investigated the potential of employing an exemplar-based sparse representation (SR) classification method for Parkinson's disease (PD) identification through speech analysis. To assess the effectiveness of their proposed approach, a series of experiments were conducted on the DDK and sentence reading tasks of the PC-GITA database, utilizing both the IS10 and combined feature sets. The findings suggest that the Proposed-NSRC method attained the utmost precision of 82.50% for the DDK task.

In a study conducted by Savita Gaur et al. [12], the potential of the harmonic-to-noise ratio (HNR) as a speech biomarker for distinguishing between normal and fatigued voices was examined using the K-nearest neighbor (KNN) machine learning method. The study involved 32 healthy young male volunteers aged 20 to 40 years who were

recorded for sustained vowel /a/ and visual reaction time after one night of sleep deprivation. The effectiveness of speech HNR as a biomarker for detecting healthy and fatigued voices was assessed using the KNN classifier. The HNR feature was extracted from an acoustic sample on three occasions, and a significant change in HNR ($p<0.05$) was observed at 3 AM. Through KNN classification, the optimal value of k-neighbors for visual reaction time was determined, resulting in a validation accuracy of 56% and a test accuracy of 78%.

Muntasir Mamun et al. [13] employed vocal features to detect Parkinson's disease (PD) and evaluated the performance of ten, unlike machine learning methods. These algorithms utilized to ascertain PD detection were LightGBM, XGBoost, Random Forest, K-Nearest Neighbor, AdaBoost, Bagging, Support Vector Machine, Decision Tree, Logistic Regression, and Naïve Bayes classifiers. The dataset comprised 195 vocal records from patients obtained from the UCI Repository dataset. The results revealed that LightGBM exhibited the highest accuracy of 95%.

In a study conducted by Rohit Lamba et al. [14], Parkinson's disease detection through speech analysis was carried out using two datasets: the Speech dataset & Italian Parkinson's Voice and the Mobile Device Voice Recordings at King's College London dataset. Voice samples from these datasets were used to generate 17 acoustic features, and a genetic algorithm was employed to select the eight most significant features. Four classifiers, namely random forest, k-nearest neighbours, logistic regression, and XGBoost, were utilized. By combining the feature selection based on the genetic algorithm and logistic regression, the study achieved 100% accuracy on one dataset and 90% accuracy on the other.

In their research, Mehrbakhsh Nilashi et al. [15] developed a hybrid method that combined unsupervised feature selection and supervised learning techniques. The study utilized data obtained from the UCI machine learning archive. The methodology began with clustering the data using the SOM clustering technique, followed by feature selection using the Laplacian score. To predict UPDRS, GPR (Gaussian Process Regression) was implemented on the generated clusters. The performance of the proposed method was evaluated using root-mean-square error (RMSE) and correlation

coefficients. The results showed significant performance, with R-squared values of Motor-UPDRS = 0.9489 and Total-UPDRS = 0.9516, as well as RMSE values of Motor-UPDRS = 0.5144 and Total-UPDRS = 0.5105.

Aditi Govindu et al. [16] focused on the utilization of machine learning methodologies within telemedicine for the early diagnosis 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 K-Nearest Neighbors (KNN), Vector Machine (SVM), Random Forest and Logistic Regression models were compared to determine their classification results, with the Random Forest classifier proving to be the most effective machine learning technique for detecting Parkinson's disease. The Random Forest classifier model had a detection accuracy of 91.83% and a sensitivity of 0.95.

Renata Guatelli et al. [17] proposed a method for data augmentation through the creation of spectrograms from voice signals using various colour palettes. A total of 13 colour palettes from the Matlab colourmap tool were utilized, and popular CNN models such as VGG 16, AlexNet, Inception v3, ResNet 50, and Squeezenet were employed to evaluate 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%.

E Gelvez-Almeida et al. [18] solve the classification task of identifying patients with Parkinson's disease using a database of 135 spectrograms, various models such as the Standard Extreme Learning Machine, Multilayer Extreme Learning Machine, and Hierarchical Extreme Learning Machine were employed. The spectrograms were preprocessed using the Local Binary Pattern operator to extract image features, which were then used to train the network. The findings demonstrate that the performance of multilayer networks is superior, with 92.59% accuracy.

Yusra Mohammed M. Salih et al. [19], the dataset of speeches was gathered from individuals both with and without Parkinson's Disease (PD) in order to identify the disease. The audio recordings were examined, and feature vectors were derived from

them. To effectively differentiate between individuals with PD and those without, a supervised ANN Multi-Layer Perceptron with a backpropagation algorithm was utilized. Several architectures were experimented with to determine the best one for PD detection, resulting in a 93% accuracy rate.

2.1. SPECTROGRAM

A spectrogram is a graphical visual portrayal of a signal that displays the diverse frequency occurrences across varying time spans. The frequency and time dimensions are represented by the two axes of the spectrogram, respectively. Compared to other time-frequency analyses, spectrograms contain more detailed information. To detect Parkinson's disease, vocal signals can be converted into various signals, which are then converted into a spectrogram. This spectrogram can then be utilized to classify and identify individuals with Parkinson's disease [20]. A sample spectrogram can be seen in Figure 2.1.

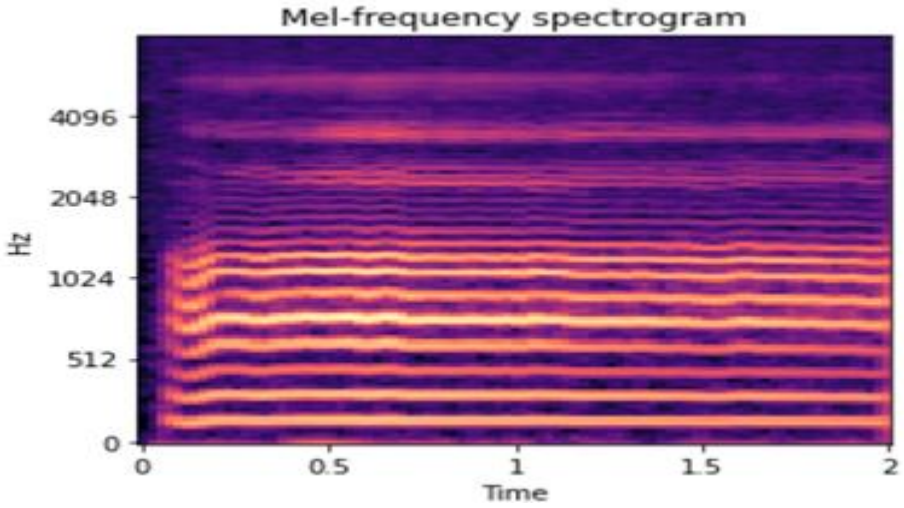


Figure 2.1. Personification of spectrogram.

2.2. MACHINE LEARNING

Recently, machine learning (ML) has become a popular tool for medical disease diagnosis due to its ease of implementation and high accuracy. A measure of dysphonia was suggested as a way to distinguish between individuals with Parkinson's disease

(PD) and healthy individuals by analyzing variations in voice frequency. Research conducted in 2013 found that sustained vowels, words, and sentences from a set of speaking tasks carried PD-specific information that could be effectively analyzed using machine learning techniques. Subsequent years saw the development of new feature extractors and data sets to improve further the analysis of voice signals in PD diagnosis [21].

There are disparities between traditional machine learning and deep learning when it comes to feature engineering. Traditional machine learning approaches often require manual feature extraction, which can be a challenging and time-consuming process. In contrast, deep learning methods leverage automatic feature extraction, enabling more precise and accurate representations of the data [22]. As illustrated in Figure 2.2.

2.3. APPROACHES UTILIZING DEEP LEARNING

Deep learning is a significant advancement in machine learning, as it enables the accuracy of the detection of objects with superior performance. However, the effectiveness of deep learning is dependent on a massive amount of data, which can be difficult to obtain from open sources. Furthermore, deep learning can also be applied to unsupervised data. It takes longer to train deep learning models, but GPU training can significantly reduce the training time [23].

The primary objective of this study is to leverage a Convolutional Neural Network (CNN) and supervised learning for dataset classification. Two main research directions are explored: firstly, the implementation of augmentation techniques to address the challenge of small dataset sizes. Secondly, utilizing the CNN model to achieve higher accuracy compared to previous studies that used the identical dataset. Our approach aims to employ the largest possible dataset for classification purposes.

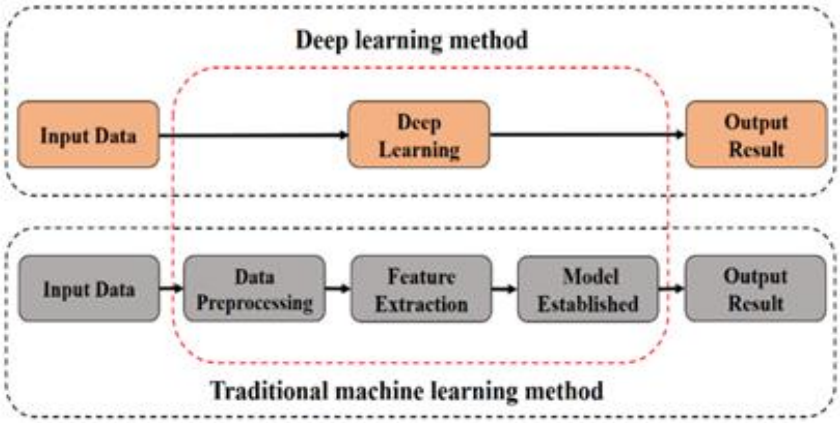


Figure 2.2. Highlighting the distinctions between traditional machine learning and deep learning methodologies [24].

2.4. CONVOLUTIONAL NEURAL NETWORK (CNN)

The popularity of CNN has increased significantly thanks to its impressive performance in various tasks such as image segmentation, object tracking, and classification. CNN's processing techniques and extensions are inspired by the structures of the human brain [25].

The structure of a Convolutional Neural Network (CNN) involves the organization of multiple layers in a specific order. The structure commences with an input layer, preceded by a series of convolutional layers and a pooling layer. Following that, there exist one or multiple fully connected layers, ultimately leading to the output layer [26]. The basic architecture of the Convolution Neural Network (CNN) is shown in Figure 2.3.

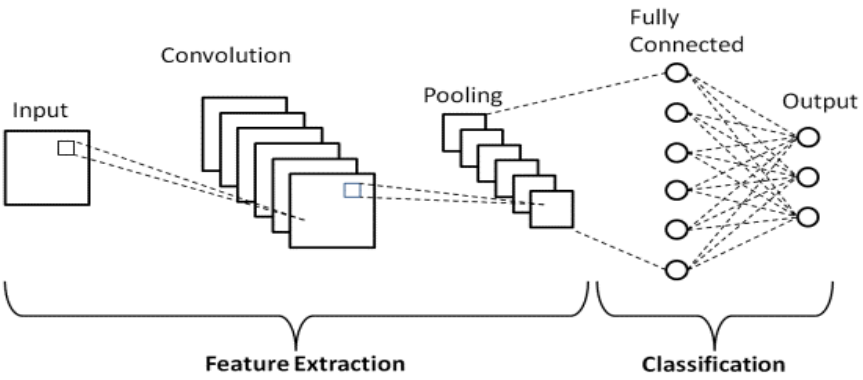


Figure 2.3. The basic CNN Architecture.

2.4.1. Convolution Layer

A convolutional layer is a core element of convolutional neural networks (CNNs) used for image recognition. It applies a mathematical equation called convolution to the input data using a small, fixed-size filter or kernel, and this process creates a new feature map that enhances and emphasizes specific patterns or features within the input data. The output of a convolutional layer is passed through an activation function and then pooled to reduce its size further and extract the most important information. Convolutional layers are important because they can learn to recognize and extract important features from images without the need for hand-engineered feature extraction [27] as the procedure of the convolution network is shown in Figure 2.4.

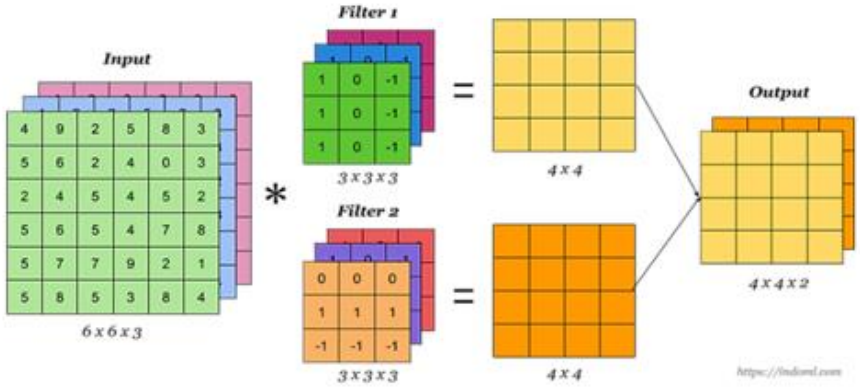


Figure 2.4. The functioning of a convolutional neural network [20].

The couple subsequent layers in a convolutional neural network called the pooling layer and fully connected layer, are critical components of the CNN architecture.

2.4.2. Pooling Layer

One of the layers in a convolutional neural network involves passing extracted feature sets to a pooling layer, which can perform operations such as maximum or average pooling depending on the specific use case. The pooling layer serves to condense the image while retaining the essential information. As a result, the dimensions of the data are reduced, and In the max pooling layer, we keep the highest value found within each window of the input data. [28], as seen in Figure 2.5.

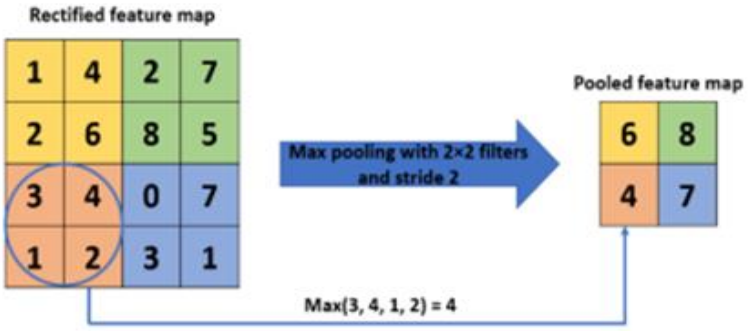


Figure 2.5. The way in which the max pooling procedure is represented [29].

2.4.3. Fully Connected Layers (fc)

The fully connected layer, also known as FC layer, is an integral part of a CNN. Its main function is to establish connections between all the neurons in the preceding layer and each neuron in the subsequent layer, facilitating the classification of features obtained from the other layers [30]. Figure 2.6 shows the fully connected layer, which comprises the input and output layers:

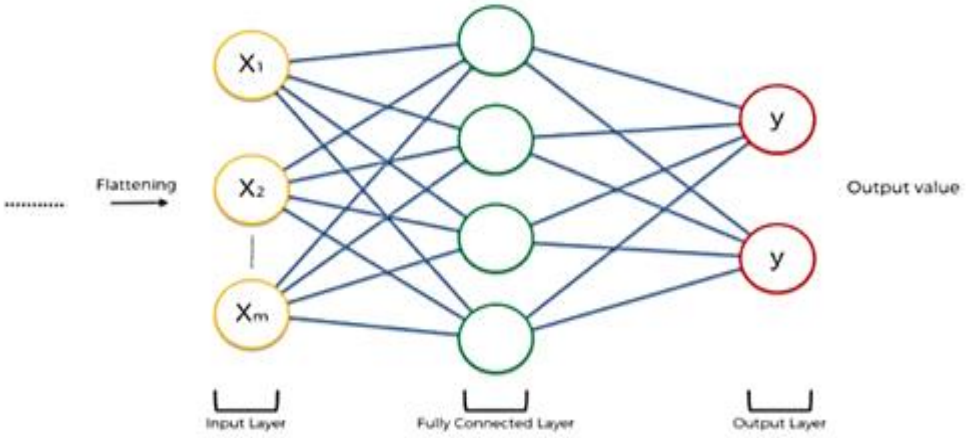


Figure 2.6. The operational mechanism of the fully connected layer [20].

2.4.4. Hyperparameters

Hyperparameters are the set of variables utilized to define the architecture of a network. These parameters facilitate optimal performance for the model. The model's deployability and speed are enhanced through the use of adjustable parameters.

2.4.5. Activation Functions

Activation functions enable non-linear transformations in deep neural networks, making it easier to learn complex tasks through the learning process [31].

2.4.5.1. Hyperbolıc Tangent

This equation is commonly utilized as a non-linear activation function when creating multiple layers in a perceptron. Its output range typically falls within the interval of (-1,1).

$$f(x) = \frac{2}{1 + e^{-2x}} - 1$$

2.4.5.2. Rectified Linear Units

Newer solutions have emerged to replace deep architectures, and one such solution is the rectified linear units (ReLUs).

$$f(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases}$$

Compared to traditional units, rectified linear units (ReLUs) offer several advantages, such as faster computation and improved gradient propagation due to their lack of saturation (a common issue with sigmoid units). Additionally, ReLUs have a range of (0, infinity) and exhibit unique properties that simplify network structures while retaining their effectiveness. Random weight initialization can cause certain units to end up in a "dead zone" with zero gradients. To mitigate this issue, this work implements Tanh and ReLU as activation functions and softmax for classification purposes.

2.4.6. Loss Function

The loss function is a useful tool for evaluating the effectiveness of a network model in handling labeled data. Two popular kinds of loss functions are cross-entropy and mean squared error (MSE), each of which calculates differently and can aid in gradient optimization [32].

2.4.7. Dropout Learning

In some methodologies of deep neural network architecture, there is a concern about overfitting due to limited sample availability. However, the utilization of dropout learning can effectively mitigate this issue. Dropout is a technique that has been widely adopted to decrease overfitting and improve overall performance [33,34].

The idea behind dropout is straightforward yet highly effective. During each training cycle, a hidden unit is randomly deactivated with a pre-defined probability (typically around 50%), and the learning process proceeds as usual. While dropout is commonly applied after the pooling layer, it can also be implemented after the convolution layer. In our study, we employed a dropout rate of 0.4 units for our specific purposes. The dropout technique is depicted in Figure 2.7.

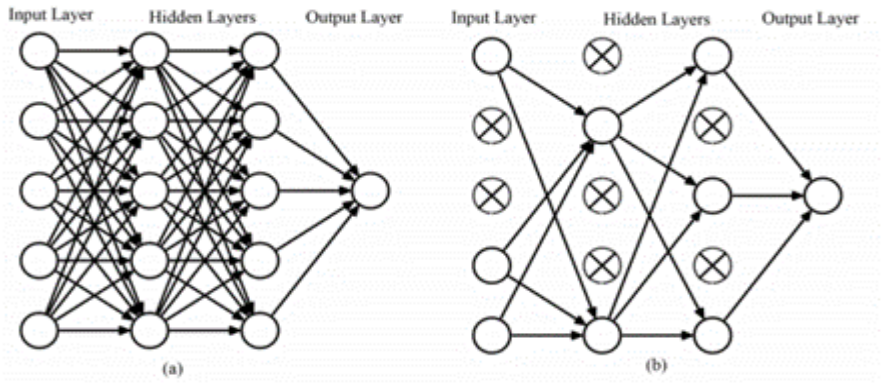


Figure 2.7. The Dropout process is shown in (a) without Dropout and (b) with Dropout [35].

2.4.8. Regularization

This is a method of decreasing the complexity of a model, which can help prevent overfitting and introduces a sanction to the loss function. The most popular types of regularization are L1 and L2. L2 is typically more effective for intricate models as it can learn complex patterns. On the other hand, L1 tends to perform better for simpler models.

2.4.9. Early Stopping

This particular method is a form of regularization that helps prevent overfitting on a validation set by monitoring the loss. Through early stopping, the process is halted when the model's loss starts increasing [36].

2.4.10. K-fold Cross Validation

In machine learning, batch size refers to the number of training examples utilized before a model update occurs. Increasing the batch size can enhance efficiency, while reducing the number of batches may be more effective with a limited number of epochs. Nonetheless, larger batch sizes often lead to greater generalization. For this particular process, the batch size has been established as 128.

2.4.11. Batch Size

In machine learning, batch size refers to the number of training examples utilized before a model update occurs. Increasing the batch size can enhance efficiency, while reducing the number of batches may be more effective with a limited number of epochs. Nonetheless, larger batch sizes often lead to greater generalization. For this particular process, the batch size has been established as 128.

2.5. MEASUREMENT AND EVALUATION

Evaluation is a crucial aspect of any experiment as it involves determining the accuracy of classification by the classifier for a given dataset. Techniques such as Confusion Precision, Recall, Matrix, Accuracy, F1-Score, among others, can be used to explain the corrected classification of objects or the accuracy of classification.

A Confusion Matrix is a tool that is employed to display and improve the accuracy of classification models. The matrix comprises columns that represent samples in predictive species and rows that represent samples in actual species. The diagonal within the matrix denotes all true representations [38].

Table 2.1. Displays the confusion matrix representation for two-class classification

		Prediction of species	
		S2	S2
Actual C	S1	True positive	False negative
	S2	False positive	True negative

The letter "S" is utilized to clarify a specific species. TP pertains to the count of True Positive values, while FP corresponds to False Positive values. Additionally, FN represents the quantity of False Negatives. TN represent the quantity of True Negative as shown in Table 2.1. The concept of Accuracy involves describing the proportion of correctly classified instances by a classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The F1-score is commonly employed to demonstrate how effectively a model classifies data using a particular algorithm, and it serves as a tool to assess the accuracy of tests [39]. To obtain the F1-score, one can compute the harmonic mean of Precision and Recall.

$$F1 = 2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)$$

Where

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

In classification, Recall represents the number of true positive predictions. On the other hand, Precision refers to the ratio of correctly predicted positive outcomes out of all positive predictions made. The F1-score, which falls between 0 and 1, will be 0 if either Precision or Recall is 0. Conversely, if both Precision and Recall are 1, then the F1-score will be 1.[40].Receiver Operating Characteristics (ROC) curves and Area Under the Curve (AUC) are important evaluation metrics in classification. ROC curves are utilized to represent probabilities, whereas AUC indicates the degree of separation between classes. Their results range between 0 and 1. A low AUC value (close to 0) indicates poor classification performance, while a high AUC value (close to 1) indicates good separation of classes by the model[20]. Figure 2.8 displays the representation of AUC-ROC.

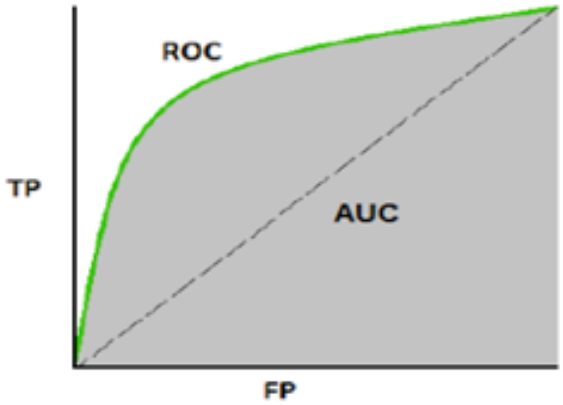


Figure 2.8. Depiction of AUC-ROC.

PART 3

METHODOLOGY

The research methodology was conducted in several stages, including dataset preprocessing, utilization of various augmentation techniques, and implementation of a CNN architecture model. This study is carried out using the Python programming language and the LIBROSA library.

3.1. DATASET DEFINITION

Lately, there has been growing interest among researchers in utilizing voice analysis as a means of assessing Parkinson's disease without the need for invasive procedures. As part of this effort, a team of neurologists from various disciplines collaborated to construct a database of voices belonging to Parkinson's patients. In 2019, voice recordings were obtained at a specialized sound laboratory located at Hospital RIVADAVIA, which had been prepared for this purpose by a sound technician and a speech therapist affiliated with Universidad Nacional de La MATANZA UNLAM. The study involved a neurological evaluation (UPDRS), voice recording, and vocal cord endoscopy for individuals with Parkinson's. The voice database comprises recordings from 55 Parkinson's patients, of which 24 are female, and 31 are male, as well as 71 non-Parkinson's individuals who underwent the same protocol under similar conditions.

A CSV file was generated to label the dataset, which included information such as file names, numbers of folds, and scientific names [41].

3.2. DATA PREPROCESSING

Data preprocessing in deep learning involves the preparation and modification of raw input data before it is utilized for training or inference in a neural network. This stage is crucial because it can heavily affect the accuracy and effectiveness of the model that is produced. With proper data preparation, deep learning algorithms can more accurately identify patterns and connections within the data, resulting in superior predictions and insights. In this context, two approaches to data preprocessing have been implemented. All steps that we implemented in our research are illustrated as a workflow, as shown in Figure 3.1.

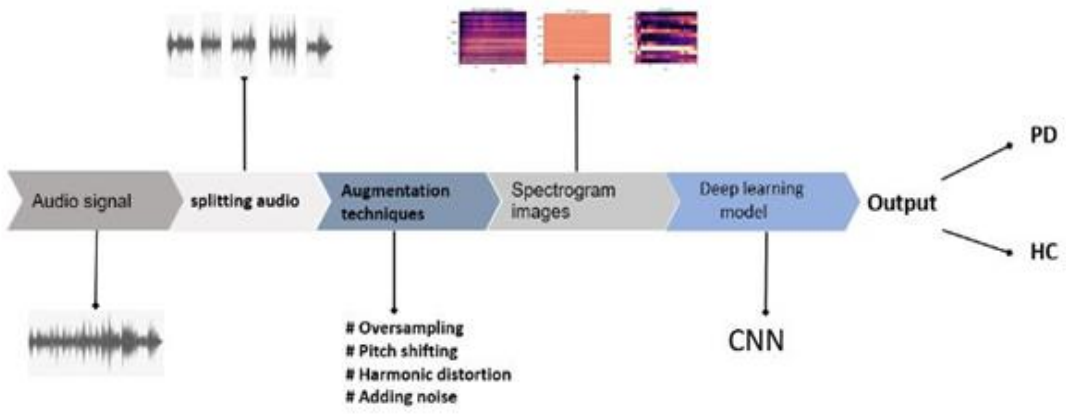


Figure 3.1. Workflow of building deep learning model for Parkinson's disease detection.

3.2.1. Splitting Preprocessing

This study employed audio splitting as a preprocessing technique, which entailed breaking down lengthy audio recordings into fixed-duration segments containing 2 seconds of audio data. The open-source Python library LIBROSA was utilized to implement this method, enabling us to read the audio data and divide it into fixed-length intervals seamlessly. To prevent the duplication of data, we made sure that the segments did not overlap. Following the audio segmentation, we used augmentation techniques to ready the data for training our deep learning model. This technique effectively produced the necessary quantity of training data for our deep learning model.

3.2.2. Separating Audio Signals Into Harmonic Components

A widely encountered issue in signal processing involves the separation of an audio signal into its percussive and harmonic components, which can be tackled with Python and the LIBROSA and sound file libraries. The use of the LIBROSA.effects.harmonic function facilitates the extraction of the harmonic parts of an input audio signal. Subsequently, we can save them into a newly-named file using the sound file library. This technique proves to be a potent instrument for scrutinizing and manipulating audio signals, affording a more comprehensive understanding of the harmonic and non-harmonic features of a sound. This, in turn, can lead to novel insights and applications in the domain of audio processing, encompassing music transcription, speech analysis, and sound separation, as displayed in Figure 3.3, Figure 3.4, and Figure 3.5 (b).

3.3. 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, typically involving 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 original data set, explicitly increasing the size of the dataset. Other methods used in this paper are oversampling, pitch Shifting, and Gaussian noise. 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, as demonstrated in Figures (3.3, 3.4, and 3.5).

3.3.1. Audio Data Augmentation With Oversampling

Oversampling is a common data augmentation technique employed in deep learning to augment the training dataset's size by replicating or generating new instances of the minority class, aiming to balance the distribution of classes. This approach is especially valuable when dealing with imbalanced datasets, where one class has

considerably fewer examples compared to other classes. Evaluating the model's performance on the original dataset is also crucial in assessing its effectiveness, the unbalanced dataset to ensure that the oversampling is indeed improving the model's performance as expressed in figures (3.3, 3.4, and 3.5) (c).

3.3.2. 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 selection of the number of steps was based on a manual evaluation conducted by two independent listeners who assessed the majority of the augmented recordings. The purpose was to ensure that the pitch-shifting process did not affect the vocal feature. 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 [17], as indicated in Figures (3.3, 3.4, and 3.5) (d).

3.3.3. 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 too 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 as x (t + 1) [42] as presented in Figures (3.3, 3.4, and 3.5) (e).

$$x(t + 1) = x(t) + \sigma$$

3.4. SPECTROGRAM CALCULATION

In order to utilize a Convolutional Neural Network (CNN), which is designed to operate on images, it is necessary to convert the audio files into an appropriate image format [43]. The properties of acoustics are capable of transformation into distinct characteristics, which can then be employed to create vectors. The extraction of relevant features is of utmost significance to the functionality of automated systems. Prior to processing, acoustics are converted into spectrograms, as shown in Figure 3.2, that provide a graphical representation of the sound frequencies as they vary over time [44]. For the purpose of our research, we applied three distinct time-frequency transformation methods - Mel-frequency cepstral coefficients (MFCCs), Chroma-STFT, and Mel-spectrogram - to convert the audios into spectrograms. The representation of these transformations, which are illustrated using sounds, is showcased in three Figures (3.3, 3.4, and 3.5).

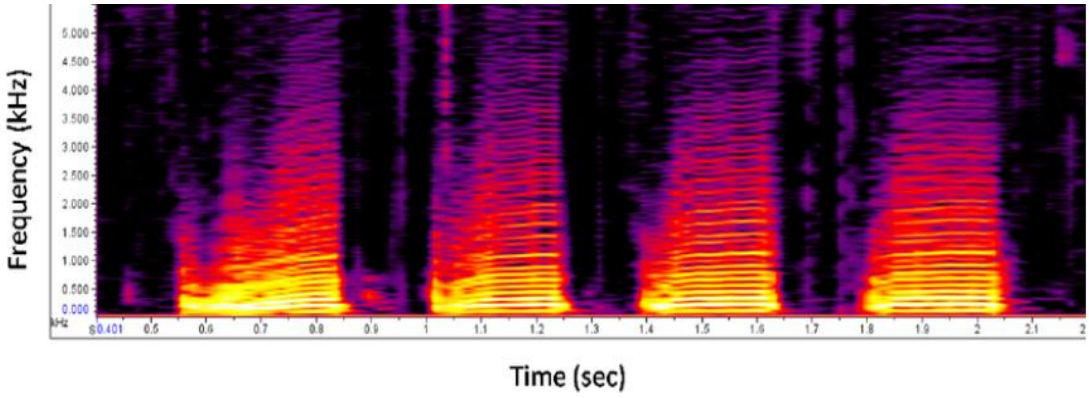
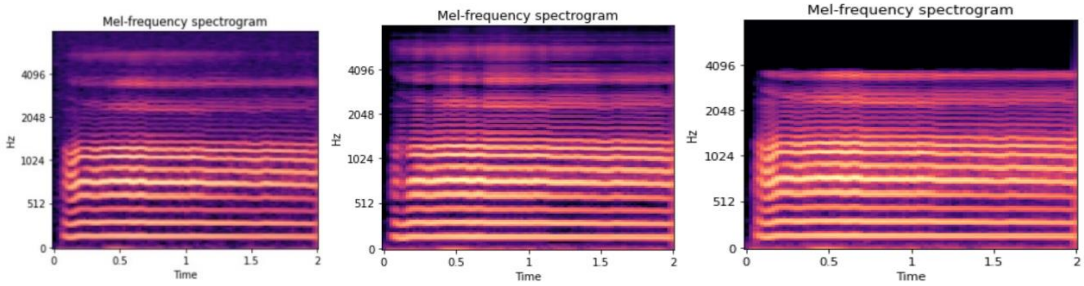


Figure 3.2. Visualization of spectrogrm

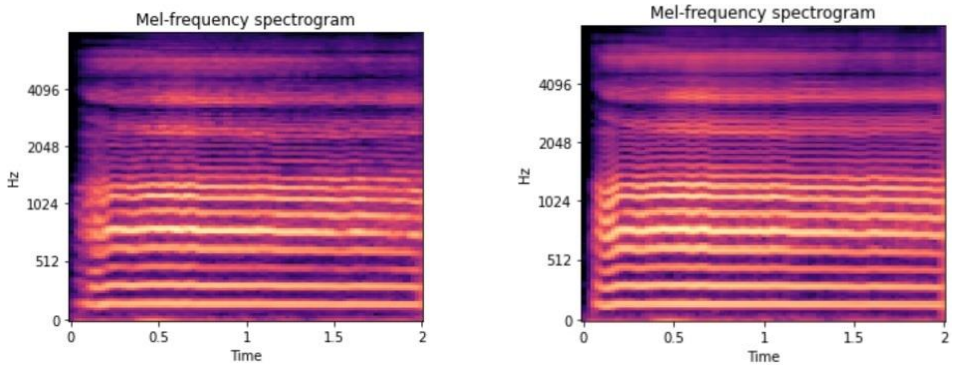
3.4.1. Chroma-Stft (Short-Time Fourier Transform)

To analyze audio samples, a Chroma-STFT is computed from the waveform, which allows for the extraction of chroma features. These features can be visualized as a spectrum in a Chroma-gram, where the tones of the audio are represented on the vertical axis. The use of Chroma-gram is particularly useful for music audio since the spectrums are displayed in 12 bins, which correspond to the 12 distinct semitones in an octave, as illustrated in Figure 3.3. This representation allows for more intuitive and

accurate analysis of the music, making it a popular tool in the field of music information retrieval [45].



(a) Original splitted audio (b) Harmonic (c) Oversampling

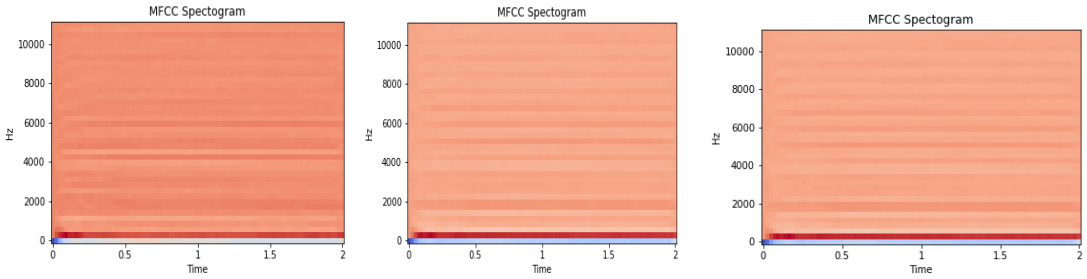


(d) Pitch (e) Gauss

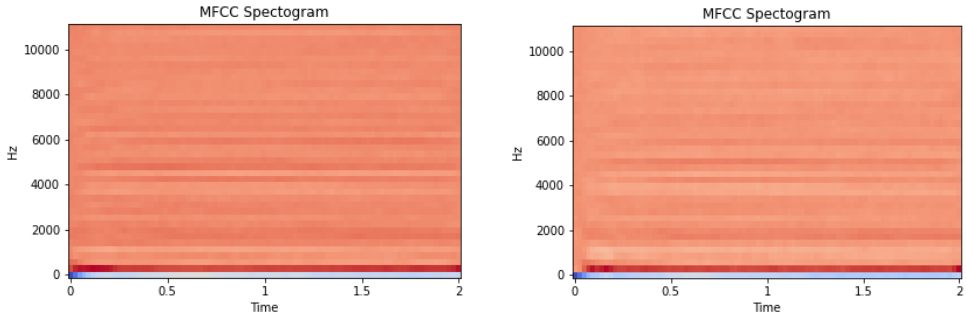
Figure 3.3. (STFT) Spectrogram of Augmented sounds.

3.4.2. Mel Frequency Cepstral Coefficients (MFCC)

The Mel Frequency Cepstral Coefficients (MFCCs) have the capability to serve as a Mel-frequency spectrogram, which is a technique frequently employed in audio-related applications. Typically, a signal consists of a small indication, usually ranging from 10 to 20 bins, and provides a comprehensive view of the spectral envelope [46]. As illustrated in Figure 3.4, the MFCC is depicted with a vector consisting of 20 dimensions. The results demonstrate a 90.7% accuracy with a loss value of 0.282.



(a) Original splitted audio (b) Harmonic (c) Oversampling

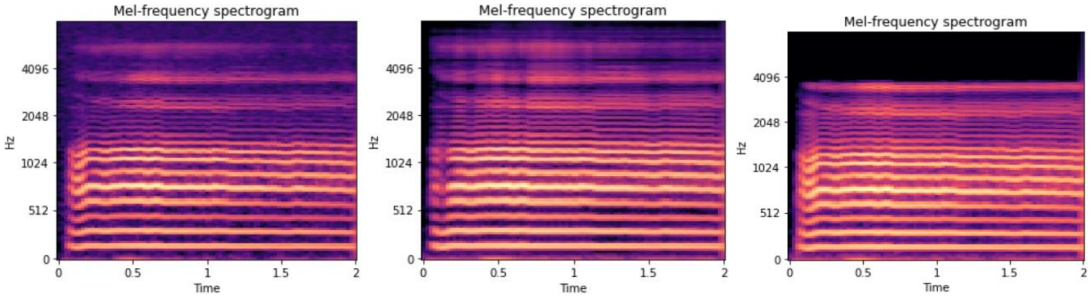


(d) Pitch (e) Gauss

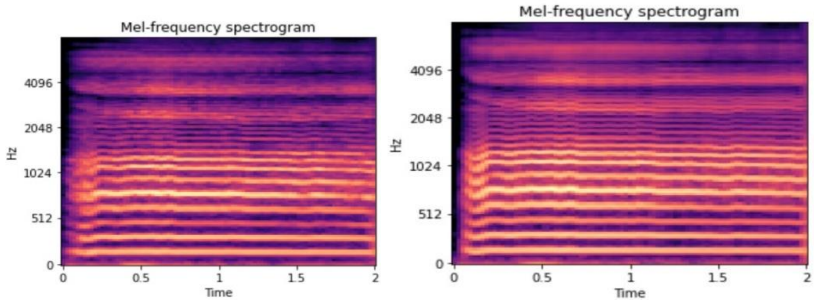
Figure 3.4. MFCC represented using a vector consisting of 20 dimensions.

3.4.3. Mel-Spectrogram

The Mel-spectrogram computes a spectrogram on a Mel scale, which allows for the representation of a signal across various frequencies and times. Typically, this representation involves partitioning the signal into 128 bins [45]. Figure 3.5 presents the Mel-spectrogram that has been represented with 128 dimensions.



(a) Original splitted audio (b) Harmonic (c) Oversampling



(d) Pitch (e) Gauss

Figure 3.5. Rendering of mel spectrogram utilizing 128 dimensions

3.5. SYSTEM MODELING

The current research employs a Convolutional Neural Network (CNN) due to its remarkable accomplishments in most classification tasks. CNNs are modelled after the interconnectivity patterns of neurons and typically include an input layer, multiple convolutional layers with a pooling layer, one or more fully connected layers, and an output layer.

3.5.1. Convolution Layer

The convolution operation is the key component of a CNN, serving as its main building block. Its primary function is to extract features. The layer comprises a set of learnable filters, also known as kernels, which we utilize in the form of 2D convolution layers with varying filter sizes.

3.5.2. Pooling Layer

Once the convolutional layer has been applied, the next step is to implement the pooling layer. This technique is used to reduce the number of parameters, as it compresses the images while still retaining the crucial information within them. Our approach involved utilizing a max pooling filter with a stride of 2, which enabled us to shrink the images down effectively.

Following the application of the convolution and pooling layers, the next step involves implementing the fully connected layers. Our model incorporates two such layers, with dimensions of (1024 and 437) respectively.

Optimizer: the optimization algorithms or strategies are responsible for minimizing losses and producing highly accurate results. In this study, we employed the Adaptive Moment Estimation (Adam) optimizer algorithm, which effectively reduces losses and ensures the attainment of highly precise outcomes. This algorithm is widely accepted within the neural network community, as it is efficient, enhances accuracy, and facilitates successful and effective training of deep neural networks.

Loss Function: the aim of utilizing this technique was to make accurate predictions about the expected outcomes. To achieve this, the method updates the weights by calculating the gradients. In our approach, we utilized categorical cross-entropy as the loss function, as it is the optimal choice for multi-class classification tasks due to the large decision boundary associated with this type of classification.

Dropout: overfitting is a common challenge in machine learning where a model becomes too closely tailored to the training data and fails to generalize well to new data. To alleviate this issue, we have implemented a regularization technique called dropout. By incorporating a dropout layer with a rate of 0.4, we randomly deactivate some of the neurons in the network during training. This approach helps prevent the neurons from becoming too reliant on specific input features and encourages them to learn more generalizable patterns. By using dropout, we aim to ensure that our model

does not overfit the training data and instead captures the underlying patterns that can be applied to new data.

Regularization: to reduce complexity and prevent overfitting in the model, we utilized a technique known as L2 (0.01) regularization. We opted for L2 over L1 regularization, as it is considered to be more effective in handling complex data tasks. The padding operation involves adding an extra pixel to an image, while the stride process is employed to regulate the convolution of filters [47].

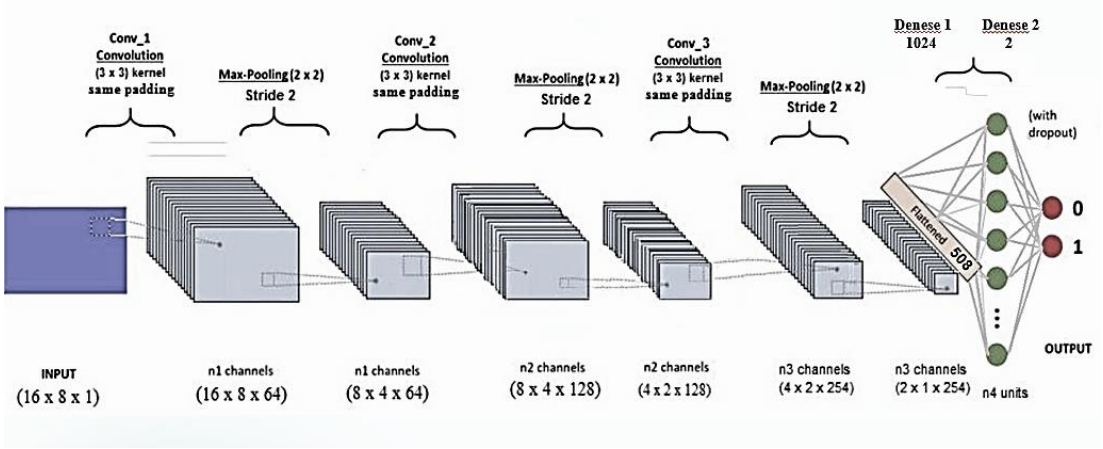


Figure 3.6. Illustrating the parameters of the CNN model employed in this research.

In this particular section, we will be thoroughly discussing the proposed approach that is being suggested for classifying Parkinson's disease. This approach is primarily based on the Deep CNN architecture, which is a complex and advanced neural network that has proven to be quite effective in dealing with intricate and challenging tasks. The proposed approach is intended to provide accurate and efficient diagnoses for individuals who are suffering from Parkinson's disease. To better understand the entire process, Figure 3.6 has been included, which will give a comprehensive overview of the entire workflow of this proposed approach.

The dataset used in this study comprises 126 voice recordings from both 71 healthy individuals and 55 patients with Parkinson's disease, which were obtained from the sound laboratory in the Hospital RIVADAVIA database [48]. In order to extract useful information from these audio signals for further analysis.

The first step involved preprocessing the voice recordings. This was done by utilizing specific techniques, which is splitting preprocessing to split the audio into 2 seconds equal duration segments, which allowed for the extraction of relevant features from each segment.

To tackle the problem of a limited dataset and to enhance the effectiveness of the proposed approach, data augmentation techniques are implemented in the second stage of the process. These procedures involve generating additional data samples from the existing dataset by applying various transformations and modifications to the original recordings. By doing so, the size of the dataset is increased, which can help to prevent overfitting and improve the accuracy of the model. The aim of data augmentation is to provide more diverse and representative samples for the model to learn from, ultimately leading to better performance and more robust results.

In the third stage of our approach, we took the preprocessed audio signals and converted them into spectrogram images. A spectrogram is a graphical display of the frequency content of a signal over time. By transforming the audio signals into spectrogram images, we are able to analyze the frequency components of the signals in greater detail. Spectrograms provide a way to visualize and analyze the changing frequency content of a signal over time. This step allows us to extract more information from the audio signals, which can be used as input for the machine learning model. The resulting spectrogram images can be thought of as a two-dimensional representation of the audio signal, where the x-axis represents time, the y-axis represents frequency, and the colour represents the intensity of the signal per point in time and frequency. This transformation from audio signals to spectrogram images is a critical step in our approach, as it allows us to analyze the audio signals in a more detailed and informative way.

In the final step of our research, we utilized a convolutional neural network (CNN) model to classify the spectrogram images of healthy individuals and patients with Parkinson's disease. CNNs are a type of neural network that are commonly used for image classification tasks, as they are able to learn and extract features from the images at various levels of abstraction.

Following that, the dataset is irregularly divided into training and testing sets, with a ratio of 80:20. Our CNN model is then trained for 150 epochs, and the model with the highest validation accuracy is selected based on its exceptional accuracy, precision, recall, and F1-score.

3.5.3. Representation of Audios

The techniques of Chroma-STFT, Mel-frequency cepstral coefficients (MFCCs), Mel-Spectrogram were utilized to transform sounds into spectrograms. The objective was to determine which technique yields the optimal results after 150 epochs.

3.5.3.1. Chroma-Stft

We employed Chroma-STFT as a method to convert audio signals into spectrograms using 12 dimensions. The model contains a total of (630,528) parameters. The outcomes indicated a mean training accuracy of 90% and a mean testing accuracy of 96.1%, as illustrated below, accompanied by a loss value of 0.250. The training and validation accuracy and loss for Chroma-STFT are presented in Figure 3.7.

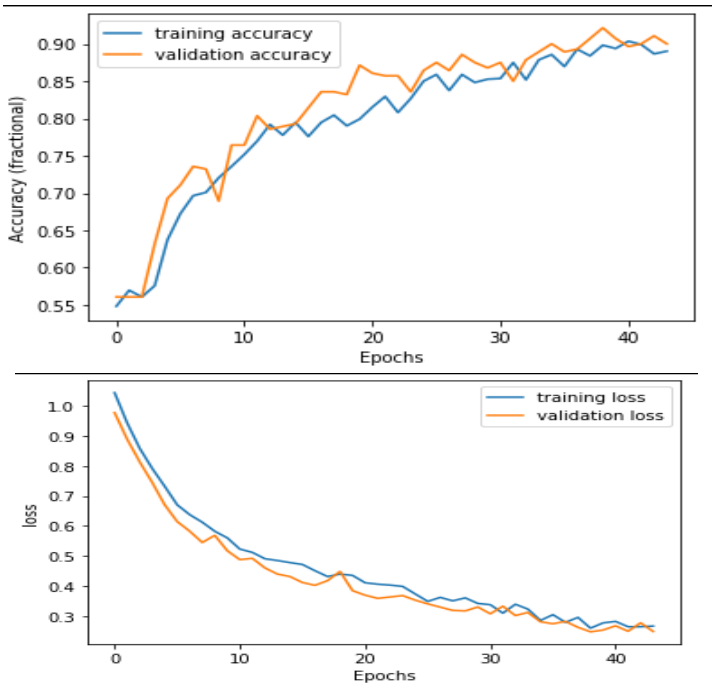


Figure 3.7. Plot illustrating the training and validation accuracy and loss using Chroma-STFT

3.5.3.2. Mel-Frequency Cepstral Coefficients

We applied Mel-frequency cepstral coefficients (MFCCs) to our dataset to generate a spectrogram with 20 dimensions, resulting in a whole parameter count of 897,026. The results demonstrate a 90.7% accuracy with a loss value of 0.282. Figure 3.8 presents the training and validation accuracy and loss achieved when employing Chroma-STFT.

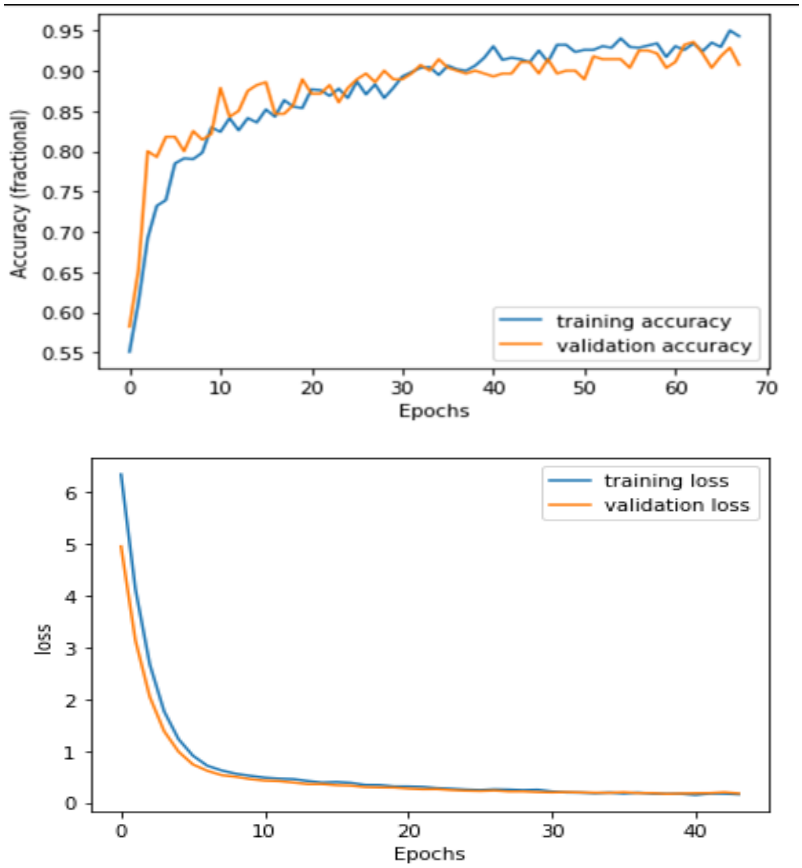


Figure 3.8. Plot illustrating the validation, training accuracy and loss using MFCC.

3.5.3.3. Mel Spectrogram

As a third technique, we employed a Mel spectrogram to transform sound into an image with 128 dimensions. The resulting spectrogram had a total parameter count of 890,624. According to the evaluation metrics, this approach yielded an average training accuracy of 99.3 and an average testing accuracy of 97.9, as depicted in Figure 3.9.

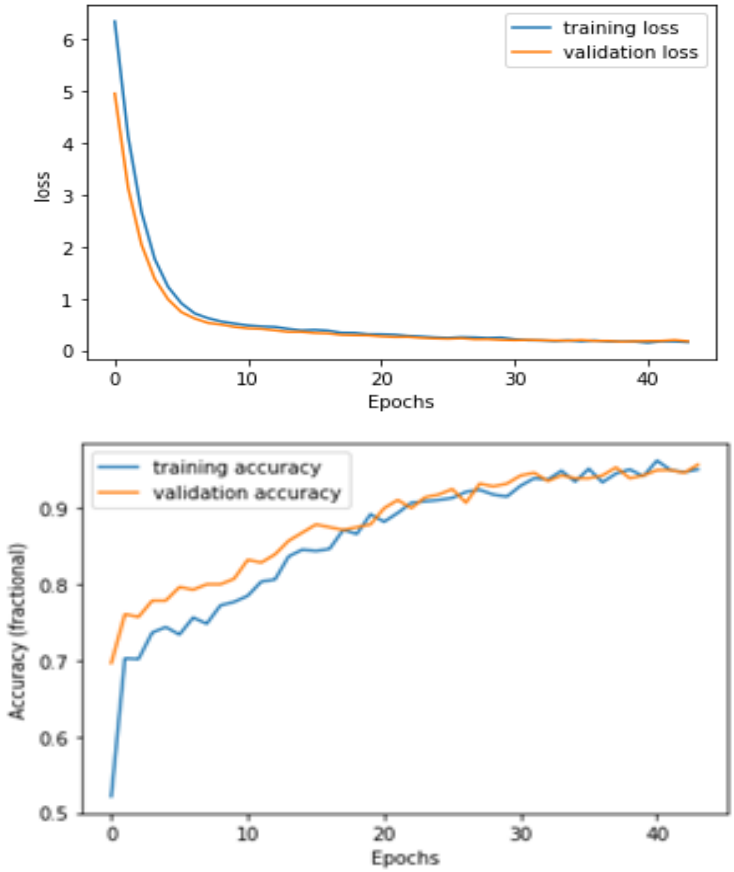


Figure 3.9. Plot illustrating the validation, training accuracy and loss using Mel spectrogram

3.6. CNN MODEL

This model employs convolutional and pooling layers, utilizing a 2D convolution operation with a 3x3 filter and channel sizes of 64, 128, and 254. The max pooling layer is set to 2x2. The network consists of two fully connected layers with sizes of 1024 and 437 sequentially. The final dense layer is used to classify into two categories. The activations used were tanh and ReLU throughout the network, while the last layer uses Softmax. A dropout rate of 0.4 is employed to prevent overfitting. The Adam optimizer and categorical cross-entropy loss function were used. The model was trained for 150 epochs and achieved accuracy and loss of 97.9%.

PART 4

RESULTS

We experimented with various techniques for converting voice recordings related to Parkinson's disease into spectrograms, such as Chroma STFT, Mel-frequency cepstral coefficients (MFCCs), 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 the Table 4.1.

Table 4.1. Results are summarized, including accuracy and loss 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
Melspectrogram	**99.3**	**97.9**	**96**	**96.5**	**96**

Once trained, Chroma-STFT, Mel-frequency cepstral coefficients (MFCCs), and Mel-Spectrogram exhibited their unique performances over 150 epochs, representing the number of times the complete dataset was employed to train the models. Typically, a summary of their accuracy scores is presented in Table 4.2. In the table, the accuracy summary reveals the percentage of accurately classified data points in the dataset.

Table 4.2. Comparison of test accuracy results.

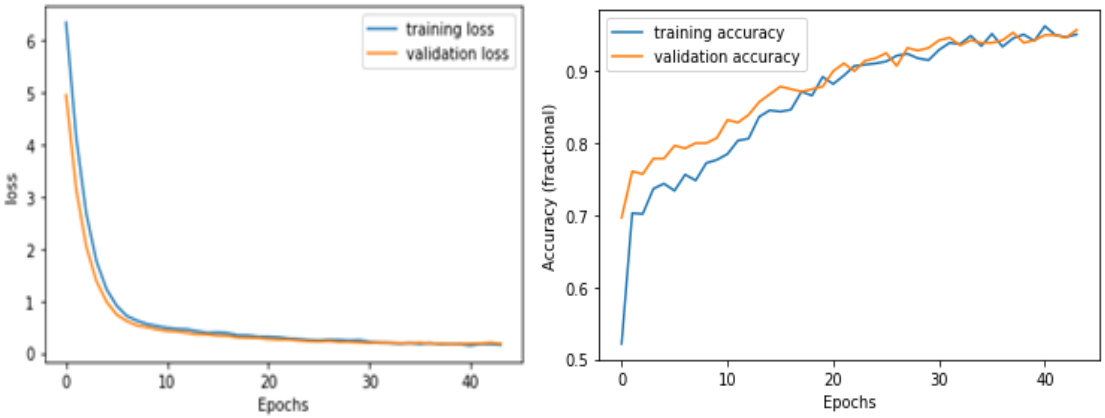
Methods	Min. Accuracy %	Max. Accuracy %	Average Accuracy %
Chroma STFT	88.6	94.3	91.9
MFCC	83.6	95.0	90.3
Mel Spectrogram	**95.7**	**99.3**	**97.9**
AlexNet [17]	81.20	91.74	87.64
VGG16 [17]	92.88	98.01	95.98
Inception V3 [17]	78.92	86.89	83.13
ResNet50 [17]	86.89	90.03	88.09
SqueezeNet [17]	73.79	84.05	80.09
H-ELM [49]	NA	NA	81.48
ML-ELM with 2 layers [49]	NA	NA	81.48
ML-ELM with 3 layers [49]	NA	NA	81.48
ELM [49]	NA	NA	77.78

The information displayed in Table 4.1 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 4.3 are commonly used in the context of training neural networks for deep learning tasks.

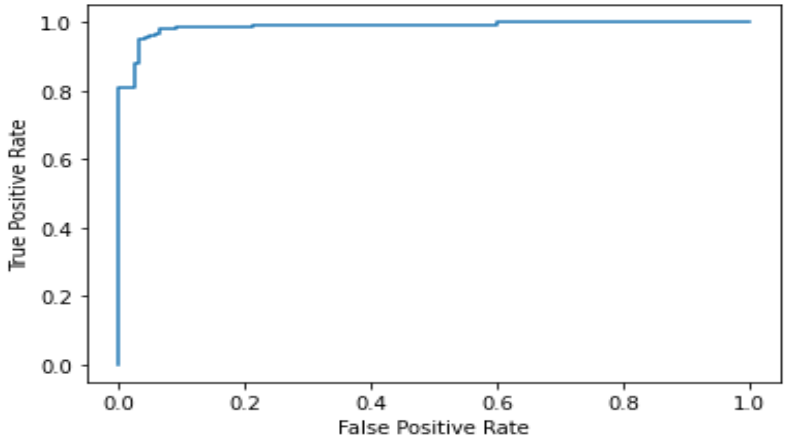
Table 4.3. Control parameter.

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

In this paragraph, we utilize the evaluation matrix to assess the performance of our model on both the training and testing datasets. Figure 4.1 showcases the strength and robustness of our model, demonstrating its remarkable accuracy surpassing the most recent study with an impressive 97.9%



a) Model loss b) Model accuracy



c) ROC Curve AUC

Figure 4.1. The evaluation of the proposed CNN model.

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 4.4, the results of the various K-fold cross-validations are presented, providing insights into the model's performance across different data splits.

Table 4.4. 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**

Furthermore, we incorporated an additional step in our process by selecting one segment from each voice recording after undergoing the splitting process. Subsequently, we applied augmentation techniques. As a result, we observed that the number of healthy control samples amounted to 385, while the number of individuals affected by Parkinson's disease was 280. Our findings indicate an average train accuracy of 0.990 and an average train loss of 0.061. Additionally, the average test accuracy is 0.961, with an average test loss of 0.176.

PART 5

DISCUSSION

Utilizing voice recordings and converting them into spectrogram images to represent signals can potentially yield more features. For instance, the process of transforming voice recordings for Parkinson's disease into spectrograms allows for the extraction of features from the sounds. However, it is crucial to carefully select the most effective feature extraction method that maximizes the information obtained from the input. By employing different techniques to convert sounds to spectrograms, varying features can be achieved.

There are a few critical points that require additional attention. First and foremost, the limited amount of available data poses a challenge. Furthermore, finding and implementing appropriate augmentation techniques can be challenging as it is essential to ensure that they do not negatively impact the original voice and alter it in any way. This was a particularly difficult task, as numerous techniques were attempted before finding the most suitable one. Despite this, further processing is still necessary, given the limited amount of data available. It is essential to extract each feature from the data, as each one possesses unique characteristics.

Convolutional Neural Networks have become a popular class of neural networks, particularly in image detection. Their success spans across a variety of processes, including classification, object tracking, and image segmentation.

Previous research indicates that in classification tasks, CNNs outperform other models such as AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet, H-ELM, and ELM. This can be attributed to CNNs' superior performance with both images and speech, as well as their enhanced power and feature extraction capabilities.

The research conducted in papers [17,49] aimed to classify audio using various algorithms, including AlexNet, VGG16, Inception V3, ResNet50, SqueezeNet, and H-ELM/ELM. The achieved accuracy is presented in Table 5.1.

Table 5.1. Comparison of test accuracy results

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 [9]	81.20	91.74	87.64
VGG16 [9]	92.88	98.01	95.98
Inception V3 [9]	78.92	86.89	83.13
ResNet50 [9]	86.89	90.03	88.09
SqueezeNet [9]	73.79	84.05	80.09
H-ELM [10]	NA	NA	81.48
ML-ELM with 2 layers [10]	NA	NA	81.48
ML-ELM with 3 layers [10]	NA	NA	81.48
ELM [10]	NA	NA	77.78

5.1. STUDY ANALYSIS

In this thesis, a dataset consisting of 1400 samples with a 2-second duration time. Wav audio samples were presented. The spectrogram was used to convert the voice recordings into images, and three types of feature extraction were employed with varying vector dimensions. Chroma-STFT, Mel-frequency cepstral coefficients (MFCCs), and Mel spectrogram were used to transform the sounds into spectrograms. After analyzing the spectrograms, it was concluded that the Mel spectrogram performed better when revealing voice recordings. Therefore, it was chosen as the best option, and its image was fed to a CNN, resulting in higher accuracy, precision, recall, and F1-score.

This research has shown that CNN is a versatile tool that can be applied to various classification tasks. Specifically, this study focuses on the development of an algorithm to classify Parkinson's disease into two distinct categories.

The CNN architecture was utilized in this study with varying parameters. In the convolution layer of the CNN, filter sizes of 64, 128, and 254 were implemented, while max pooling was performed with a 2x2 size. The widely popular Adam optimizer was chosen due to its ability to enhance accuracy and expedite the training process. For the loss function, categorical cross entropy was selected as it provides a larger decision boundary and performs better for multi-class classification tasks.

To mitigate overfitting, a range of techniques were employed, including L2 regularization and dropout. The ultimate goal was to identify the optimal approach that would enhance accuracy, precision, recall, and F1-score, while also minimizing loss and delivering superior performance.

The primary objective of the study is to attain the highest possible accuracy, which is currently set at 97.9 %.

PART 6

CONCLUSION

The results indicate that the precision attained on the validation set is over 97.9%, which is on par with state-of-the-art performance. It is remarkable that the results are even more impressive given that only frequency-based features extracted from spectrograms were used to classify the voice recordings. The proposed method offers an additional advantage of employing the sustained vowel /a/. This vowel's pronunciation is an effortless task for patients as it doesn't require extensive instructions or prior exercises. Additionally, it's a quick task that takes only a few seconds. As a result, the suggested technique may aid physicians in diagnosing Parkinson's disease during the diagnostic process [50].

This article discusses the implementation of augmentation techniques and the utilization of a CNN algorithm for detecting Parkinson's disease by analyzing voice recordings of pronouncing the vowel letter /a/. The dataset used for evaluating the algorithm comprised 126 patients, and the audio recordings were converted into image-based representations that exclusively captured frequency characteristics. A pre-trained network was employed for classification, and the algorithm achieved an accuracy of more than 97.9% for distinguishing between a healthy control group and a group of individuals diagnosed with Parkinson's disease. This study forms a solid groundwork for future research on deep learning architectures that can automatically or semi-automatically extract features from voice recordings to diagnose Parkinson's disease.

6.1. 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.

REFERENCES

[1] S. L. Oh *et al.*, "A deep learning approach for Parkinson's disease diagnosis from EEG signals," *Neural Comput Appl*, vol. 32, no. 15, pp. 10927–10933, Aug. 2020, doi: 10.1007/s00521-018-3689-5.

[2] J. Silver, M. E. Schwab, and P. G. Popovich, "Central nervous system regenerative failure: Role of oligodendrocytes, astrocytes, and microglia," *Cold Spring Harb Perspect Biol*, vol. 7, no. 3, 2015, doi: 10.1101/cshperspect.a020602.

[3] S. L. Oh *et al.*, "A deep learning approach for Parkinson's disease diagnosis from EEG signals," *Neural Comput Appl*, vol. 32, no. 15, pp. 10927–10933, Aug. 2020, doi: 10.1007/s00521-018-3689-5.

[4] Mass. IEEE MIT Undergraduate Research Technology Conference 2017 Cambridge, Institute of Electrical and Electronics Engineers, Massachusetts Institute of Technology, Mass. IEEE MIT Undergraduate Research Technology Conference 2017.11.03-05 Cambridge, and Mass. URTC 2017.11.03-05 Cambridge, *2017 IEEE MIT Undergraduate Research Technology Conference (URTC) MIT Stata Center, Building 32, 32 Vassar Street, Cambridge, MA 02139 : November 3-5, 2017*.

[5] M. Kundu, M. A. Nashiry, A. K. Dipongkor, S. S. Sumi, and M. A. Hossain, "An optimized machine learning approach for predicting parkinson's disease," *International Journal of Modern Education and Computer Science*, vol. 13, no. 4, pp. 68–74, Aug. 2021, doi: 10.5815/IJMECS.2021.04.06.

[6] A. W. Michell, S. J. G. Lewis, T. Foltynie, and R. A. Barker, "Biomarkers and Parkinson's disease," *Brain*, vol. 127, no. 8. pp. 1693–1705, Aug. 2004. doi: 10.1093/brain/awh198.

[7] Z. Bosnić and I. Kononenko, "An overview of advances in reliability estimation of individual predictions in machine learning," *Intelligent Data Analysis*, vol. 13, no. 2. pp. 385–401, 2009. doi: 10.3233/IDA-2009-0371.

[8] O. S. S. Alsharif, K. M. Elbayoudi, A. A. S. Aldrawi, and K. Akyol, "Evaluation of Different Machine Learning Methods for Caesarean Data Classification," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 5, pp. 19–23, Sep. 2019, doi: 10.5815/ijieeb.2019.05.03.

[9] J. Fatih, A. Thesis, A. Assoc, and H. Kutucu, "SPECTROGRAM IMAGES BASED IDENTIFICATION OF BIRD SPECIES USING

CONVOLUTIONAL NEURAL NETWORKS 2021 MASTER THESIS COMPUTER ENGINEERING."

[10] X. Wang *et al.*, "A Parkinson's Auxiliary Diagnosis Algorithm Based on a Hyperparameter Optimization Method of Deep Learning," *IEEE/ACM Trans Comput Biol Bioinform*, 2023, doi: 10.1109/TCBB.2023.3246961.

[11] M. K. Reddy and P. Alku, "Exemplar-based Sparse Representations for Detection of Parkinson's Disease from Speech," *IEEE/ACM Trans Audio Speech Lang Process*, pp. 1–11, 2023, doi: 10.1109/TASLP.2023.3260709.

[12] S. Gaur, P. Kalani, and M. Mohan, "Harmonic-to-noise ratio as speech biomarker for fatigue: K-nearest neighbour machine learning algorithm," *Med J Armed Forces India*, 2023, doi: 10.1016/j.mjafi.2022.12.001.

[13] M. Mamun, M. I. Mahmud, M. I. Hossain, A. M. Islam, M. S. Ahammed, and M. M. Uddin, "Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms," in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 566–572. doi: 10.1109/UEMCON54665.2022.9965732.

[14] R. Lamba, T. Gulati, A. Jain, and P. Rani, "A Speech-Based Hybrid Decision Support System for Early Detection of Parkinson's Disease," *Arab J Sci Eng*, vol. 48, no. 2, pp. 2247–2260, Feb. 2023, doi: 10.1007/s13369-022-07249-8.

[15] M. Nilashi, R. A. Abumalloh, S. Alyami, A. Alghamdi, and M. Alrizq, "Parkinson's Disease Diagnosis Using Laplacian Score, Gaussian Process Regression and Self-Organizing Maps," *Brain Sci*, vol. 13, no. 4, p. 543, Mar. 2023, doi: 10.3390/brainsci13040543.

[16] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," *Procedia Comput Sci*, vol. 218, pp. 249–261, 2023, doi: 10.1016/j.procs.2023.01.007.

[17] R. Guatelli1, V. Aubin1, M. Mora2, J. Naranjo-Torres2, and A. Sinopoli1, "SAIV, Simposio Argentino de Imágenes y Visión Detección de Parkinson mediante Espectrogramas en Color y Redes Neuronales Convolucionales".

[18] E. Gelvez-Almeida, A. Vásquez-Coronel, R. Guatelli, V. Aubin, and M. Mora, "Classification of Parkinson's disease patients based on spectrogram using local binary pattern descriptors," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2022. doi: 10.1088/1742-6596/2153/1/012014.

[19] "Parkinson's Disease Detection by Processing Different ANN Architecture Using Vocal Dataset," *Eurasian Journal of Science and Engineering*, vol. 9, no. 1, 2023, doi: 10.23918/eajse.v9i1p161.

[20] J. Fatih Awrahman and H. Kutucu, "Spectrogram Images Based Identification of Bird Species Using Convolutional Neural Networks."

[21] J. S. Almeida *et al.*, “Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques,” *Pattern Recognit Lett*, vol. 125, pp. 55–62, Jul. 2019, doi: 10.1016/j.patrec.2019.04.005.

[22] M. M. Mijwil, I. E. Salem, and M. M. Ismaeel, “The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review,” *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 1. College of Education, Al-Iraqia University, pp. 87–101, 2023. doi: 10.52866/ijcsm.2023.01.01.008.

[23] C. Yin, Y. Zhu, J. Fei, and X. He, “A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,” *IEEE Access*, vol. 5, pp. 21954–21961, Oct. 2017, doi: 10.1109/ACCESS.2017.2762418.

[24] Q. Chen, Q. Xie, Q. Yuan, H. Huang, and Y. Li, “Research on a real-time monitoring method for the wear state of a tool based on a convolutional bidirectional LSTM model,” *Symmetry (Basel)*, vol. 11, no. 10, Oct. 2019, doi: 10.3390/sym11101233.

[25] N. M. S. Thesis and Y. Ş. Günaydin, “ANKARA YILDIRIM BEYAZIT UNIVERSITY GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES SAR IMAGE DESPECKLING USING CONVOLUTIONAL NEURAL,” 2019.

[26] M. B. Er, E. Isik, and I. Isik, “Parkinson’s detection based on combined CNN and LSTM using enhanced speech signals with Variational mode decomposition,” *Biomed Signal Process Control*, vol. 70, Sep. 2021, doi: 10.1016/j.bspc.2021.103006.

[27] S. S. Iyengar, V. Saxena, IEEE Computer Society. Technical Committee on Parallel Processing, Institute of Electrical and Electronics Engineers, Jaypee Institute of Information Technology University, and University of Florida. College of Engineering, *2019 Twelfth International Conference on Contemporary Computing (IC3-2019) : 8-10 August 2019, Jaypee Institute of Information Technology, Noida, India*.

[28] S. L. Oh *et al.*, “A deep learning approach for Parkinson’s disease diagnosis from EEG signals,” *Neural Comput Appl*, vol. 32, no. 15, pp. 10927–10933, Aug. 2020, doi: 10.1007/s00521-018-3689-5.

[29] H. Gholamalinezhad and H. Khosravi, “Pooling Methods in Deep Neural Networks, a Review.”

[30] “CLASSIFICATION OF EEG SIGNALS BY”.

[31] K. J. Piczak, “Environmental sound classification with convolutional neural networks; Environmental sound classification with convolutional neural networks,” 2015, doi: 10.5281/zenodo.12714.

[32] S. Mallat, “Understanding deep convolutional networks,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*

Sciences, vol. 374, no. 2065. Royal Society of London, Apr. 13, 2016. doi: 10.1098/rsta.2015.0203.

[33] R. S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network," *Sci Program*, vol. 2019, 2019, doi: 10.1155/2019/7213717.

[34] M. Sarigül and M. Gök, "A NEW DEEP LEARNING APPROACH: DIFFERENTIAL CONVOLUTIONAL NEURAL NETWORK."

[35] Z. S. Wang, J. Lee, C. G. Song, and S. J. Kim, "Efficient chaotic imperialist competitive algorithm with dropout strategy for global optimization," *Symmetry (Basel)*, vol. 12, no. 4, pp. 1–16, Apr. 2020, doi: 10.3390/SYM12040635.

[36] M. Mahsereci, L. Balles, C. Lassner, and P. Hennig, "Early Stopping without a Validation Set," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.09580

[37] T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015, doi: 10.1016/j.patcog.2015.03.009.

[38] V. M. Patro and M. Ranjan Patra, "Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy," *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, no. 4, Aug. 2014, doi: 10.14738/tmlai.24.328.

[39] A. Fujino, H. Isozaki, and J. Suzuki, "Multi-label Text Categorization with Model Combination based on F 1-score Maximization."

[40] H. Huang, J. Wang, and H. Abudureyimu, "Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, pp. 814–817. doi: 10.21437/interspeech.2012-248.

[41] M. Giuliano *et al.*, "Construction of a Parkinson´s voice database."

[42] S. Wei, S. Zou, F. Liao, and W. Lang, "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar. 2020. doi: 10.1088/1742-6596/1453/1/012085.

[43] J. Fatih Awrahman and H. Kutucu, "Spectrogram Images Based Identification of Bird Species Using Convolutional Neural Networks."

[44] X. Jiang-jian, D. Chang-qing, L. Wen-bin, and C. Cheng-hao, "Audio-only Bird Species Automated Identification Method with Limited Training Data Based on Multi-Channel Deep Convolutional Neural Networks."

[45] R. S. Alkhawaldeh, “DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network,” *Sci Program*, vol. 2019, 2019, doi: 10.1155/2019/7213717.

[46] S. Rajesh and N. J. Nalini, “Musical instrument emotion recognition using deep recurrent neural network,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 16–25. doi: 10.1016/j.procs.2020.03.178.

[47] A. S. Almryad and H. Kutucu, “Automatic identification for field butterflies by convolutional neural networks,” *Engineering Science and Technology, an International Journal*, vol. 23, no. 1, pp. 189–195, Feb. 2020, doi: 10.1016/j.jestch.2020.01.006.

[48] M. Giuliano *et al.*, “Construction of a Parkinson´s voice database.”

[49] E. Gelvez-Almeida, A. Vásquez-Coronel, R. Guatelli, V. Aubin, and M. Mora, “Classification of Parkinson’s disease patients based on spectrogram using local binary pattern descriptors,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2022. doi: 10.1088/1742-6596/2153/1/012014.

[50] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, *Deep Learning Approach to Parkinson’s Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification; Deep Learning Approach to Parkinson’s Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification*. 2019. doi: 10.0/Linux-x86_64.

RESUME

Saja Murtadha HASHIM, acquired her foundational education in the same city, successfully completing both her first and elementary schooling. She proceeded to pursue her high school education at Al-Mithaq High School in Yemen. Subsequently, she attained a bachelor's degree in Software Engineering from Al-Mansour University College in July 2006.

In her quest for further academic advancement, Saja Murtadha HASHIM relocated to Karabuk, Turkey, in 2021 to pursue her M.Sc. degree. She embarked on her master's program at the Department of Computer Engineering at Karabuk University, Turkey.