PAPER NAME

BlackBook_final_report.pdf

WORD COUNT	CHARACTER COUNT
7047 Words	**38719 Characters**
PAGE COUNT	FILE SIZE
40 Pages	**1020.4KB**
SUBMISSION DATE	REPORT DATE
Apr 18, 2022 10:24 AM GMT+5:30	**Apr 18, 2022 10:25 AM GMT+5:30**

● **34% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 31% Internet database
- Crossref database
- 17% Submitted Works database

- 20% Publications database
- Crossref Posted Content database

Explainable AI for Time Series Model Interpretability

Submitted in partial fulfilment of the requirements
of the degree of

B.E. Computer Engineering

By

Bhavi Dave	**60004180013**
Nirali Parekh	**60004180065**
Raj Shah	**60004180076**

Guide:

Dr. Kriti Srivastava
Head of Department

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

University of Mumbai
2021-2022

CERTIFICATE

This is to certify that the mini project entitled **"Explainable AI for Time Series Model Interpretability"** is a bonafide work of **"Bhavi Dave (60004180013), Nirali Parekh (60004180065), and Raj Shah (60004180076)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering

Dr. Kriti Srivastava
Guide

Dr. Meera Narvekar
Head of Department

Dr. Hari Vasudevan
Principal

Major Project Report Approval for B.E.

This mini project report entitled **Explainable AI for Time Series Model Interpretability** by **Bhavi Dave, Nirali Parekh, and Raj Shah** is approved for the partial fulfillment of the degree of *B.E. in Computer Engineering.*

Examiners

1.---
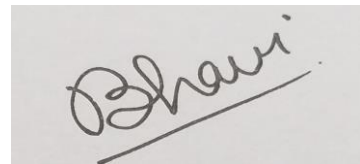
2.---

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Bhavi Dave   60004180013

Nirali Parekh   60004180065

Raj Shah   60004180076

Abstract

Model Interpretability is critical for analyzing the potential risks inherent to the utilization of black-box neural networks in the medical domain. The current frameworks and methodologies for Parkinson's disease detection require considerable expertise from medical professionals alongside an expensive and time-consuming procedure. On the other hand, Artificial Intelligence (AI) driven Deep Learning models have generated state-of-the-art performance across the medicinal domain and have the potential to improve the existing process for PD detection by leveraging the early stage symptom of vocal deterioration to enable early identification of the disease in a cost effective and accurate manner. However, the potential of Deep Learning for Parkinson's Disease selection has remained untapped thus far because of the lack of model interpretability and the consequent potential of even an accurate model to fail to capture the correct reasoning and the causal relationship between the input signal and the prediction. This leaves the model susceptible to both deliberate attacks using engineered noises, or even a co-incident minor noise to push the model to generate an output with fatally high confidence. A thorough understanding of model bias, accountability and security is critical to the adoption of AI in the realm of medicine. This work implements a black-box neural network trained on the "Synthetic vowels of speakers with Parkinsonism" dataset that contains vowel phonations of PD patients and a healthy control group. The black-box model trained on this classification task achieved a classification accuracy of 90.32% with a precision, recall and F1 Score of 0.91, 0.90 and 0.905 respectively. A novel model interpretation algorithm is proposed that divides the audio file into logical chunks and provides a category label identifying the productivity of each chunk towards correct inference. This is based on its interactions with every other audio region in the file, and the effective change of model performance and confidence with its inclusion. A particular audio instance within a clip is significant for desired output only in conjunction with other audio regions that together generate a meaningful pattern indicating the existence of Parkinson's disease. Alongside this category label for each audio chunk, the algorithm also ranks all audio chunks in the order of their importance to the correct prediction by analyzing the degree in which their impact varies when combined with other audio regions. This paper successfully demonstrates that Deep Learning algorithms can be leveraged to create a reliable, accurate, and efficient method for PD Detection.

Contents

List of Figures

List of Abbreviations

Sr. No.	Abbreviation	Expanded form
i	PD	Parkinsons Disease
ii	LSTM	Long Short-Term Memory
iii	DNN	Deep Neural Networks
iv	CDCML	Cross-modal deep continuous metric learning
v	SVM	Support Vector Machine
vi	CV	Cross Validation
vii	FFT	Fast Fourier Transform
viii	DCT	Discrete Cosine Transform
ix	MSE	Mean Square Error
x	CNN	Convolutional Neural Networks
xi	RNN	Recurrent Neural Networks
x	VMD	Variational Mode Decomposition
xi	AUC	Area under the ROC Curve
xii	XAI	Explainable AI
xiii	SHAP	Shapley Additive Explanations

List of Tables

Sr. No.	Table no.	Label
i	2.1.1	Literature Review
ii	4.2.1	Category Classification for Audio Chunks
iii	4.4.1	Results
iv	4.4.2	Rank Ordering of Audio Chunks

Chapter 1. Introduction

1.1 Description

Machine Learning (ML), a branch of Artificial Intelligence (AI), has gained increasing interest from the scientific community, since it allows humans to support decision-making processes in several fields. However, despite their success, ML based approaches have some limitations and disadvantages. Among them, the poor explainability of ML complex models has hampered their extensive application in those fields where the impact of the decision is critical, as it is the case for medical application. Indeed, the more a decision can influence people's lives, the more important it is to understand which are the factors that lead to that particular decision. Explainable AI is artificial intelligence in which the results of the solution can be understood by humans. It contrasts with the concept of the "black box" in machine learning where even its architecture designers cannot explain why an AI arrived at a specific decision.

The medical field sees a vast range of applications of Machine Learning like tumor detection and disease classification. However, physicians cannot simply use the predictions of the model but should also trust the results obtained. To trust them, physicians have to understand how and why these decisions were made and compare these reasons with their previous domain knowledge. Consequently, the black-box approach implemented by most of ML classification systems, which do not clearly indicate the adopted decision rules, has hampered their use by clinicians. With the growing number of the aged population, the number of Parkinson's disease (PD) affected people is also mounting. Unfortunately, due to insufficient resources and awareness in underdeveloped countries, proper and timely PD detection is highly challenged. Besides, all PD patients' symptoms are neither the same nor they all become pronounced at the same stage of the illness. Mostly used diagnostic tools include psychometric tests, neuroimaging, and cerebrospinal fluid examination, though these methods are time-consuming, expensive and invasive, respectively.

An effective screening process, particularly one that doesn't require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and

1

non-invasive tool for diagnosis. If explainable algorithms could be applied to a voice recording dataset to accurately diagnose PD, this would be an effective screening step prior to an appointment with a clinician. With the growing number of the aged population, the number of Parkinson's disease (PD) affected people is also mounting. Unfortunately, due to insufficient resources and awareness in underdeveloped countries, proper and timely PD detection is highly challenged. Besides, all PD patients' symptoms are neither the same nor they all become pronounced at the same stage of the illness. Mostly used diagnostic tools include psychometric tests, neuroimaging, and cerebrospinal fluid examination, though these methods are time-consuming, expensive, and invasive, respectively. Techniques that analyze speech and vocal patterns might be effective tools to diagnose Parkinson's disease, and possibly at earlier stages. There are no laboratory biomarkers that can detect Parkinson's, and brain imaging scans do not allow for a definitive diagnosis. The clinical diagnosis of the disease is currently based on the manifestation of two to three motor symptoms, including muscle stiffness, resting tremors, slowness of movement, and balance issues. These criteria can identify Parkinson's with 90% accuracy, but it takes, on average, 2.9 years to reach a diagnosis. Speech is a complicated skill, and it's often affected by Parkinson's-associated motor changes. Between 60 and 80% of patients may experience reduced vocal loudness, harsh or breathy vocal quality, and abnormal speaking rates. The early detection of Parkinson along with the anticipation of the start of treatment would have a relevant effect on both the quality of life of patients and the healthcare system. An interpretable algorithm can help encourage applicability since physicians are required to understand how and why model decisions were made and compare these reasons with their previous domain knowledge. An algorithm that clearly indicates the adopted decision rules, can help overcome the previously hampered adoption of ML systems.

1.2 Problem Formulation

As our major project, we propose to impart explainability to the black-box decision-making of the Deep Learning models. The problem of Parkinson's disease classification from audio data is addressed to test our proposed novel XAI algorithm. Considering audio signals as time-series data, we propose a novel algorithm that imparts interpretability by masking audio chunks in the series at a time and analyzes the change in model performance to gauge their importance in decision making.

1.3 Motivation

XAI can simply be described as aiming to make AI systems more understandable to humans. "Black box" AI systems that give predictions without any explanation are problematic for numerous reasons, not only because of their lack of transparency but also because they hide potential biases within the system. Techniques that analyze speech and vocal patterns might be effective tools to diagnose Parkinson's disease, and possibly at earlier stages . There are no laboratory biomarkers that can detect Parkinson's, and brain imaging scans do not allow for a definitive diagnosis. The clinical diagnosis of the disease is currently based on the manifestation of two to three motor symptoms, including muscle stiffness, resting tremor, slowness of movement and balance issues. These criteria can identify Parkinson's with 90% accuracy, but it takes, on average, 2.9 years to reach a diagnosis. Speech is a complicated skill, and it's often affected by Parkinson's-associated motor changes. Between 60 and 80% of patients may experience reduced vocal loudness, harsh or breathy vocal quality and abnormal speaking rates. The early detection of Parkinson along with the anticipation of the start of treatment would have a relevant effect on both the quality of life of patients and the healthcare system.

Medicine has long reached an overwhelming consensus on the importance of detecting Parkinson's disease (a degenerative neurological disorder marked by decreased dopamine levels in the brain) in a timely manner. With the advent of the domain of HealthTech, Deep Learning approaches have generated State Of The Art performance in solving several problems in the medicinal arena. Black box neural networks give predictions without any explanation and are problematic for numerous reasons, including their lack of transparency and potential hidden biases within the system. We propose the creation of an algorithm that can interpret black-box neural networks trained to detect Parkinson's Disease from voice recordings. It leverages characteristic vocal features and explains the model's reasoning for this inference.

Our architecture is modeled on a typical black-box neural network and extends the architecture to provide interpretability to model the aforementioned model inferences. This would overcome the shortcomings of the traditional method of diagnosis where a physician is required to perform a tedious analysis of a person's motor skills in various situations. Vocal deterioration is among the

first symptoms of the disorder and accurate and reliable models for the same can help with early detection and therefore promote better treatment. Our model can serve as an effective non-invasive screening tool. As deep learning continues to be adopted, the prominence of assistance and automated decisions made by neural networks in high stake situations is an undeniable fact that stakeholders and academicians have to grapple with. An XAI model that detects PD leverages the performance of a neural network while providing reasoning and accountability to inferences, reducing bias. Consequently, our algorithm can help the adoption of neural networks for Parkinson's disease identification by proving interpretability and therefore reasoning for model inferences. The potential for human scrutiny, interpretation and verification can make the model robust and reliable, increasing confidence.

1.4 Proposed Solution

With the advent of the domain of HealthTech, Deep Learning approaches have generated State Of The Art performance in solving several problems in the medicinal arena. We propose the creation of a novel, explainable sequence learning problem that can detect Parkinson's Disease from voice recordings of characteristic vocal features and explain the reasoning for this inference. Our sequence learning architecture is modeled on the basis of a Convolutional Neural Network, and extends the architecture to provide interpretability to inferences. This would overcome the shortcomings of the traditional method of diagnosis where a physician is required to perform tedious analysis of a person's motor skills in various situations. Our model can serve as an effective non-invasive screening tool, promoting early detection.

1.5 Scope of the Project:

The first stage is associated with a review and analysis of both XAI and Deep Learning approaches alongside the creation of a pipeline for PD audio data that can later be leveraged in XAI modeling. A novel algorithm is then to be designed to impart explainability and interpretability to the inference provided by the 'black-box' neural network architecture.

Chapter 2. Review of Literature

Various review works have been focused on the use of speech recognition in the problem of Parkinsons Disease prediction. In DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification (H. Zang et al) [1] propose a voiceprint-based PD identification application that integrates deep learning and mobile health. They use a convolutional neural network (CNN) for the final identification. In Machine Learning using speech utterances for Parkinson Disease Detection (Ondřej Klempíř et al.) [10], use changes in speech and articulation as significant biomarkers for detecting PD. They trained several recognition methods including AdaBoost, Bagged trees, Quadratic SVM and k-NN.

In Robust automated Parkinson disease detection based on voice signals with transfer learning [2], A Klempir et Al trained Convolutional neural networks (CNN) like SqueezeNet1_1, ResNet101, and DenseNet161 for automated PD identification based on biomarkers-derived voice signals. O Karaman et Al in Robust Feature Engineering for Parkinson Disease Diagnosis: New Machine Learning Techniques [3] used traditional ML technique of support vector machine and evaluated it with 10-fold cross-validation (CV), with stratification for balancing the number of patients and controls for each CV fold.

In Detection of Parkinson Disease Using Variational Mode Decomposition of Speech Signal [4], M Yang et Al used variational mode decomposition (VMD) for extracting relevant information of speech signals and used various statistical features (mean, variance, skewness and kurtosis), energy and energy entropy for Parkinson disease detection.

Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques [5], K Biswajit et al propose eighteen feature extraction techniques and four machine learning methods to classify PD data obtained from sustained phonation and speech tasks[6]. Their main contribution is speech signal processing: both traditional and novel speech signal processing technologies have been employed for feature engineering, which can automatically extract a few linear and nonlinear dysphonia features. [7]-[9]

Table 2.1.1: Literature Review

Ref	Paper	Methodology	Performance
[1]	DeepVoice : H. Zang et al.	CNN + Joint Time-Frequency Analysis algorithm	Accuracy : 90.45 ± 1.71%
[2]	A. Klempíř et al.	AdaBoost, Bagged trees, Quadratic SVM and k-NN	Average overall accuracy : 82.3 % Averaged AUC = 0.88
[3]	O. Karaman et al.	DenseNet-161 (CNN) + transfer learning	Accuracy: 89.75% Sensitivity: 91.50%, Precision: 88.40%
[4]	M. Yang et al.	Support Vector Machine + 10-fold cross-validation (CV)	Accuracy > 58%.
[5]	K. Biswajit et al.	Variational Mode Decomposition (VMD)	Accuracy : 96.29%
[6]	J. Almieda et al.	18 feature extraction techniques + 4 machine learning methods	Accuracy : 94.55% AUC 0.87 EER 19.01%.
[7]	L. Zhang et al.	Novel speech signal processing + classical regression and classification algorithms	Accuracy : 88.74% Recall : 97.03% Mean absolute error : 3.7699

| [8]	S. Grover et al.	ResNet architecture + artificial images representing the spectrogram of the voice signal	Accuracy : 90 %
[9]	C. Abdullah et al.	Satacked autoencoder and a softmax classifier	Accuracy rate : 86.095 %
[10]	Ondřej Klempíř et al.	[11] AdaBoost, Bagged trees, Quadratic SVM and k-NN	Accuracy : 82.3 % Average AUC : 0.88

2.1 Research Gaps:

- Currently there are no explainable algorithms that are designed specifically for audio. All existing XAI architectures are focused towards explaining images.
- The existing XAI models can be used for audio files by converting them into images. But this requires the files to be of the same duration. It cannot deal with audio files having variable length.
- There is no generalizable model that can be applied across several other problem statements including temporal data or that require sequence modeling.

Chapter 3. Design

3.1 Architectural Design of the Proposed System

We broach the problem statement with a two-step approach - create a deep neural network model that accurately classifies an audio sample into the two relevant classes of Healthy control and Parkinsons's disease, and then proceed to add explainability by using our novel XAI algorithm.

In order to be able to extend our neural network to integrate interpretability, we also have to model our audio data in a novel manner. In our pipeline, audio data is converted to tensors using sampling. The sample rate is a hyperparameter to be tuned. After the audio is sampled, a tensor from every second of audio is collected, standardizing the representation of one second of audio. Our model can work with an audio of any given length, because of the aforementioned standardization. A multi-second audio file is simply treated as a collection of the one second standardizations. If an audio is not large enough to completely fill a chunk, zeros are padded at the trailing end to maintain the generalization. Each of the corresponding audio files is converted to model input using this pipeline.
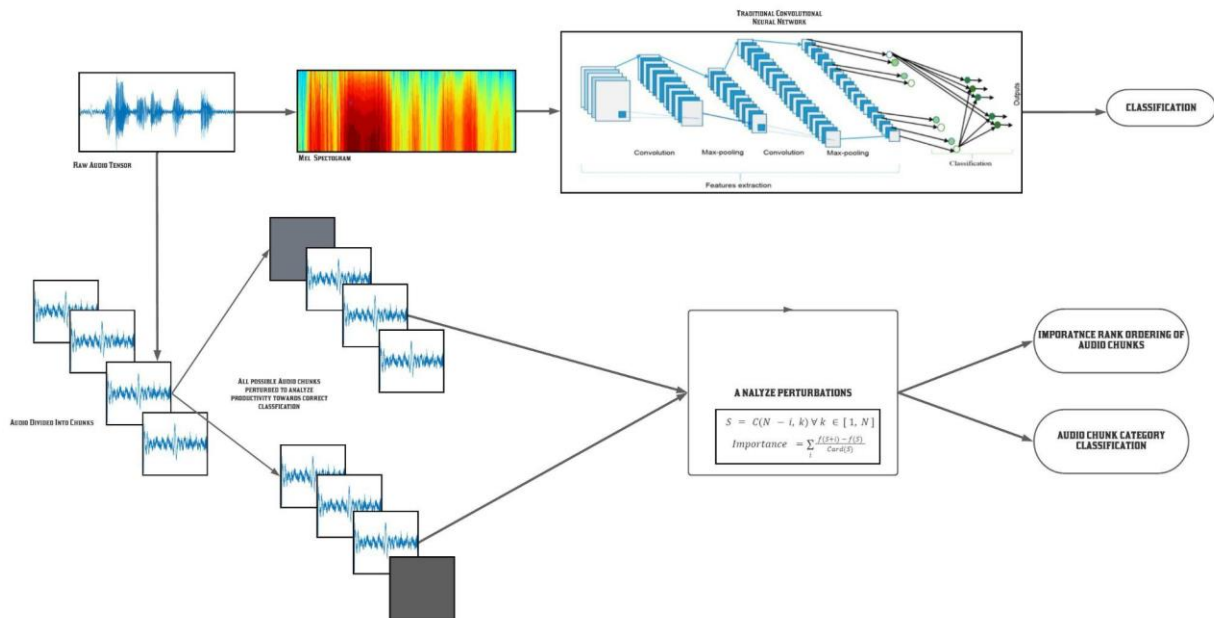
The proposed algorithm is a model agnostic surrogate model that leverages the variance in changes in the prediction to interpret black-box neural networks for audio classification. The trivial and simplified version of the model intuition is that if the model prediction does not change much by tweaking or removing the value of an audio chunk, the particular audio chunk is likely not an important predictor because its effect on the desired model output is negligible. Conversely, the model relies most on audio regions that, when tweaked or removed, generate maximal deviance in the ability to correctly classify the audio file.

This hypothesis is further developed to include cases where a particular audio instance within a clip is significant for desired output only in conjunction with other audio regions that together generate a meaningful pattern indicating the existence of Parkinson's disease. That is, an audio chunk might not alone be indicative of Parkinson's disease, but when seen in the context of other audio regions from the same clip, indicate the same. It is for this reason that the model output cannot be attributed to a single audio chunk without analyzing it with all the other audio chunks. Furthermore, since every person has a distinct and recognizable voice, certain sections of audio might be indicative of degeneracy only within the context of the entire audio clip and the unique benchmarks that other regions of the voice recordings provide. Without this context and interdependence, even a potentially important second loses its significance. Therefore, a model cannot attribute importance to an audio region solely based on its presence or absence - we must take into account all the possible interactions with other audio regions that lend credence to a particular datapoint. The proposed algorithm aims to generate a sophisticated model of these interdependencies in a way where the true contribution of each audio region can be estimated, along with the confidence of the estimation.
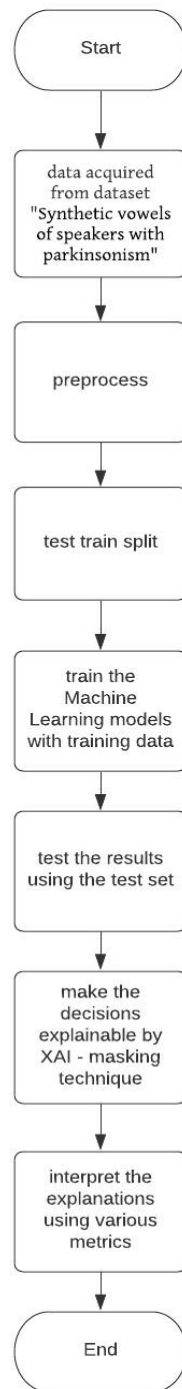
3.2 Process Pipeline

Fig 3.2.1: End-to-end pipeline of the proposed XAI system.

Chapter 4. Implementation

4.1 Data Description:

When it comes to machine learning, no element is more essential than the quality of the training data. The preliminary description of data is examined for implementing this project. We explored various datasets related to voice recordings of Parkinson's patients. And the dataset named "Synthetic vowels of speakers with parkinsonism" that was released in 2019 suits our purpose the best. This dataset contains recordings of sustained vowels 'A' and 'I' performed by two sets of subjects: healthy people, and patients with Parkinson's disease. The dataset contains synthesized replicas of sustained vowels 'A' and 'I' performed by healthy controls and patients with Parkinson's disease. The dataset can be used as a reference for extracting features such as elevation, pitch, frequency, and subharmonics from the voices that can be learned by the deep learning models. The image given here is a sample of such voice features from the dataset instances. Along with the numeric features, the dataset also contains the raw voice recordings collected during experimentation and its corresponding sound files that are synthetically cleaned of any external noises. This cleaned signal is to be provided to the Machine Learning and explainable AI models described in previous slides to make them more versatile.

The dataset can be used as a reference for the evaluation of pitch detectors, detectors of the modal fundamental frequency, and detectors of subharmonics.

patientID_vowel.wav = waveform of the synthesized replica. This is the raw signal collected. Parameters of jitter, shimmer, and harmonic to noise ratio (HNR) are also included in the dataset.

patientID_vowel_clean.wav = waveform of the synthesized replica without added noise. This signal is to be provided to the Machine Learning and XAI models to make them more versatile.
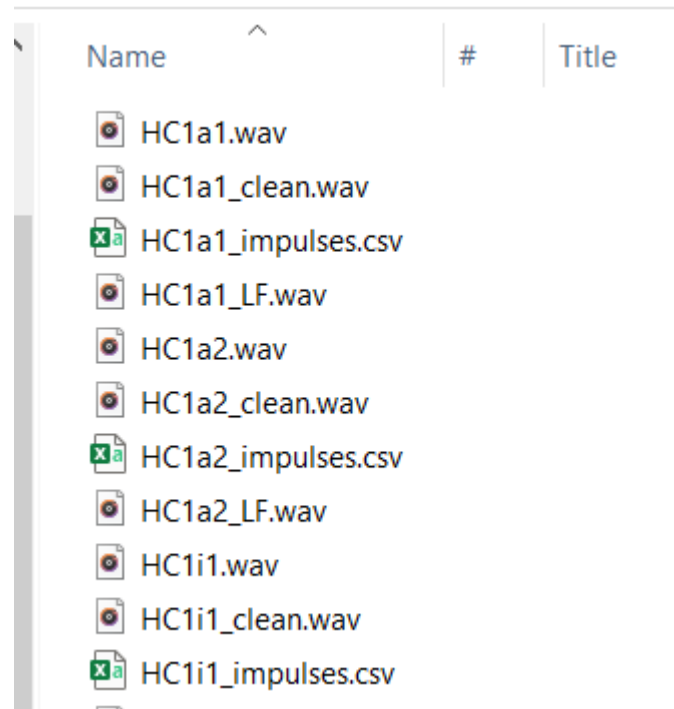
Fig 4.1.1: Sample files in the dataset "Synthetic vowels of speakers with parkinsonism"

record	subject	group	vowel	jitter	shimmer	HNR	SHR	
HC1a1	HC1	HC	a	0.332226	3.225	22.942	0	
HC1a2	HC1	HC	a	0.344828	2.73	23.677	0	
HC1i1	HC1	HC	i	0.4329	1.567	20.838	0	
HC1i2	HC1	HC	i	0.383142	1.759999	22.681	0	
HC2a1	HC2	HC	a	0.412371	3.732	20.715	0	
HC2a2	HC2	HC	a	0.407332	3.436001	19.767	0	
HC2i1	HC2	HC	i	0.403226	2.064999	27.445	0	
HC2i2	HC2	HC	i	0.4	2.077999	27.348	0	
HC3a1	HC3	HC	a	0.638978	4.337	16.459	0	
HC3a2	HC3	HC	a	0.634921	5.248	18.517	0	
HC3i1	HC3	HC	i	0.340136	3.717	24.534	0	
HC3i2	HC3	HC	i	0.344828	4.297	25.796	0	
HC4a1	HC4	HC	a	0.426439	2.483001	22.641	0	
HC4a2	HC4	HC	a	0.213675	2.851	20.317	0	
HC4i1	HC4	HC	i	0.428266	2.852	24.325	0	
HC4i2	HC4	HC	i	0.639659	3.155	23.522	0	
HC5a1	HC5	HC	a	0.492611	3.449	19.533	0	
HC5a2	HC5	HC	a	0.980392	6.837	16.879	0	
HC5i1	HC5	HC	i	0.246305	3.769	22.439	0	
HC5i2	HC5	HC	i	0.977995	7.543001	16.919	0	

Fig 4.1.2: Sample of the tabular data obtained from the dataset "Synthetic vowels of speakers with parkinsonism"

Exploratory data analysis of the audios revealed quite some differences between the clarity of speech of healthy subjects and Parkinson's patients. Subjects with Parkinson's delivered slurred words, mumbling or even stuttering. As we can notice, the speech of Parkinson's' patients sounds breathy or hoarse. Automated techniques to analyze such vocal patterns can give an early diagnosis of Parkinson's disease. Also, research shows that **89% of people** with Parkinson's disease (PD) experience speech and voice disorders, including soft, monotone, breathy and hoarse voice and uncertain articulation.

Fig 4.1.3: Voice analysis of a healthy person.

Fig 4.1.4: Voice analysis of Parkinson's' patient

Along with the audio files and various voice features, we also analyzed their speech synthesis by the use of spectrograms. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. The two images below depict the spectrograms that we created of sample recordings we heard in the previous slide. As it's evident through the voiceprints, the frequency distribution of a patient with Parkinson's disease varies significantly compared to a healthy patient. For our major project, we aim to utilize various media obtained through this dataset like the audio signals, and the voice features to impart explainability to deep learning models.
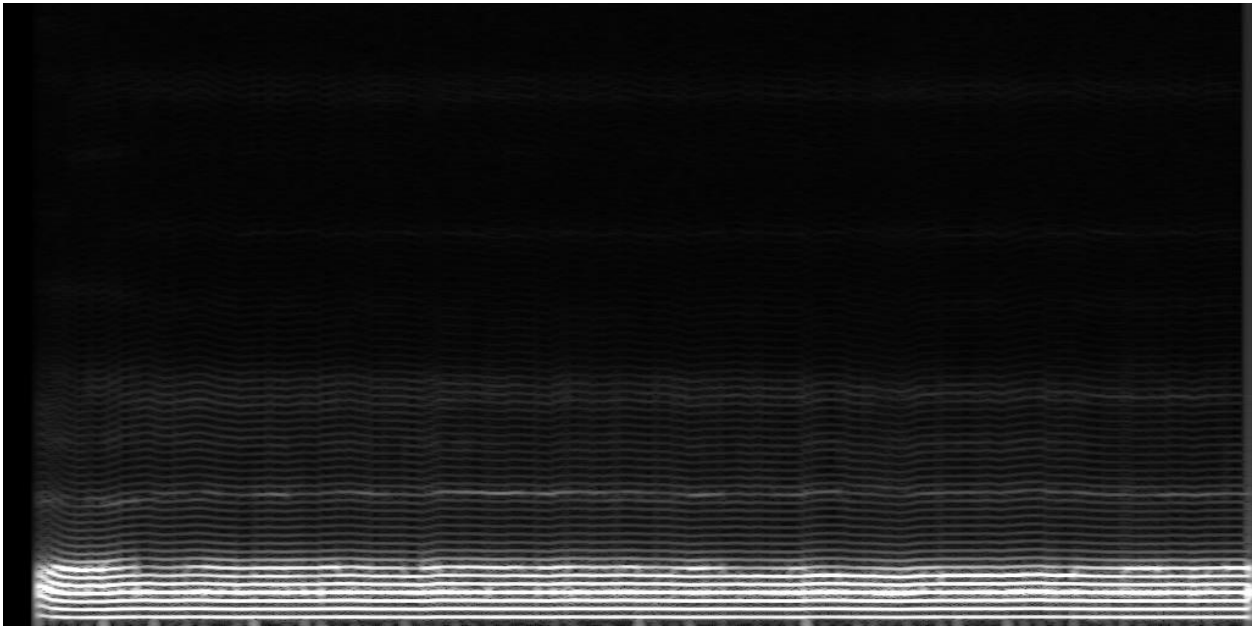
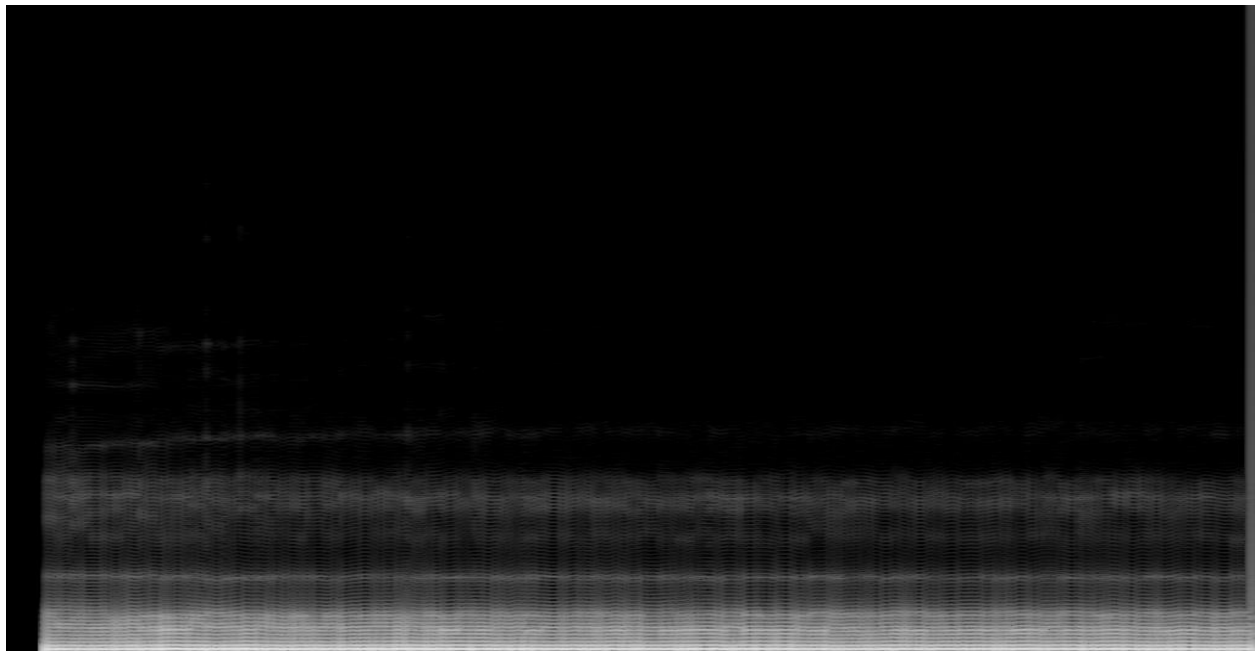Fig 4.1.5: Spectrograms of a healthy subject speaking the vowel 'A'

Fig 4.1.6: Spectrograms of a Parkinsons' patient speaking the vowel 'A'

4.2 Methodology

[6] Each sample is the amplitude of the wave at a particular time interval, where the bit depth determines how detailed the sample will be, also known as the dynamic range of the signal. We have sampled the audio data. [6] In signal processing, sampling is the reduction of a continuous signal into a series of discrete values. The sampling frequency or rate is the number of samples taken over some fixed amount of time. A high sampling frequency results in less information loss but higher computational expense, and low sampling frequencies have higher information loss but are fast and cheap to compute. We have sampled audio files with a sample rate of 22,050. . [22] 22050 kHz (often called "22 kHz") has been a reasonably popular sample rate for low bit rate MP3s such as 64 kbps in years past. After the audio is sampled, a tensor from every second of audio is collected, standardizing the representation of one second of audio. Our model can work with an audio of any given length, because of the aforementioned standardization. A multi-second audio file is simply treated as a collection of the one second standardizations. If an audio is not large enough to completely fill a chunk, zeros are padded at the trailing end to maintain the generalization.

Figure 6 shows an audio sample from the dataset. The length of the audio tensor of this sample is 3,75,962. As described earlier, the audio is sampled with the Sample Rate of 22,050.
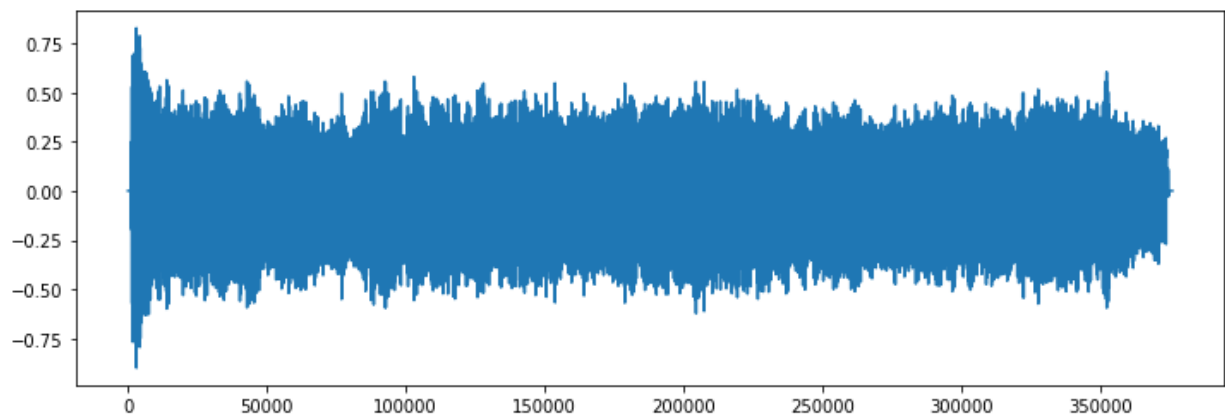

Figure 4.2.1: Sample Audio File

After converting this audio file into chunks using the process described above, we generate 18 chunks with 22,050 points in each one because of the audio file length. Since the audio length is

greater than 17 chunks, but not large enough to complete chunk 18, padding of zeros is utilized to maintain uniformity of representation in 1 second of audio data.
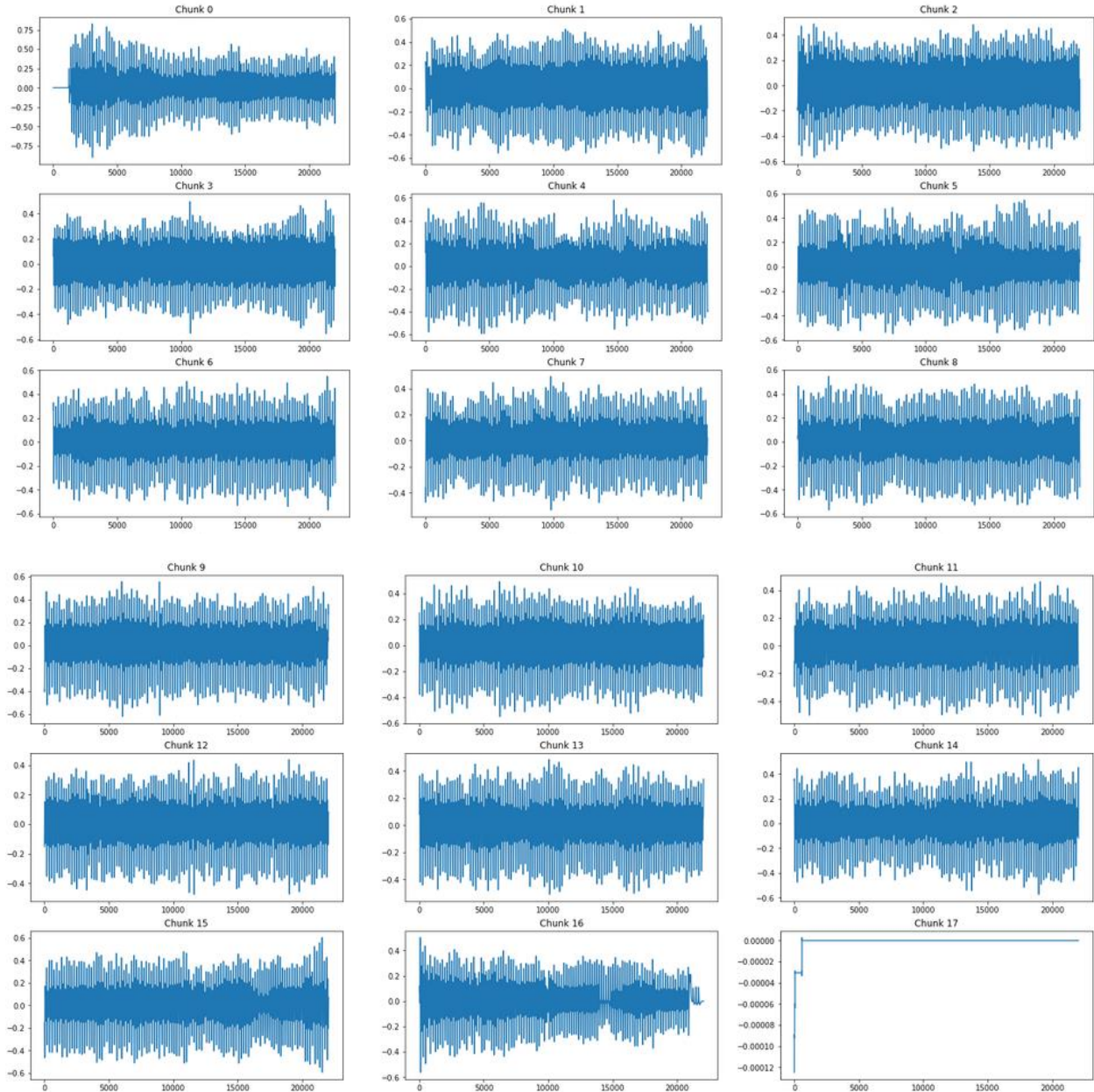

Figure 4.2.2: Sample Chunks corresponding to Audio File

This model will therefore be trained to perform binary audio classification - given an audio file, it predicts whether the speaker belongs to the Parkinson's disease class or the healthy control class.

A black-box deep neural network was trained on the dataset since the review of the literature demonstrated that the methodology achieved state-of-the-art performance albeit at the expense of interpretability. This flaw will later be corrected with our algorithm. A convolutional neural network (CNN) was selected because it gains relevant, abstract, and location invariant features while the max-pooling layer helps reduce complexity while allowing prominent features to pass through to the upper layers of the network. This greatly reduces the number of parameters while achieving a similar degree of performance because of sparse connectivity and weight sharing.

The audio files are read as a tensor and resampled to a fixed sample rate of 22050 kHz. The time-series signal is then transformed into the image domain by converting it into a spectrogram that represents the frequency content in the audio file as image colors. The work employs a log-scaled Mel-spectrogram that converts frequencies to the mel scale that takes into account human sensitivity to different ranges of audio frequency. This neural network is therefore trained to perform binary audio classification - given an audio file, it predicts whether the speaker belongs to the Parkinson's disease class or the healthy control class. The model was trained on the same partition of the dataset to avoid overlap, ensuring that the test data points were unseen by all three models. 15% of the dataset was reserved for testing, and the same other 80% was used to train the model. Stochastic gradient descent was used as an optimizer with a momentum of 0.9. The criterion used to evaluate loss for backpropagation was Cross-Entropy Loss. The network utilized a learning rate scheduler that decayed the learning rate of each parameter group by 0.1 every 3 epochs. The performance of the black-box model is summarized in the results section.

After the neural network is trained, the proposed algorithm for model interpretability is applied. Every audio file is sampled and a tensor from every second of audio is collected, standardizing the representation of one second of audio as an "audio chunk".   The algorithm can provide interpretability to model predictions for audio of any given length, because of the aforementioned standardization. A multi-second audio file is simply treated as a collection of one-second

17

standardizations. If audio is not large enough to fill a chunk, zeros are padded at the trailing end to maintain the generalization.

Since every region of audio can interact with every other audio region of varying size, the interpretable algorithm considers each of these interactions and their effect on the black-box neural network's output. While predicting masked the class of an audio file with a masked chunk, there are 6 cases possible as is highlighted in the Table below-

Table 4.2.1: Category Classification for Audio Chunks

Case	Initial Prediction	Prediction after Masking	Change in Confidence	Category
1.	Correct	Correct	Higher	Unproductive
2.	Correct	Correct	Lower	Marginally productive
3.	Incorrect	Incorrect	Higher	Productive
4.	Incorrect	Incorrect	Lower	Marginally unproductive
5.	Correct	Incorrect	NA	Crucially Productive
6.	Incorrect	Correct	NA	Extremely Unproductive

We can therefore immediately label every audio second as productive or unproductive with a granularity of 6 classes. The model should predict the correct class with high confidence and the incorrect class with low confidence. Therefore, if a once correctly predicted class is predicted with higher confidence after an audio chunk is removed, that audio chunk was unproductive (Case 1) whereas if the correct class was predicted with lower confidence after an audio chunk's removal the performance deterioration can be attributed to the audio chunk being at least marginally productive (Case 2). Similarly, if the confidence of the incorrect class increases after an audio

chunk is removed, the chunk is productive (Case 3) and if the confidence of the incorrect class decreases after removal, it is only marginally unproductive (Case 4). In case 5, an audio chunk when masked changes the model prediction from the correct class to the incorrect class, signifying that the removed audio chunk was 'Crucially Productive'. Conversely, in case 6, an audio chunk when masked changes the model prediction to the correct class from the incorrect class, implying that it is a counterproductive audio region and extremely productive to the model output. Values that most significantly hinder the model's ability to predict correctly are classified to be 'Crucially Productive' and the seconds that actively hinder a model's ability to classify are classified as 'Extremely Unproductive'.

We define the productivity of an audio chunk as the percentage difference in the confidence of the correct class when the section of audio is excluded.

$$Productivity\ of\ chunk\ = \frac{New\ Confidence\ of\ correct\ class\ -\ Original\ Confidence\ of\ correct\ class}{Original\ Confidence\ of\ correct\ class} \quad (1)$$

Since every second of audio data does not enhance or degrade performance in isolation, the delta in model performance is attributed to its interaction with any other second in the audio in groups of all possible sizes. That is, the importance of an audio chunk is calculated based on its productivity value calculated in association with unique combinations of sizes ranging from one to the audio length. Therefore, all possible combinations of subsets S are considered. The algorithm, therefore, considers all possible interactions and combines the productivity results in all cases. Before aggregation, these values are normalized by the number of audio data points in the group (i.e. the cardinality of S) since the deviance in performance can be attributed to all the chunks in the group. Therefore, the lesser the number of chunks in the group, the more the impact of the particular chunk being considered. These values are then aggregated to generate a single value indicating the importance of this audio chunk. The audio chunk that has the highest value for this metric is the most important second.

$$Importance\ of\ an\ Audio\ Chunk\ = \frac{1}{Card(S)}\sum_{\forall S}\ Productivity(S) \quad (2)$$

Consequently, all the audio chunks that compose the audio file can be ranked according to the black-box neural network's interpretation of their productivity, and consequently, their importance. The deviation in these values for every second shows the consistency with which the interpretable algorithm produces the results and can therefore be attributed to the XAI model's confidence in the productivity value. The most productive seconds are the most important ones.

4.3 Algorithm

If N is the length of the Audio file, f is a probability function demonstrating the output of a black-box neural network, S denotes sunsets and C denotes combinations then the importance of second i can be calculated in by iteratively forming subsets and aggregating a weighted importance value normalized by the cardinality of the subset.

$$S = C(N - i, k) \, \forall \, k \in [\, 1, N \,] \tag{3}$$

$$Importance = \sum_i \frac{f(S+i) - f(S)}{Card(S)} \tag{4}$$

The algorithm, therefore, provides interpretability to a black box neural network output.

4.4 Results

This section outlines the evaluation measures and the results obtained by the black-box neural network as well as the interpretability provided by the novel algorithm. The Evaluation measures include -

1. Accuracy: It is the fraction of classifications the model got correct.

$$Accuracy = \frac{True \ Positive + True \ Negative}{True \ Positive + True \ Negative \ False \ Positive + False \ Negative} \tag{5}$$

2. Precision: Proportion of the positive identifications which are correct.

$$Precision = \frac{True \ Positive}{True \ Positive + False \ Positive} \tag{6}$$

3. Recall: Proportion of actual positives which were identified correctly.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{7}$$

4. F1-score: It is calculated as the harmonic mean of precision and recall.

$$F1\ Score = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall} \tag{8}$$

The results are summarized in the Table. All metrics are calculated on unseen test data.

Table 4.4.1: Results

Metric	Model Performance
Precision	0.91
Recall	0.90
F1 Score	0.905
Accuracy	90.32%

Furthermore, an example of the per chunk analysis performed by the model can be seen in the Figure. As described in the methodology section, the model-interpretability algorithm considers all possible interactions between subset combinations of audio chunks and generates a category label for each audio chunk using majority voting depending on the weighted values of change in model performance.
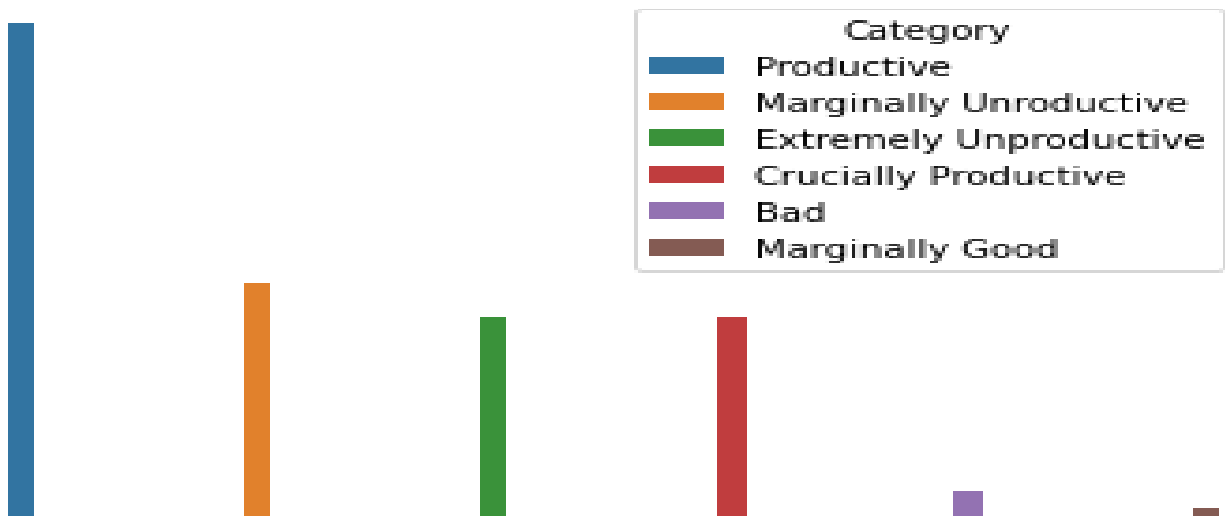
Figure 4.4.1: Audio Chunk Category Classification

As described in the methodology section, for every 'n' second audio, the model also ranks all audio chunks in the order of importance and associates a confidence value associated with the rank. This helps determine the relative importance of audio chunks and the degree by which their impact varies when combined with other audio regions. An example output for a 10 second audio can be seen below in Table 4.4.2.

Table 4.4.2: Rank Ordering of Audio Chunks

Audio Chunk Number	Importance Rank	Confidence
1	1	0.82
4	2	0.71
7	3	0.54
2	4	0.26
9	5	0.84
0	6	0.91
6	7	0.76
8	8	0.79
5	9	0.46
3	10	0.53

5. Conclusion

To address the pitfalls associated with the detection of Parkinson's disease, our project has implemented a novel, interpretable algorithm. Given voice recordings of characteristic vocal features, we predict whether a voice sample belongs to the set of Parkinson's Disease patients or is amongst the group of healthy control. This non-invasive tool for diagnosis can help promote early detection and treatment, overcoming the tedious traditional method. Adding explainability can help the implementation of the algorithm at scale, especially in a sensitive domain like medicine. Since we can follow the model's chain of reasoning, the likelihood of bias is low and is susceptible to human verification. Our novel XAI approach will be among the first Explainable architectures for audio and does not place restrictions on the input audio data, coping with variable lengths. Our method of modeling Sequence data and revising traditional sequence models are highly generalizable and can be implemented across several other problem statements that include temporal data or require sequence modeling, and hence is a step toward the integration of XAI in temporal data. The algorithm can be improved by further analyzing ways of masking audio chunks to further explore how the absence of a signal affects the output and consequently deriving its importance. There is also the possibility of improving the speed of inference by using only an intelligently sampled subset of audio chunk combinations instead of all possible combinations. This has the potential of significantly reducing both training and inference time without compromising performance.

The successful validation of the hypothesis outlined in this paper not only suggests an important application for Deep Learning algorithms in the realm of medicine but also paves way for better actual applicability since decisions taken by the model with regard to Parkinson's disease classification can be interpreted. The method can also be generalized and extended to interpret several other neural network architectures that classify temporal data and can consequently create a huge, positive disruption in the way medical networks make crucial decisions, allowing for optimized and unbiased analysis that can save tremendous amounts of resources that otherwise would go to waste on account of an inaccurate or incomplete understanding of the domain.

6. Papers Published :

- *"Automatic Sleep Stage Scoring on Raw Single-Channel EEG: A comparative analysis of CNN Architectures"*- Accepted at the 4th International Conference on Information Technology, Computer and Electrical Engineering

- *"Convolutional Neural Networks for Audio Classification: An Ensemble Approach"*- Accepted at the 6th International Conference on Advanced Computing and Intelligent Engineering

7. References

[1] Hlavnička, Jan; Čmejla, Roman; Klempíř, Jiří; Růžička, Evžen; Rusz, Jan (2019): Synthetic vowels of speakers with Parkinson's disease and Parkinsonism. figshare. Dataset. https://doi.org/10.6084/m9.figshare.7628819.v1

[2] H. Zhang, A. Wang, D. Li and W. Xu, "DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification," 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018, pp. 214-217, doi: 10.1109/BHI.2018.8333407.

[3] Klempíř, O. and Krupička, R., 2018. Machine learning using speech utterances for parkinson disease detection. Lékař a technika-Clinician and Technology, 48(2), pp.66-71.

[4] Karaman, O., Çakın, H., Alhudhaif, A. and Polat, K., 2021. Robust automated Parkinson disease detection based on voice signals with transfer learning. Expert Systems with Applications, 178, p.115013.

[5] Wang, M., Ge, W., Apthorp, D. and Suominen, H., 2020. Robust Feature Engineering for Parkinson Disease Diagnosis: New Machine Learning Techniques. JMIR Biomedical Engineering, 5(1), p.e13611.

[6] Karan, B., Mahto, K. and Sahu, S.S., 2018, April. Detection of Parkinson disease using variational mode decomposition of speech signal. In 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 0508-0512). IEEE.

[7] Almeida, Jefferson S., et al. "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques." Pattern Recognition Letters 125 (2019): 55-62.

[8] Zhang, Liang, et al. "An intelligent mobile-enabled system for diagnosing Parkinson disease: Development and validation of a speech impairment detection system." JMIR Medical Informatics 8.9 (2020): e18689.

[9] Zhang, Liang, et al. "An intelligent mobile-enabled system for diagnosing Parkinson disease: Development and validation of a speech impairment detection system." JMIR Medical Informatics 8.9 (2020): e18689.

[10] Grover, S., Bhartia, S., Yadav, A. and Seeja, K.R., 2018. Predicting severity of Parkinson's disease using deep learning. Procedia computer science, 132, pp.1788-1794.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

[13] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," arXiv preprint arXiv:1707.07328, 2017.

[14] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger, "The harms of demographic bias in deep face recognition research," in 2019 International Conference on Biometrics (ICB), 2019, pp. 1–6.

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language

[16] Z. Yang, R. Al-Bahrani, A. C. E. Reid, S. Papanikolaou, S. R. Kalidindi, Wk. Liao, A. Choudhary, and A. Agrawal, "Deep learning based domain knowledge integration for small datasets: Illustrative applications in materials informatics," in 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[17] R. Hoehndorf, N. Queralt-Rosinach et al., "Data science and symbolic ai: Synergies, challenges and opportunities," Data Science, vol. 1, no. 1-2, pp. 27–38, 2017.

[18] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," arXiv preprint arXiv:1710.00794, 2017.

[19] J. Chen, M. Xie, Z. Xing, C. Chen, X. Xu, L. Zhu, and G. Li, "Object detection for graphical user interface: Old fashioned or deep learning or a combination?" in Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ser. ESEC/FSE 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 1202–1214. [Online]. Available: https://doi.org/10.1145/3368089.3409691

● **34% Overall Similarity**

Top sources found in the following databases:

- 31% Internet database
- Crossref database
- 17% Submitted Works database

- 20% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	parkinsonsnewstoday.com Internet	4%
2	weasychair.org Internet	3%
3	doctorpenguin.com Internet	3%
4	coursehero.com Internet	3%
5	Nirali Parekh, Bhavi Dave, Raj Shah, Kriti Srivastava. "Automatic Sleep ... Crossref	2%
6	kdnuggets.com Internet	1%
7	figshare.com Internet	1%
8	res.mdpi.com Internet	<1%

33 University of Strathclyde on 2020-03-27
Submitted works
<1%

34 thesai.org
Internet
<1%

35 Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, Wei We...
Crossref
<1%

36 City University on 2019-09-02
Submitted works
<1%

37 Zhengzhong Liang, Gregory Ditzler. "The Impact of an Adversary in a L...
Crossref
<1%

38 WWW.MDPI.COM
Internet
<1%

39 Fen Xie, Xiaoya Gao, Wanlin Yang, Zihan Chang et al. "Advances in the ...
Crossref
<1%

40 Onur Karaman, Hakan Çakın, Adi Alhudhaif, Kemal Polat. "Robust auto...
Crossref
<1%

41 chato.cl
Internet
<1%

42 "Pattern Recognition. ICPR International Workshops and Challenges", S...
Crossref
<1%

43 biomedical-engineering-online.biomedcentral.com
Internet
<1%

44 docs.h2o.ai
Internet
<1%

45 Hanbin Zhang, Aosen Wang, Dongmei Li, Wenyao Xu. "DeepVoice: A vo... <1%
Crossref

46 peerj.com <1%
Internet

47 lf1.cuni.cz <1%
Internet

48 slideshare.net <1%
Internet

49 Brandenburgische Technische Universität Cottbus on 2020-12-29 <1%
Submitted works

50 Coventry University on 2019-09-19 <1%
Submitted works

51 University of London External System on 2011-04-29 <1%
Submitted works

52 University of West London on 2022-01-23 <1%
Submitted works

53 ieeexplore.ieee.org <1%
Internet

54 researchgate.net <1%
Internet

55 Malaviya National Institute of Technology on 2021-06-20 <1%
Submitted works

56 University of Bradford on 2020-09-18 <1%
Submitted works