

**Hybrid Resampling and XGBoost Prediction Model using patient's
information and drawings as features for Parkinson's Disease
Detection**

**A DISSERTATION
SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF**

**MASTER OF TECHNOLOGY
IN
Signal Processing and Digital Design**

**Submitted by
Aishwarya Keller
(2K19/SPD/02)**

**Under the supervision of
Dr. Anukul Pandey (Asst Prof, ECE Dept.)**



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi -110042
JUNE 2021**

CONTENTS

CANDIDATE’S DECLARATION	iv
CERTIFICATE	v
ACKNOWLEDGEMENT	vi
ABSTRACT	vii
LIST OF FIGURES	viii
LIST OF TABLES	ix
LIST OF ABRREVIATIONS	xi
CHAPTER 1 - INTRODUCTION	1
1.1 Overview	1
1.1.1 Symptoms seen in PD patients	1
1.1.2 Causes of the disease	2
1.1.3 Major factors that pose as risk in PD	3
1.1.4 Complications encountered by the patients	3
1.2 Impact of PD on motor system	4
1.3 Literature Review	6
1.3.1 Handwriting as a tool for detection	8
1.4 Research Gap	9
1.5 Research Objective	10
1.6 Structural organisation of the dissertation	10
CHAPTER 2 – PROPOSED RESAMPLING TECHNIQUE FOR DATA IMBALANCE WITH AGE BASED PARKINSON DISEASE DETECTION	11
2.1 Introduction	11
2.2 Impact of imbalance on a prediction model	12
2.2.1 Class imbalance	12
2.2.2 Solutions for class imbalance	14
2.3 Resampling techniques for data imbalance	15
2.4 Adopted methodology	17
2.5 Reason for choosing SMOTE and ENN over other resampling techniques	18
2.5.1 Why SMOTE ?	19
2.5.2 Why ENN ?	21
2.5.3 Hybridisation : SMOTE+ENN	22
2.6 Impact of age on PD	23
2.7 Proposed methodology	27
2.7.1 Data acquisition	27
2.7.2 Flow diagram	29
2.8 Concluding remarks	34
CHAPTER 3 – IMPACT OF GENDER AND DOMINANT HAND ON PD DETECTION	35
3.1 Introduction	35
3.2 Impact of gender on PD	35
3.2.1 Clinical differences	36
3.2.2 Non-motor symptoms	37
3.2.3 Impact of gender on PD pathophysiology	38
3.3 Impact of handedness on PD	39

3.4 Proposed methodology	40
3.5 Concluding remarks	41
CHAPTER 4 -RESULTS AND DISCUSSION	43
4.1 Introduction	43
4.2 ROC analysis	44
4.3 Advantages	47
4.4 Limitations	48
CHAPTER 5 – CONCLUSION AND FUTURE SCOPE	49
5.1 Conclusion	49
5.2 Scope for future work	50
REFERENCES	51
LIST OF PUBLICATIONS	57

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I **Aishwarya Keller** student of M.Tech (Signal Processing and Digital Design), hereby declare that the project Dissertation titled “**Hybrid Resampling and XGBoost Prediction Model using patient's information and drawings as features for Parkinson's Disease Detection**” which is submitted by me to the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.



Place: Delhi

Date: 31st July 2021

Aishwarya Keller

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Report titled “**Hybrid Resampling and XGBoost Prediction Model using patient’s information and drawings as features for Parkinson’s Disease Detection**” which is submitted by **Aishwarya Keller, 2K19/SPD/02** of Electronics and Communication Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 31st July 2021



Dr. Anukul Pandey
SUPERVISOR

ACKNOWLEDGEMENT

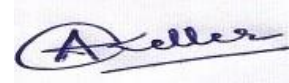
A successful project can never be prepared by the efforts of the person to whom the project is assigned, but it also demands the help and guardianship of people who helped in completion of the project.

I would like to thank all those people who have helped me in this research and inspired me during my study.

With profound sense of gratitude, I thank Dr. Anukul Pandey, my Research Guide, for his encouragement, support, patience and his guidance in this research work.

Furthermore, I would also like to thank the Head of the Department, Electronics and Communication, Prof N. S. Raghava and Prof. S. Indu, who gave me the permission to use all required equipment and the necessary format to complete the report.

I take immense delight in extending my acknowledgement to my family and friends who have helped me throughout this research work.



Aishwarya Keller

ABSTRACT

In the list of most commonly occurring neurodegenerative disorders, Parkinson's disease ranks second while Alzheimer's disease tops the list. It has no definite examination for an exact diagnosis. It has been observed that the handwriting of an individual suffering from Parkinson's disease deteriorates considerably. Therefore, many computer vision and micrography-based methods have been used by researchers to explore handwriting as a detection parameter. Yet, these methods suffer from two major drawbacks, i.e., the prediction model's biasedness due to the imbalance in the data and low rate of classification accuracy. The proposed technique is designed to alleviate prediction bias and low classification accuracy by use of hybrid resampling (Synthetic Minority Oversampling Technique and Wilson's Edited Nearest Neighbours) techniques and Extreme Gradient Boosting (XGBoost). Additionally, there is proof of innate neurological dissimilarities between men and women and the aged and the young. There is also a significant link of the dominant hand of the person and the side of the body where initial manifestation begins. Further, the gender, age, and handedness information have not been utilized for Parkinson's disease detection.

In this research work, a prediction method is developed incorporating age, gender, and dominant hand as features to identify Parkinson's disease. The proposed hybrid resampling and XGBoost method's experimental results yield an accuracy of 98.24% highest so far when age is taken as a parameter along with nine statistical parameters (root mean square, largest value of radius difference between ET and HT, smallest value of radius difference between ET and HT, standard deviation of ET and HT radius difference, mean relative tremor, maximum ET, minimum HT, standard deviation of exam template values, number of instances where the HT and ET radius difference undergoes a change from negative value to positive value or vice versa) achieved on the HandPD dataset. The conventional accuracy is 98.24% (meanders) and 95.37% (spirals) when age is used along with nine statistical parameters extracted from the dataset. It becomes 97.02% (meanders) and 97.12% (spirals) when age, gender and handedness information are utilised. The proposed method results were compared with existing methods, and it is evident that the method outperforms its predecessors.

LIST OF FIGURES

Figure No.	Title	Page No.
2.1	a) Example of class overlapping	13
	b) Example of small disjuncts	13
2.2	Pictorial representation of a two- class imbalanced dataset.	15
2.3	Schematic representation of undersampling and oversampling method.	16
2.4	Block diagram of generalised method for imbalance issue removal.	18
2.5	SMOTE working process.	19
2.6	Python code for SMOTE	20
2.7	SMOTE resampling in sample space.	20
2.8	a) ENN editing with 1-NN classifier with misclassified samples marked by dotted regions.	22
	b) Removal of misclassified samples.	22
2.9	Python code for the hybrid resampling.	23
2.10	SMOTE+ENN resampling in sample space.	23
2.11	Impact of age on SN neurons.	24
2.12	Effects of ageing on the person's body and consequent changes responsible for PD.	25
2.13	Variations seen in processes due to age advancement which led to SN cell death.	26
2.14	Dataset description	28
2.15	Sample form filled by a PD patient	28
2.16	Proposed method's flowchart.	29
2.17	(a) HT and ET for a Spiral image	29
	(b) HT and ET for a Meander image.	29
2.18	Arbitrary points of spiral and meander images. Every vector starts from the central point of meander or spiral and ends up at the point selected randomly.	30
2.19	Columns heads of the .csv file generated after feature extraction (10 features are used for prediction).	31

Figure No.	Title	Page No.
3.1	PD risk factors and symptomatology variations in men and women.	37
3.2	PD pathophysiology variations in both genders.	38
3.3	Columns heads of the .csv file generated after feature extraction (12 features are used for prediction).	41
4.1	(a) ROC comparison of the proposed method with Chi2-Adboost using Meander data with 10 features	46
	(b) ROC comparison of the proposed method with Chi2-Adboost using Meander data with 12 features	46
4.2	(a) ROC comparison of the proposed method with Chi2-Adboost using Spiral data with 10 features	47
	(b) ROC comparison of the proposed method with Chi2-Adboost using Spiral data with 12 features	47

LIST OF TABLES

Table No.	Title	Page No.
1.1	Parkinson's Disease - Signs and Symptoms.	5
2.1	Confusion Matrix for the misclassification cost.	15
2.2	ENN algorithm	22
2.3	Repeated ENN algorithm	22
2.4	Characteristics and mechanisms of cells common to both ageing and PD.	27
2.5	Samples before and after the application of SMOTE on meander and spiral data when 10 features are used for prediction	32
2.6	Samples before and after the application of ENN on meander and spiral data when 10 features are used for prediction	33
3.1	PD influenced by laterality at onset along with age and duration of motor symptoms.	39
3.2.	p-value (level of significance) of link between PD features and laterality	40
3.3	p-value (level of significance) of link between lateralisation of syptom and handedness with PD features.	40
3.4	a) Samples before and after the application of SMOTE on meander and spiral data when 12 features are used for prediction.	41
	b) Samples before and after the applying ENN on meander and spiral data when 12 features are used for prediction.	41
4.1	Comparison of the proposed model with other models. Here IBT: Imbalanced Training, BT: Balanced Training, ACC _b : Balanced Accuracy, MCC: Mathews Correlation Coefficient.	45

LIST OF ABBREVIATIONS

Abbreviation	Full form
PD	- Parkinson's Disease
DNN	- Deep Neural Network
PCA	- Principal Component Analysis
OFS	- Official Feature Set
RF	- Random Forest
FP-CIT SPECT	- Single-Photon Emission Computed Tomography
PPMI	- Parkinson's Progression Markers Initiative
MLP	- Multi-Layer Perceptron
EEG	- Electro Encephalo Graphy
MRI	- Magnetic Resonance Imaging
PET	- Positron Emission Tomography
ALFF	- Amplitude of Low Frequency Fluctuations
fALFF	- fractional Amplitude of Low Frequency Fluctuations
ReHo	- Regional Homogeneity
ROS	- Random Over Sampling
SMOTE	- Synthetic Minority Oversampling Technique
ADASYN	- ADaptive SYNthetic sampling
RUS	- Random Under Sampling
CUS	- Cluster based Under Sampling
SN	- Substantia Nigra
DNA	- Deoxyribo Nucleic Acid
ET	- Exam Template
HT	- Handwriting Template

CHAPTER 1

INTRODUCTION

1.1 Overview

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects motion. The indicators/symptoms have slow manifestation. It can begin with an imperceptible tremor manifestation in merely one hand. The most common indicators are tremors. Although slowing of movement or stiffness is also seen in many patients.

In the preliminary stages of PD, the patients' facial expressions can be hardly noticeable or none at all, arms of the patients may not swing while walking. The speech may be affected and become slurred or soft. PD symptoms get worsen as one's condition progresses over time.

Despite the fact that PD is incurable, medications might help improve the symptoms significantly. Surgery can be done (occasionally) to control certain regions of the brain which might rectify the symptoms.

1.1.1 Symptoms seen in PD patients

PD signs and symptoms can vary from one person to another [1]. In the early stages, the signs/indications may be mild and inconspicuous. The symptoms usually start on one side of the patients' body and worsen on that side, even after they start affecting both sides.

These include:

1. **Tremor** – It generally originates in a limb (hand or fingers). A pill-rolling tremor (i.e., rubbing one's thumb and forefinger back and forth) or trembling hands while they are at rest are also examples of tremors.
2. **Bradykinesia (Slowed movement)** – PD progression affects movement such that day-to-day tasks become difficult and tedious like getting out of a chair. One's steps may become shorter while walking and/or feet might drag while trying to walk.

3. **Rigid muscles** – Patients may feel stiffness in muscles which is painful and restrict the motion range.
4. **Reduced balance and stooped posture** - One might find it difficult to balance his/her body weight while standing and end up being stooped because of PD.
5. **Deprivation of reflexes** - The potential to accomplish unconscious movements, like swinging arms while walking, smiling or blinking may deteriorate.
6. **Changes in speech** – Soft, staggered, quick, or slurred speech while talking are the changes seen in speech patterns. It becomes a monotone instead of the usual inflections.
7. **Changes in writing** – Handwriting of the patient diminishes. It becomes hard to write as the grip on the pen/pencil loosens due to the disease.

1.1.2 Causes of the disease

In PD, the neurons (nerve cells) of the brain perish or gradually break down. Most indicators are a result of the deprivation of neurons that manufacture dopamine, a chemical messenger in the brain. When the level of dopamine decreases, the activity of the brain becomes abnormal. Consequently, the movement of the body is impaired and other indicators of PD become visible.

The following factors play a key role in causing PD:

1. **Genes** – There are certain mutations in the genes that can lead to PD as identified by researchers. These are however less frequent except in cases of family history of PD. Although some gene changes seem to heighten the chances of PD but with a relatively smaller risk of PD for each of these genetic markers.
2. **Triggers due to environmental elements** - Being exposed to specific toxins or environmental components might heighten the chances of later PD, however, the probability is negligible.

Researchers have found certain changes that happen in the brain of PD patients, though it is no clear evidence as to why they occur. These are listed as follows:

- i. **The presence of Lewy bodies** – The microscopic markers of PD are the clumps of certain elements present inside brain cells known as Lewy bodies. Researchers claim that Lewy bodies are significant in the context of the cause of PD.

- ii. **Alpha-synuclein present within Lewy bodies** - There are many elements within Lewy bodies. It is believed that the most significant one is known as alpha-synuclein (a-synuclein) protein. This is a significant factor to focus on among the recent research in PD.

1.1.3 Major factors that pose as risk in PD

PD has the following risk factors:

1. **Age** – PD rarely affects young people. It usually starts in adulthood or late life and the probability rises with age. The disorder generally develops around age of 60 or elderly.
2. **Heredity** – A family history of PD may hike the chances that a person might develop it. Nevertheless, the risk is negligible unless most of the relatives in the family have been afflicted with the disease.
3. **Sex** - Males are more vulnerable to PD as compared to females.
4. **Exposure to toxins** - Being exposed to pesticides and herbicides may heighten the chances of PD.

1.1.4 Complications encountered by the patients

PD is often accompanied by the following issues (which might be curable):

1. **Difficulty in thinking** – Dementia and thinking difficulties might occur in patients. These are usually seen in the later phases of PD and are unresponsive to medications.
2. **Emotional changes and depression** - One might experience depression in the preliminary phases. Therapy for depression can help manoeuvre the other issues of the disease. A person might also experience changes in emotions, such as loss of motivation, fear, and anxiety. However, these symptoms are curable with medications.
3. **Swallowing problems** – It might become difficult to swallow as the disease progresses. This results in the accumulation of saliva causing the person to drool.
4. **Eating and chewing difficulties** - PD starts affecting the muscles of the mouth in the later stages. This makes it difficult for the person to chew food which can end up in gaging and nutrition deficiency.

5. **Sleep disorders/problems** - PD patients often have problems with sleep. This includes staying up the whole night, staying up early, or sleeping throughout the day. Patients may have brisk eye motions or sleep behaviour disruptions which includes acting out your dreams. Medications may aid in alleviating the sleep issues.
6. **Bladder issues** - Inability in controlling urine or having difficulty urinating are the problems faced by PD patients.
7. **Constipation** – PD patients suffer from constipation due to a slower digestive tract.

Besides the above-mentioned signs, patients can experience:

1. **Changes in blood pressure** – A person gets giddy or lightheaded when he/ she stands. This is because there is an instant drop in blood pressure (orthostatic hypotension).
2. **Dysfunction of the sense of smell** – There can be issues with the sense of smell. It might be difficult to recognize a particular odour or distinguish between odours.
3. **Fatigue** - Persons suffering from PD no longer have energy and feel tired, particularly later in the day.
4. **Pain** - PD patients encounter pain, either in certain regions of their bodies or in their entire bodies.
5. **Sexual dysfunction** - Few PD patients exhibit reduced sexual desire or performance.

1.2 Impact of PD on motor system

PD causes a variety of signs and symptoms listed in Table 1.1 below. Symptoms labelled as "General" are very frequent and are seen in several patients at some stage or the other. Universal symptoms include bradykinesia, akinesia, and hypokinesia as they occur in most patients. They are generally seen in the starting stages of the disease. All the three terms i.e., “bradykinesia,” “akinesia,” and “hypokinesia” generally point to a reduction in slower. Walking becomes time-consuming. Additionally, there is a reduction in the recurrence of spontaneous movements like smiling, blinking and shrinking. This results in a subtle or dull looking face. In the Table 1.1 shown below, the term “bradykinesia” is used to

refer to slow walking; "hypokinesia" for a reduction in mobility (and/or strength); and "akinesia" for two conditions - of delayed onset of movement and reduced mobility. Poor management when the balance is disrupted is called "Temporary instability". In a simple situation like stumbling on an uneven road, the patient can cause a fall as he/she is not responsive is not enough to restore the balance. An increase in muscle tone is termed as "Strength". The body shows more resistance than usual when the organ is removed (e.g., through operation). The unusual posture with a shoulder fall and a head bend is called "Stooped posture". "Relaxing vibration" is a recurring movement back and forth of any organ, or head, jaw or trunk, which happens when that portion of the body is motionless. Typical kinds of vibrations include callousness - raising the arm and twisting - finger extension. Some patients with PD have never experienced tremors.

Table 1.1 Parkinson's Disease - Signs and Symptoms.

General	Shortage/Delay of movement (Akinesia)
	Slowness in motion (Bradykinesia)
	Decreased amplitude of movement (Hypokinesia)
	Inability to regain posture control when balance is disturbed
	High resistance is seen in joint motion (passive)
	Stooped posture
	Tremor at rest
Occasionally present	Severe trunk flexion(Camptocormia)
	Reduced arm swing
	Reduced dexterity
	Reduction in range of repetitive movements
	Struggling to get up from a chair
	Trouble performing tasks simultaneously
	Drizzling
	Slurred speech (Dysarthria)
	Problem in swallowing (Dysphagia)
	Unusual body posture (Dystonia)
	Tiredness/Fatigue
	Shortened steps/ getting stuck in a place (Festination)
	Sudden, temporary, short episodes of moving ability (Freezing of gait)
	Decrement in expressions of the face (Hypomimia)
	Decreased volume of voice (Hypophonia)
	Diminishing size of writing (Micrographia)
	Gait shuffling with short steps
	Rapid speech/erratic rhythm(Tachyphemia)

1.3 Literature review

In the practice of medical sciences, the detection of impairments in motion of the patients is based on the neurological investigation. This is done at the time of doctor consultations and home records made by the patients or their nurses. Although, a small duration of inspection may not reveal the data significant enough in the determination of the manifestation of the disease. Moreover, the data of the daily records may be subjective. Many types of research have been conducted for PD detection which has their focus on symptoms like impaired voice, loss of olfactory senses, etc.

The observed symptoms are usually linked to vocal impairment and speech issues. Abdullah Caliskan made use of a deep neural network (DNN) to exploit this fact [2]. Initially, the noise is omitted and then segmented making use of time windows while filtering the speech signal. In the next step, several attributes are extracted from every segment. Afterward, DNN classification is done making use of Stacked Auto Encoders (SAE). DNN was performed on Parkinson's Telemonitoring Voice Dataset taken from UCI ML Repository by Srishti Grover [3]. The data was then classified uniquely into two categories - "severe" and "not severe". The input layer of the neural network consists of 16 units, 3 hidden layers with 10, 20, 10 units in each layer while the output is a 2 neuron layer. The model yielded 81.6667% accuracy. The researchers [4] tried to categorise the PD group based on various sets of features. Feature sets were created based on PCA and OFS. Non-linear features were generated for the dataset taken from Max little University Oxford. Regression tree (Bagging CART), Bagging classification, Random Forest (RPART) were used as nonlinear classifiers for classification with 96.83% as the classification accuracy using a combination of RF and PCA. There is a decrease in the levels of dopamine, which is a liquid produced by neurons (brain cells). This is another pathological symptom of PD. The dopamine levels can be measured by FP-CIT SPECT i.e., a dopamine transporter imaging technique. Therefore, the researchers constructed a deep-learning model which was automated and interpreted using the FP-CIT SPECT image dataset [5] taken from PPMI repository. Images in SPECT are given as input to the 3D convolutional layer initially. Following the progression through $7 \times 7 \times 7$ convolutional filters, max-pooling, and ReLU activation layer along with output layer, 16 3D outputs are produced in the next step. Subsequently, Shu Lih Oh [6] proposed the first ever distinctive automatic

model for detection PD using EEG signals with CNN. It was seen that techniques of non-linear features extraction could be utilized to distinguish a typical (normal) EEG signal from a PD EEG signal. The CNN model comprised of 13 layers with ReLU activation within hidden layers succeeded by SoftMax function in the output layer. This CNN model was 88.25% accurate and 84.71% sensitive. Sleep Behaviour Disorder (RBD), loss of olfactory senses and Rapid Eye Movement (REM) along with the information from PPMI were used in later research [7]. Self-operating diagnostic models built using machine learning methods namely, Boosted Random Forest, MLP and Boosted Logistic Regression having 97.16% accuracy were also executed. The neuro-imaging methods like PET (non-invasive methods), MRI and EEG (invasive methods) have been significant in studying neural activities in the human brain [8]. They make use of ALFF, Functional Connectivity, fALFF and ReHo as features. Additionally, swarm intelligence was used to speed up the performance of existing neural methods [9]. Parallel methods also enhance the performance by training on large datasets [10] [11].

Ornelas Vences et al. [11] developed a fuzzy inference system which was established on the assessors' insight of turning rate formed on 4 biomechanical features derived from sensors placed on lower limbs. Detection and rating bradykinetic gait employing waist-worn sensor was proposed by Sama et al. [12]. MashhadiMalek et al. [13] calculated the link between rigidity and tremor in PD. The most commonly occurring disorder is tremor and thus it is one of the most studied features in literature. The reason behind this is that it is quite challenging to manually capture the subtle tremor features. Rigas et al. [14] studied both actions and resting tremors based on the information gathered using accelerometers. Two, equivalent Hidden Markov Models were used to assess the posture, severity and action. The research done by Abdulhay et al.[15] described the tremor and gait features obtained while deep stimulation of the brain. The data was acquired as a result of sensors set beneath the patient's feet and the forefingers. In addition, the methods of machine learning were fed to the automatic diagnosis system of PD. Methods of deep learning have been employed time and again in PD diagnosis. Kim et al. [16] developed some Convolutional neural networks to differentiate between the gravity of the symptoms.

1.3.1 Handwriting as a tool for detection

Speech and gait analysis for PD detection are simple yet they have some drawbacks. The speech recording needs high-quality recording with negligible background noise while the gait monitoring needs specialized instrumentation with sufficient space to enable walking. The fear of falling while walking in PD restricts the utilisation of gait analysis in PD identification. Unusually small and constricted handwriting is called micrographia and it is proved to be associated with the disease. Handwriting omits the requirement of a distortion/noise-free environment and gait-related problems in measurement. It has also proved to be a potential marker in the diagnosis of PD. It was observed [17] that in-air movements during handwriting significantly affect the detection accuracy for PD. Handwriting exams were used [18] to distinguish the patients from the healthy ones. A study involving 20 patients and 20 healthy ones were conducted. Each person wrote his/her name and address on a page. Later, mean pressure and related velocity parameters were calculated, and the method yielded an accuracy of 97.5%, 95% sensitivity, and a specificity of 100%. The main disadvantage of the methods above was the limited data size. Consequently, their significance was limited. Therefore, data from 37 PD patients and 38 healthy ones were used for research [19]. The data was collected from eight different handwriting tasks. Three models, namely Adaboost (Adb) Ensemble model, Support Vector Machine (SVM) and k-Nearest Neighbors (KNN), were developed, and the classification 81.3% accuracy was achieved. Recently, Pereira gathered data from 18 healthy persons and 37 PD patients [20]. This data had spiral drawings that were used to differentiate between healthy and PD patients using Naïve Bayes (NB), SVM and Optimum Path Forest (OPF). The optimum accuracy of 78.9% was achieved with the NB model. Their next research developed the infamous HandPD dataset taken from 74 patients and 18 healthy ones [21]. This dataset contains spiral and meander drawings. The drawings' features were extracted using NB, SVM, and OPF, and a classification accuracy of 67% was obtained. There are two major drawbacks of working with the HandPD dataset – biasedness of the models and low PD detection rate. It comprises 19.56% data of healthy subjects and 80.44% data of PD patients. Thus, the models that train on this data are biased for the majority class as the minority class instances occur rarely.

Consequently, test samples of minority classes are usually misclassified. The model is highly sensitive to the majority class (PD patients in this case) and shows a low specificity rate (for healthy subjects of the minority class). Recently, a model was designed [22] which tried to solve these two issues and developed four separate models, namely, Linear Discriminant Analysis (LDA), Decision Tree (DT), Gaussian Naïve Bayes (GNB), and KNN, to show the impact of biasedness. They employed random undersampling to overcome biasedness and a cascade of Chi2-Adaboost ensemble models to improve accuracy. However, the accuracy still needed much improvement. In literature, many methods were employed to deal with data imbalance [23], [24], and [25]. However, the accuracy score was not significant.

It is seen that PD detection is associated with the gender and age [26] of the person. The dominant hand is also significant in the context of prediction as to the side of the initial manifestation; that is, the dominant side is impacted first in the majority of both right and left-handed patients [27].

1.4 Research Gap

In the methods listed under the handwriting-based detection of PD, there are two issues: the imbalanced nature of data and low classification accuracy. These two drawbacks have the following consequences:

1. When one trains machine learning models based on imbalanced data, the models are biased as they neglect the minority class and favour the majority one. This happens because minority class instances are a rare occurrence which also makes their predictions infrequent or undiscovered. As a result, the minority class test instances are wrongly interpreted more often as compared to the majority ones. Therefore, while handling binary classification cases (like in the case of HandPD data), the model is highly sensitive (if the patient is of the majority class) and depicts low specificity (when healthy subjects are in minority). This is a clear indication of biasedness for the majority class.

Remedy of imbalanced data: The general method used to remove the issue of imbalance in the data is resampling. This can be done in two ways:

- a) Oversampling – The class having a lesser number of samples has its sample replicated to balance the size of every class in the training data.

b) Undersampling – The class having a majority number of samples has some of its samples omitted to balance the size of each class in the training data.

Remedy for low accuracy rate: An ensemble model is used for prediction or classification. The most significant advantage of using ensembles is to enhance the average prediction performance over any contributing member in the ensemble.

1.5 Research Objective

This dissertation has the following objectives:

1. To explore the impact of model bias (in prediction) and
 - i. Age as a feature in PD detection using Spiral and Meander images from the HandPD dataset
 - ii. Gender and dominant hand of the patient as parameter for PD

and propose a hybrid resampling with XGBoost as the prediction model.

2. To test the validation of the model proposed here in terms of accuracy, sensitivity, specificity, F-score, Mathew Correlation Coefficient and ROC curve.

1.6 Structural organisation of the dissertation

The remaining part of the dissertation is arranged in the following manner: Chapter 2, describes the problem of class imbalance in detail. It also gives the solutions developed over the years through research to solve the imbalance issue of datasets. Further, it highlights the importance of age as the factor in influencing the spread, symptoms, and severity of the disease. It also covers the aspects like dataset used and step by step execution of the proposed method. Chapter 3 describes the effects of gender and laterality (dominant side) on the disease manifestation. Chapter 4 includes the performance parameters and validation of the method. It also consists of that enlists the work's advantages and limitations, while Chapter 5 is the conclusion and future aspects.

CHAPTER 2

PROPOSED RESAMPLING TECHNIQUE FOR DATA IMBALANCE WITH AGE BASED PARKINSON'S DISEASE DETECTION

2.1 Introduction

This chapter is focused on three major factors:

1. The issue of class imbalance.
2. The importance of age as a parameter for PD detection.
3. Introduction and step by step execution of the proposed methodology.

Class imbalance is encountered where datasets have unequal distribution in between the classes. If this is neglected during the preprocessing tasks, the prediction model built based on learning from an imbalanced dataset is biased. This biasedness is termed as the inclination of the prediction model shown towards the majority class samples. This is due to the fact the since there is a higher sample count of the majority class, the model learns better about this class in comparison to that of the minority one. This is a serious hindrance when the studies aim towards learning crucial information about the minority class. Resampling can aid in alleviating imbalance. Moreover, a classifier ensemble can significantly improve the accuracy of the prediction which was earlier degraded due to imbalance.

The latter half of the chapter emphasizes the role of age advancement in PD. There are various changes in the person's body as he/she advances in age. Many such changes contribute to symptoms which can be key factors of PD manifestation. All the variations in the brain and cells are described in detail and help understand how the disorder can be well recognized in the context of these indicators.

Data acquisition/mining describes in detail about the dataset that was used in the research. It also consists of the distribution of the dataset. The flowchart represents the step by step execution of the prediction model. It includes the crucial stage of feature extraction and how the nine statistical features namely, root mean square, largest HT and ET radius difference, smallest HT and ET radius difference, HT and ET's radius difference's standard deviation, mean relative tremor, minimum HT, maximum ET, exam template values' standard deviation, total instances where the HT and ET radius

difference undergoes a change from negative value to positive value or vice versa were calculated. It also gives the implementation of resampling and boosting.

2.2 Impact of imbalance on a prediction model

This section illustrates the details of what exactly is class imbalance issue, its remedies and why the SMOTE and ENN techniques are used for the research.

2.2.1 Class Imbalance

Imbalanced classification is an issue with the datasets having skewed spread of data samples. It has the attributes listed below:

1. **Class overlapping:** Data instances of various classes overlap (Fig.2.1 a)). In such scenarios, the classifiers have difficulty in accurately differentiating between various classes. This leads to the misclassification of samples associated to the minority class into the majority class.
2. **Small sample size:** Gathering a sufficient amount of details for imbalanced datasets is quite difficult in reality. A remedy of this issue is to counterbalance the ratios of imbalance in the datasets to diminish the misclassification error.
3. **Small disjuncts:** Minority class data instances are scattered in many feature spaces, as seen in Fig. 2.1 b). This increases complexity in the classification stage.

There is a relevant distinction between the size of samples of two separate classes (large ratio of imbalance). The classifiers may take few data instances of the minority class as aberrations which results in a large misclassification rate of error for the minority class. With the increase in the magnitude of the data, the impact of the class imbalance issue becomes larger.

Class imbalance is encountered while handling real-world datasets, in which one class (i.e., the minority class) consists of a smaller number of instances and the other (i.e., the majority class) consists of a greater number of instances. It is a Herculean task to construct an optimal model using conventional data mining and machine learning methods without preprocessing step. The main function of preprocessing is to balance the datasets.

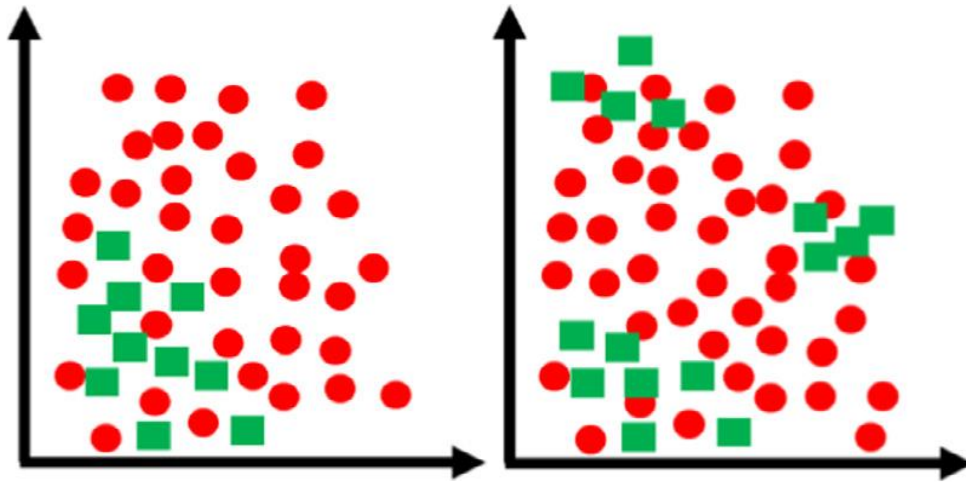


Fig. 2.1 a) Example of class overlapping b) Example of small disjuncts

Additionally, the majority of data mining algorithms find it difficult to identify rare objects among the common ones [28][29][30]. This issue is seen in applications like financial troubles, forecasting medical diagnosis and filtering e-mail [31]. Moreover, the focus of the data accumulation task is often the rare class.

If the class imbalance problem is neglected, the constructed models or learning algorithms can become swamped with the majority class and consequently neglect the minority class. Let us consider a two-class dataset with the ratio of imbalance as 99%, where the majority class makes up 99% of the data set while the minority class has only 1%. To reduce the error rate to a minimum, the learning method categorises the entire sample set into the majority class, which has the rate of error as 1%. In such a situation, the minority class instances are of prime importance and must be recognized as misclassified [32]. A practical class imbalance problem is seen in bankruptcy prediction [33][34]. To be specific, the minority class is described by the number of bankruptcy cases whereas the majority class is made of non-bankruptcy cases. When a prediction model misclassifies a bankruptcy case into the non-bankruptcy class, it is termed as type I error rate. This is more critical in comparison to the average classification accuracy rate as higher type I error rates can probably raise bad debts for organizations.

The methods to solve this imbalance issue are classified as follows:

1. Algorithmic-level techniques
2. Data-level techniques
3. Cost-sensitive techniques
4. Classifier ensembles

The data-level techniques concentrate on preprocessing the imbalanced datasets prior to the classifier development. In this way, data preprocessing and process of training the classifier can be uncorrelated and thus performed independently. Galar et al. [31], did a compared multiple famous approaches along with an amalgamation of data preprocessing techniques with classifier ensembles. Data preprocessing methods form their basis in resampling techniques. It is done before the model enters the training stage.

2.2.2 Solutions for class imbalance

There are three ways to remove the class imbalance issue on the basis of data, algorithms, and cost sensitivity [35]. Classifier ensembles are also used to alleviate the imbalance issue [36] [31].

a. Data-level solutions: These include balancing/preprocessing the acquired imbalanced training dataset by using either of the two resampling techniques: undersampling or oversampling. Undersampling decreases the instances of majority class, whereas the oversampling extends the instances in the minority class. This aids the classifier training processes and sampling to become independent of each other. Thus, various sampling techniques can be amalgamated with classifiers. It was deduced by Batista et al. [28] that the methods of sampling method can efficiently elucidate the class imbalance issue and enhance the performance of the classifier. Galar et al. [31] claimed that sampling closes in and various classifier ensemble combinations have been contemplated for the class imbalance issue.

b. Algorithm-level solutions: These include developing new or transforming existing algorithms to manoeuvre imbalanced datasets. The most common ones include the threshold technique and the one-class learning technique. The former method involves initializing various thresholds for several classes while the classifier is in the learning stage [30] while the latter technique necessitates training the classifier using a training set that consists of only one particular class [37][38]. Other techniques include clustering personalized modelling, clustering in neuro-fuzzy systems, clustering through quantum-inspired evolutionary methods, developing clustering of dynamic data in skewer neural networks, and clustering through quantum-inspired progressive algorithms also deal with imbalanced data [39][40][41][42].

c. Cost-sensitive solutions: Their main aim is to define various cost of misclassification of classifiers for various class labels. Following that, a confusion matrix is developed for the computation of the cost of misclassification as seen in Table. 2.1 below. Here, the correct categorisation cost is 0; otherwise, if the instances of data whose actual class is j is misclassified into the i class, λ_{ij} is its cost of misclassification . Hence, the risk of a_i to reduce the cost of misclassification [43] is evaluated as shown below in equation 2.1.

$$R(a_i|x) = \sum_i \lambda_{ij} P(v_j|x) \quad (2.1)$$

Table 2.1 Confusion Matrix for the misclassification cost.

		Prediction	
		Class i	Class j
True	Class i	0	λ_{ij}
	Class j	λ_{ji}	0

2.3 Resampling techniques for data imbalance

Class imbalance is witnessed when there is a differing distribution or spread of classes in a dataset i.e., the number of data points in the majority/ negative class very large compared to that of the positive class (minority class).

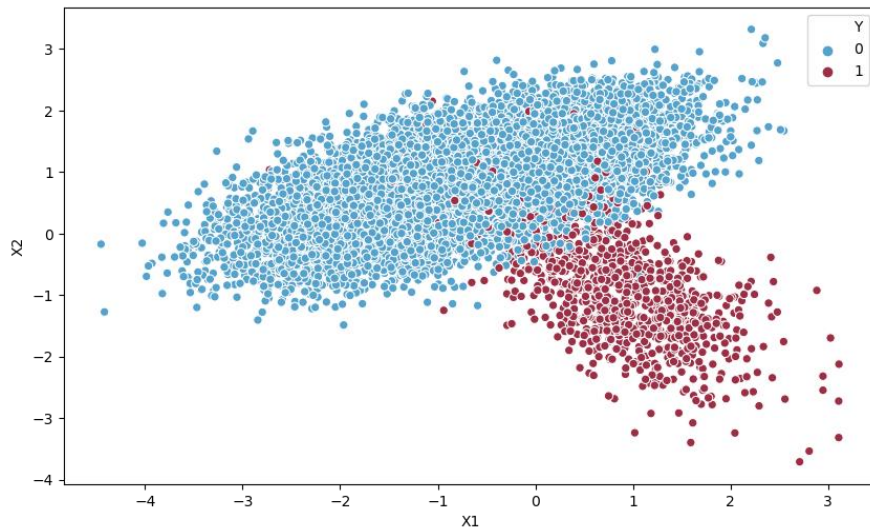


Fig. 2.2 Pictorial representation of a two- class imbalanced dataset.

Usually, the positive or minority class is the primary focus and thus the model needs to achieve optimum outputs for this class. When the imbalanced issue is not resolved during preprocessing steps, then it significantly reduce the efficiency of the classifier. The predictions mainly correspond to the majority class and treat the minority class features as distortions in the data and neglect them. This leads to a large bias in the model.

Resampling is a significant preprocessing task [28] used to solve the imbalanced issue. This is achieved by adding samples to the minority class or dropping samples from the majority class. The former process is termed oversampling while the latter is termed undersampling. The schematic representation of the aforementioned methods is given in Fig. 2.3. Three oversampling and three undersampling techniques are discussed briefly below for the empirical study. ROS (Random Over Sampling), SMOTE (Synthetic Minority Oversampling TEchnique), and ADASYN (ADaptive SYNthetic sampling) are oversampling techniques. RUS (Random Under Sampling), CUS (Cluster-based Under Sampling), and Near Miss are the under-sampling techniques chosen for discussion.

a. **Random Over Sampling**

The simplest preprocessing technique is random oversampling. Samples of the minority class are randomly selected and replicated, such that the sample count of the minority/rare class becomes equal to that of the majority class.

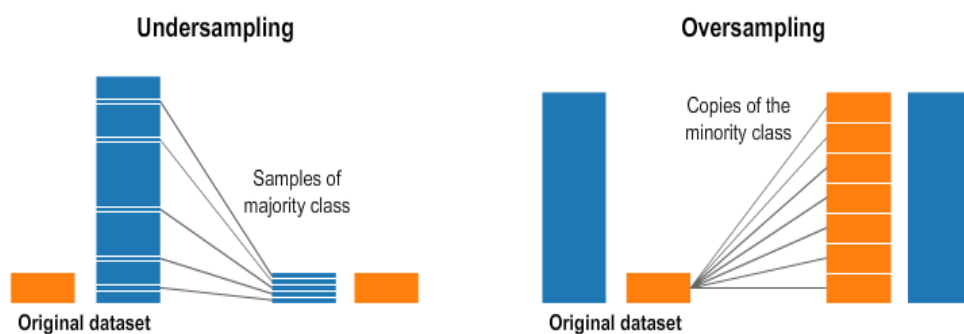


Fig. 2.3 Schematic representation of undersampling and oversampling method.

However, the repetition of the same data increases the chances of overfitting. Thus, synthetic samples generation of minority class is done to avoid this issue.

b. SMOTE

SMOTE was first developed by Chawla et.al [44]. It helps generate synthetic samples rather than duplicating the same data. This is done by considering the k-nearest neighbors of the smaller class. Here 'k' is based on the count of new instances that are required to be formed. The difference between the neighboring sample and the featured vector is computed. In the next step, the difference is multiplied with an arbitrary number between 0 and 1. This new vector value is added to the selected featured vector to generate a new instance of the minority class.

c. ADASYN

It is an upgraded version of SMOTE. It makes use of a weighted distribution for different minority class samples depends on the toughness to learn them. More synthetic samples are generated for complicated minority samples than the minority samples which are less complex to learn. This adopted strategy reduces the bias due to the class imbalance [45].

d. RUS

It selects the samples of the majority class arbitrarily. The chosen instances are then omitted from the actual set of data. Nevertheless, its main disadvantage is that it may remove certain significant data in the context of the learning process.

e. CUS

Lin et.al [24] developed a clustering-based undersampling method to resolve the issues with RUS. In this method, the K-means algorithm aids in clustering the majority class instances. A cluster centroid represents each cluster. This centroid adds to the newly generated instance list of the majority class.

f. NearMiss

This method chooses the majority class samples which are nearer to the minority class samples i.e., it retains majority class samples whose distances are short to the minority class samples [46].

2.4 Adopted methodology

The steps for the method used for the prediction of class labels using imbalanced datasets is shown below:

i) Selecting the imbalanced datasets.

- ii) Pre-processing the datasets to convert string valued attributes to integer type Panda's data frame facility.
- iii) Splitting the dataset into the Training set and Test set in the ratio.
- iv) Resampling the training set.
- v) Learning and fitting the model using an ensemble of classifiers like AdaBoost, XGBoost, etc.
- vi) Evaluating the classifier's performance after sampling.

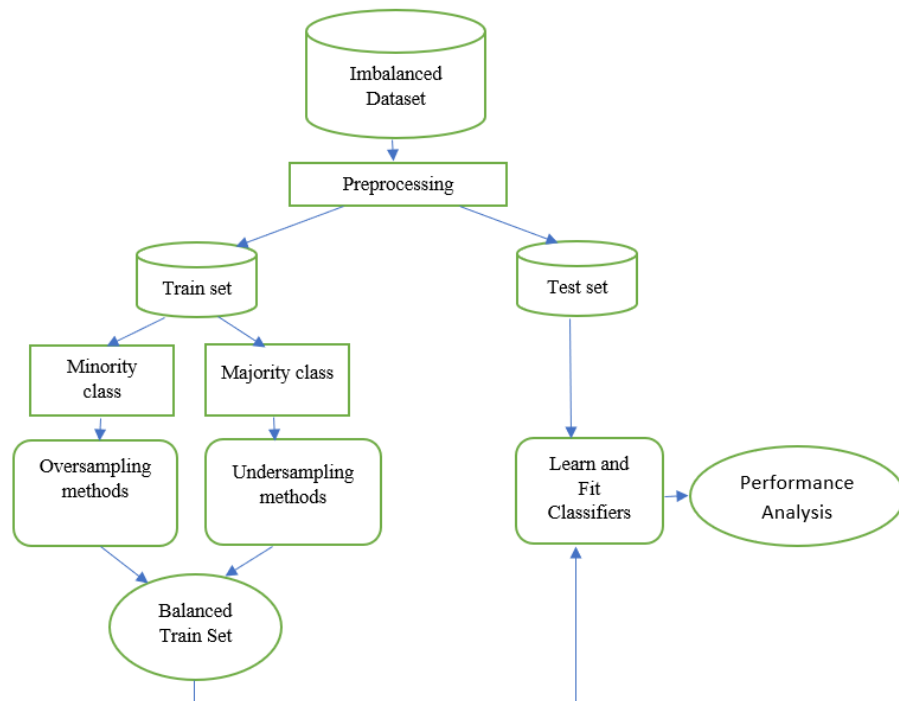


Fig. 2.4 Block diagram of generalised methodology for imbalance issue removal.

2.5 Reason of choosing SMOTE and ENN over other resampling techniques

This section describes SMOTE and ENN in detail. It also gives the reason behind the preference of these two techniques over other resampling techniques for removing imbalance. Hybrid resampling using a combination of these two is also discussed briefly.

2.5.1 Why SMOTE?

SMOTE generates synthetic instances for minority classes making use of information from the dataset and not by random instance duplication. Synthetic samples are created using a linear combination of two similar samples of the minority class, in which one sample is arbitrarily selected among the minority class nearest neighbors of another. The basis of sample generation is the operations in feature space rather than data space.

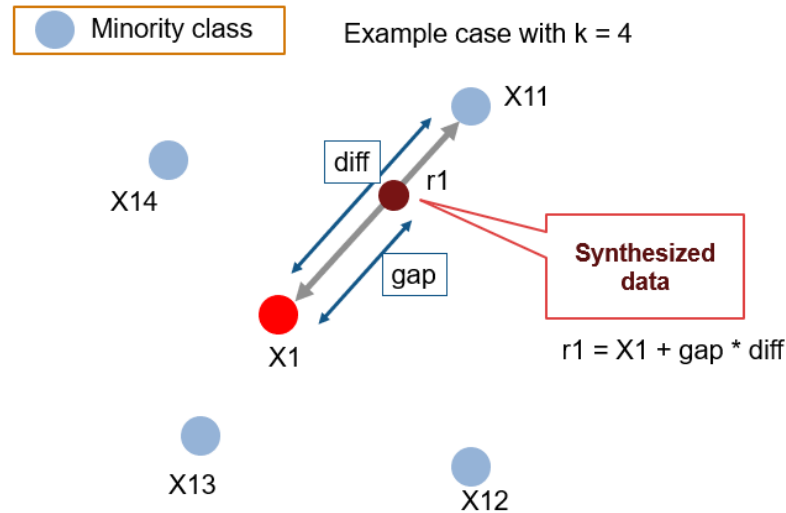


Fig. 2.5 SMOTE working process.

In the first step, the total count of oversampling instances N is initialized. It is usually chosen to make the ratio of binary class distribution as 1:1. However, this ratio can be changed as per the requirement. Subsequently, the iteration begins with choosing a positive class instance arbitrarily being the first step of the process. Afterward, KNN's (the default value is 5) of that particular instance are obtained. In the last step, N number of these K instances are chosen to interpolate new synthetic ones. Any distance metric which computes the difference in distance between the feature vector and its neighbors is used for this purpose. The difference is multiplied by any arbitrary value in $(0,1]$ and added to the preceding feature vector. This process is described in Fig. 2.5.

The results of this research reveal that the SMOTE approach can enhance the accuracy of classifiers for a minority class as it is an up-gradation of over-sampling.

```

from imblearn.over_sampling import SMOTE

counter = Counter(y_train)
print('Before', counter)
# oversampling the train dataset using SMOTE
smt = SMOTE()
#X_train, y_train = smt.fit_resample(X_train, y_train)
X_train_sm, y_train_sm = smt.fit_resample(X_train, y_train)

counter = Counter(y_train_sm)
print('After', counter)

```

```

Before Counter({0: 18497, 1: 4208})
After Counter({0: 18497, 1: 18497})

```

Fig. 2.6 Python code for SMOTE.

SMOTE, in amalgamation with under-sampling, works more efficiently than the conventional under-sampling process. When there is mere replication of the minority class sample (oversampling), the decision region that yeilds a classification decision for the minority class can be diminished and concise. This is due to the repetition of minority samples in that area. This is just the reverse of the desired result. SMOTE causes the classifier to develop wider decision regions that contain nearby minority class samples within them. This is the reason why SMOTE is more efficient than Naive Bayes and Ripper's loss ratio. It yields more relevant minority class instances to learn and interpret from , thus permitting a learner to sculpt broader decision regions, yielding greater description of the minority class.

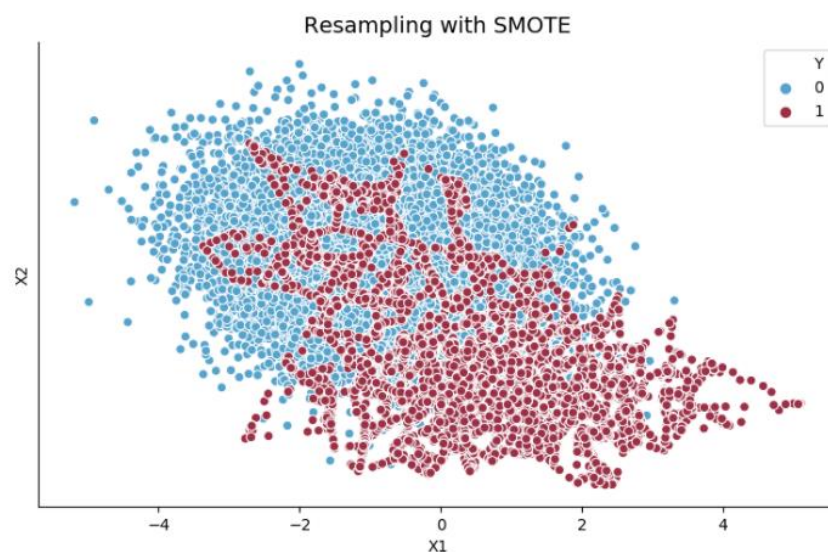


Fig. 2.7 SMOTE resampling in sample space.

2.5.2 Why ENN?

KNN is a relevant algorithm used for classification purposes. It finds k nearest neighbors of every targeted instance based on a dissimilarity estimate. Afterward, a judgement is made based on the recognised categorisation of these neighbors. This is done by designating the category of the highest voted class amidst these k neighbors. When the amount of training data n reaches infinity tends to the optimal Bayes error as $k/n \rightarrow 0$, this is termed as asymptotic classification error of k -NN. When $k = 1$, the error is circumscribed by approximately double that of the Bayes error [47]. k -NN is a lazy learning algorithm. Here, the function is locally estimated and the entire calculation is deferred until classification. Thus, k -NN learns more easily in comparison with other classifiers (which need to be trained) as it only requires reading in the training set without further processing. Yet, the high degree of local sensitivity makes k -NN highly vulnerable to noisy samples in the training set [48]. The samples with errors in the input vector or those not representatives of typical cases or with errors in the output class are termed as noisy instances. Such samples can negatively impact the ability of generalization of k -NN [49]. To decrease the effect of noisy instances in k -NN either an indirect or direct approach is used. While the direct approach tries to omit the noisy samples, the indirect method tries to alleviate the disadvantage of noisy instances without omitting them. The indirect methods might involve increasing the nearest neighbors' count or using distance-weighted voting. One of the most relevant of these methods is nearest neighbor editing [49]. It removes error-prone samples from the training set and avoids chances of overlapping amidst the classes. ENN omits all the misclassified samples (by k -NN rule) from the training set. Fig. 9 illustrates the impact of ENN. The hollow circles and the solid circles in the figure represent samples that are associated to two separate classes. Fig 2.8 a) depicts a hypothesis training set in which misclassified samples using the 1-NN rule are marked with dotted circles around them. Fig 2.8 b) depicts the decreased training set post ENN application.

ENN is based on the aberrations can be removed efficiently and probable overlapping amidst classes from a given training set can also be avoided. This makes the training of the corresponding classifier easy. Penrod and Wagner [12] claim that the accuracy of the ENN classifier closes in on the Bayes error as the number of samples tends to infinity.

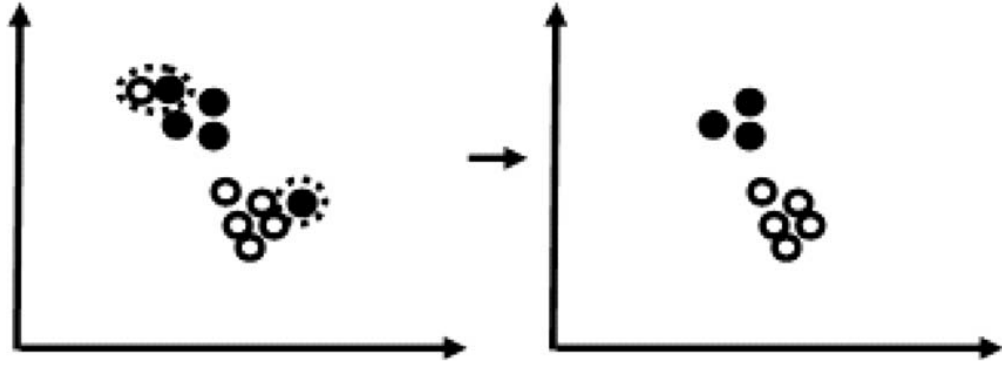


Fig. 2.8 a) ENN editing with 1-NN classifier with misclassified samples marked by dotted regions. b) Removal of misclassified samples.

The steps of this ENN algorithm are listed in Table 2.2 below.

Table 2.2: ENN algorithm

<ol style="list-style-type: none"> 1. Let $T_e = T // T$ is the original training set, and T_e is the edited set 2. For each $x_i \in T_e$, do: Discard x_i from T_e if it is misclassified using the k-NN rule with prototypes in $T_e / \{x_i\}$
--

Table 2.3: Repeated ENN algorithm

<ol style="list-style-type: none"> 1. Let $T_e = T // T$ is the original training set, and T_e is the edited set <p>REPEAT</p> <ol style="list-style-type: none"> 2. At iteration t, for each $x_i \in T_e^t$, do Discard x_i from T_e^t if it is misclassified using the k-NN rule with prototypes in $T_e^t / \{x_i\}$; <p>UNTIL $T_e^t = T_e^{t-1} // T_e^t$ and T_e^{t-1} denote the edited data set of T at iteration t and $t - 1$ respectively</p>

ENN removes samples based on the voting of their nearest neighbors. A misclassified/noisy sample is removed when its label varies from that acquired by the voting of its nearest neighbors (called a voting label). This data editing is efficient only when the real label of a sample is the same as its voting label. In particular, the performance of data editing seems to improve when the training samples' labels are anticipated accurately by their nearest neighbors.

2.5.3 Hybridization: SMOTE + ENN

SMOTE + ENN is a hybrid resampling method in which more samples are omitted from the sample space. Integrating the ENN method with oversampling done using SMOTE aids in extensive data cleaning. In this process, the nearest neighbors' samples

from both classes are removed when misclassified. This yields in a clearer and concise class separation.

```
from imblearn.combine import SMOTEENN

counter = Counter(y_train)
print('Before', counter)
# oversampling the train dataset using SMOTE + ENN
smenn = SMOTEENN()
X_train_smenn, y_train_smenn = smenn.fit_resample(X_train, y_train)

counter = Counter(y_train_smenn)
print('After', counter)
```

```
Before Counter({0: 18497, 1: 4208})
After Counter({1: 14818, 0: 8961})
```

Fig. 2.9 Python code for the hybrid resampling.

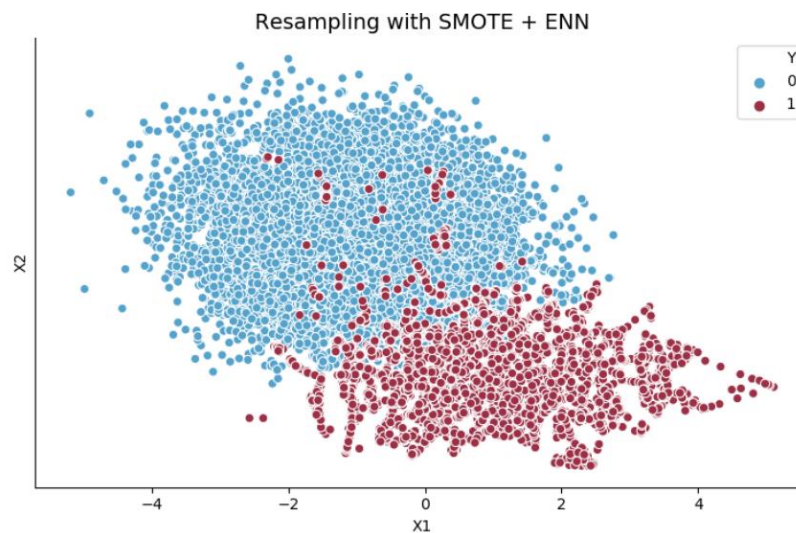


Fig. 2.10 SMOTE+ENN resampling in sample space.

2.6 Impact of age on PD

Aging is the greatest risk factor in the context of PD. PD is accompanied by a complex array of symptoms. A significant region of the brain (the substantia nigra (SN)) is impacted by severe cell loss in PD. This causes the associated motor symptoms due to the disorder. The dopaminergic neurons of the pars compacta within the SN perish. Additionally, this brain region changes pathologically more with aging as compared to other regions as seen in Fig. 2.12 below. Buchman et al. [51] studied over 750

elderly individuals with a mean age of 88.5 years. These people did not have clinically defined PD. However, the results showed that approximately one-third of the total showed mild to severe neuronal loss within the SN. 10% showed Lewy body pathology.

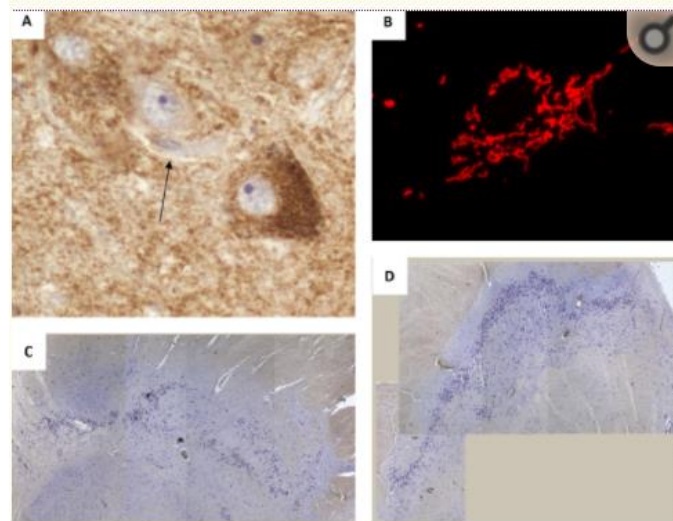


Fig. 2.11 Impact of age on SN neurons.

Fig. 2.11 shows the changes seen in SN neurons with advancing age. The portion labeled as (A) depicts the increase in cells with dysfunctional mitochondria accompanied by the loss of major mitochondrial proteins like complex I subunits are shown by the arrow. Neuronal survival image marked as (B) is significantly affected by changes in network dynamics and the potential of the mitochondrial membrane. (B) describes the mitochondrial network of a healthy neuron within the culture. This network's segregation is following variations seen in the potential of mitochondrial membrane and before the deterioration through mitophagy. The image labeled as (C) displays SN of a 69-year-old patient while image (D) displays SN of a 53-year-old. The loss of pigmented cells (even with a low magnification) is a representation of neurons lost in the SN of the 69-year-old PD patient.

The reason for the death of SN cells with age advancement can give significant insight into the cause of cells lost due to PD. The ventral tegmental area (VTA), pedunculopontine nucleus (PPN), the dorsal motor nucleus of the vagus nerve (DMV), and the locus coeruleus (LC) are several other regions of the brain which are also affected in PD. These are like the neurons of the SN which may emphasize not only their susceptibility in PD but also the processes important factors for the loss of SN neurons. Hence, vulnerable neuronal populations display few characteristics which are

common to those of SN neurons. This reflects that the unique combination of such features within the SN has an important role in PD development.

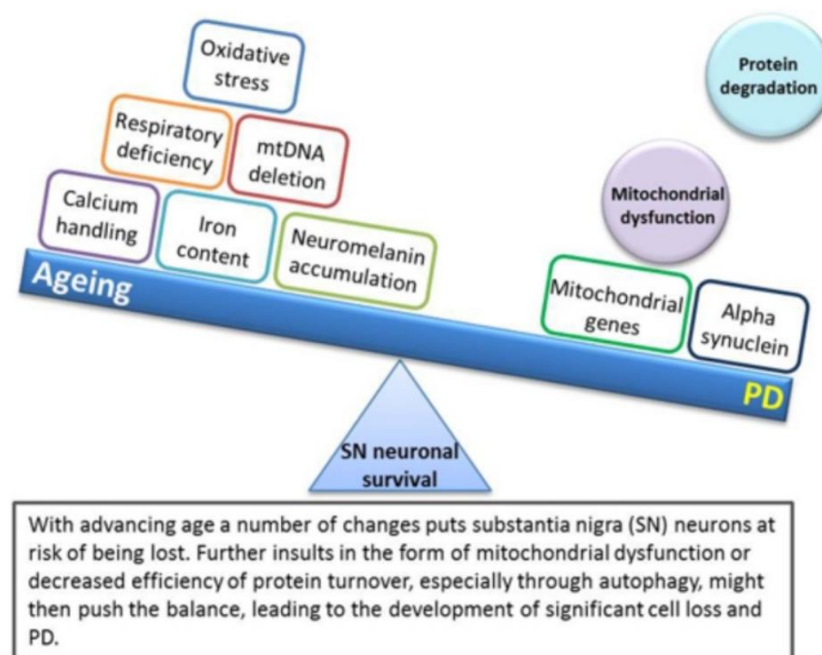


Fig. 2.12 Effects of ageing on the person's body and consequent changes responsible for PD.

Several processes are vital for the function of SN neurons like protein degradation decline, dopamine metabolism, and wild-type mitochondrial DNA copy number are seen in patients as they advance in age. An ample amount of reactive oxygen species is generated by dopamine metabolism that will influence many different processes within the neurons. Reduction in wild-type mtDNA copy number will result in a lower ATP production and a reduction in efficient protein degradation will affect the functioning of neurons. The accumulation of neuromelanin along with the ability of neurons and mitochondria to manoeuvre calcium and the levels of iron within these neurons will also be influenced. Mitochondrial complex I and IV deficiencies and aggregating alpha-synuclein cause the loss of vulnerable neurons, once this cell loss reaches a certain threshold, the symptoms of PD become visible.

The changes in cells due to progressive age involves oxidative damage, accumulation of neuromelanin and that of mitochondrial DNA defects (Fig.2.13). Therefore, there is rise in susceptibility of SN neurons. This further leads to mitochondrial dysfunction and toxic alpha-synuclein causing cell death. All these processes acting simultaneously cause neuron loss inside the brain. Several such processes are sufficient to cause death

of SN in PD. Thus, developing new information about PD could be accelerated if the research on aging and PD were planned together, and the perspective provided by gerontology gains relevance in this field [13].

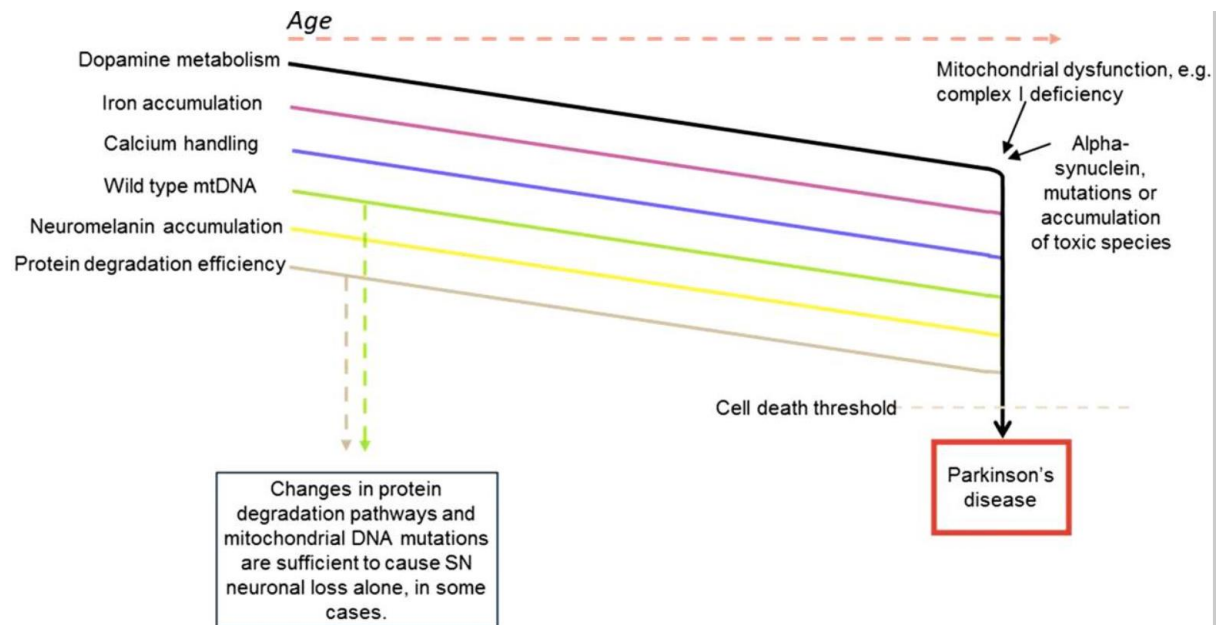


Fig. 2.13 Variations seen in processes due to age advancement which led to SN cell death.

Table 2.4 Characteristics and mechanisms of cells common to both ageing and PD.

Cellular changes				Etiology			
		PD	Age			PD	Age
SNn loss	Melanin+ cells	↓	↓	Pathogeny	Multifactorial	YES	YES
	TH+ cells	↓	↓	Oxidative stress	ETC disruption	YES	YES
	DD+ cells	↓	↓		ROS/RNS	↑	↑
	DAT+ cells	↓	↓		SOD activity	↓	↓
	Lat/Postdistribution	YES	YES		GPx activity	↓	↓
nsDAn adaptation	DA turnover	↑	↑		Lipid peroxidation	↑	↑
	DA receptors	↑	↑		Protein damage	↑	↑
	DAT activity	↓	↓	Mitochondrial dysfunction	mtDNA mutations	↑	↑
	DAn differentiation	↓	↓		mtDNA deletions	↑	↑
PPN	Ach cells	↓	↓		Mitophagy	↓	↓
Locus coeruleus	NA cells	↓	↓		Transport	↓	↓
Thalamus	GLU cells	↓	↓	Genes/Proteins	Polygenic low penetrance	YES	YES
Astrocytes	Astrogliosis	YES	?		α-synuclein aggregation	YES	YES
	Glutation release	↓	?		Parkin activity	↓	↓
	Trophic factor release	↓	↓		UCH-L1 activity	↓	↓
	Cytokines, TNFα release	↑	↑		PINK1 activity	↓	↓
	GLU release	↑	?		DJ-1 activity	↓	↓
	Gliogenesis	↓	↓	Silent toxics	MPTP/paraquat ...	↓	?
Microglia	μgliosis	YES	YES	Premotor disturbances	Olfactory dysfunctions	YES	YES
	IL-6/TNFα release	↑	↑		Sleep fragmentation	YES	YES
Stem cells	SVZ proliferation	↓	↓		Constipation	YES	YES
					Mood disorders	YES	YES

2.7 Proposed Methodology

This section gives the step by step process explanation of the method used.

2.7.1 Data acquisition

The experiments were done using the HandPD dataset accessible from the internet publicly. The dataset was introduced by Pereira et al. [14]. It consists of drawings made in a form with a prototype for guideline uses (Fig. 2.15), showing tasks specially

orchestrated to evaluate distinctive symptoms of PD patients. Fig. 2.14 below shows the distribution of the dataset.

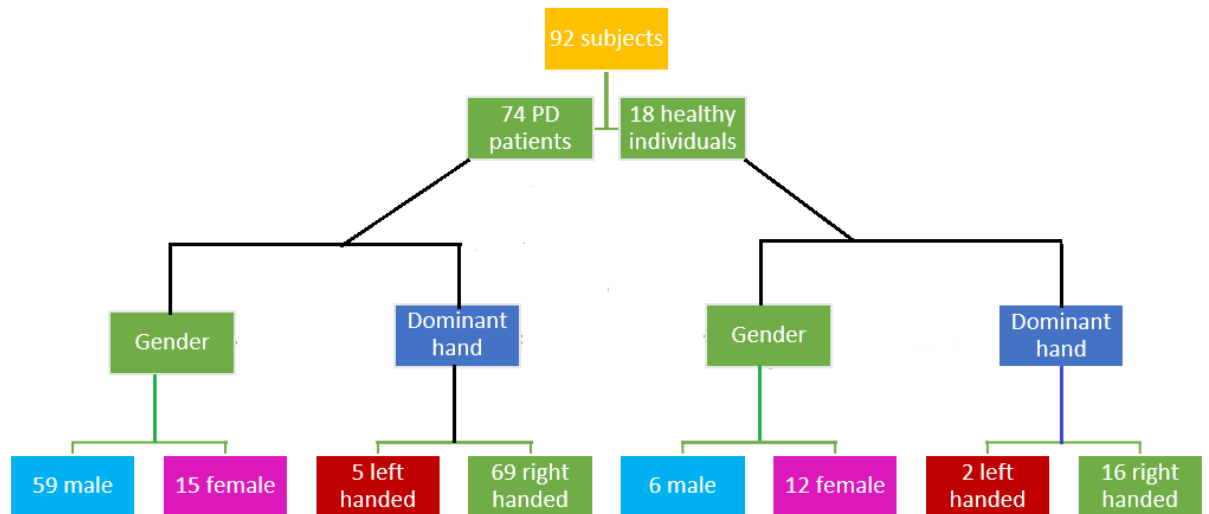


Fig. 2.14 Dataset description

Only 19.56% of the entire dataset comprises healthy individuals and while 80.44% is of PD patients. Thus, this dataset is highly imbalanced. Each person performed six different tasks shown in Fig. 2.15 below. Among these tasks, meander and spiral drawings were recorded. Each subject drew four spirals and four meanders. Thus, there were $92 \times 4 = 736$ drawings from which 368 were spirals and 368 meanders.

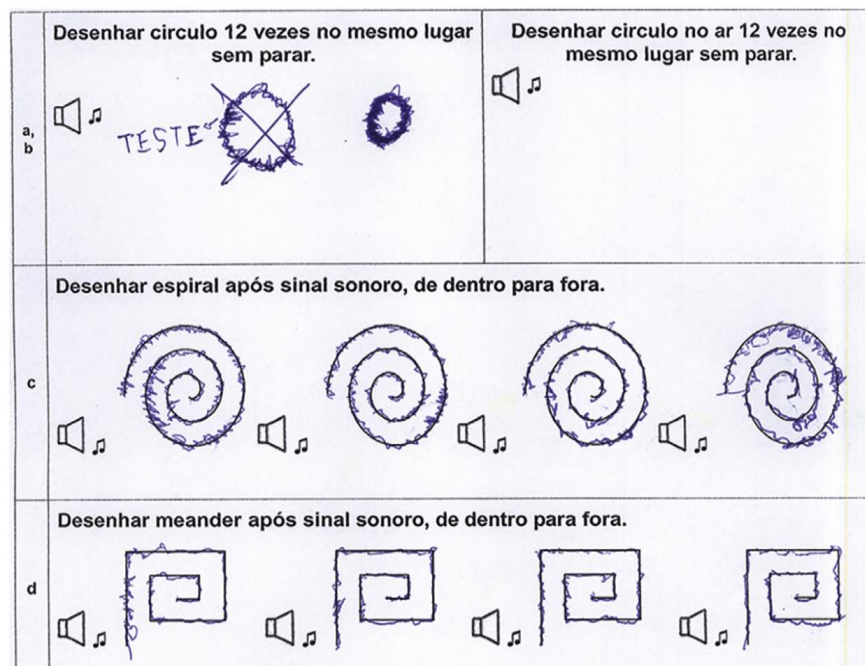


Fig. 2.15 Sample form filled by a PD patient

2.7.2 Flow Diagram

The proposed method's flowchart depicting its various stages is seen in Fig. 2.16 below.

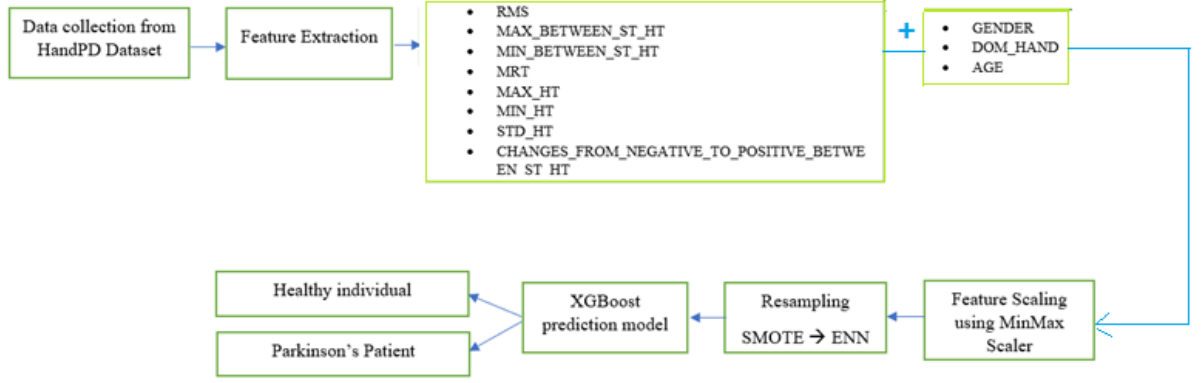


Fig. 2.16 Proposed method's flowchart.

- a.** In the first step, the HandPD dataset is retrieved. The dataset was collected from 92 people at the Sao Paulo State University's Faculty of Medicine of Botucatu, Brazil.
- b.** In the next step, feature extraction is done. Handwritten trace made by a person (HT) was separated from the exam template given in the form (ET) using an automated method, as shown in Fig. 2.17 below. Consequently, nine numerical features, namely, root mean square, largest ET and HT radius difference, smallest ET and HT radius difference, standard deviation of ET and HT radius difference, mean relative tremor, maximum ET, minimum HT, standard deviation of exam template values, number of instances where the HT and ET radius difference undergoes a change from negative value to positive value or vice versa.

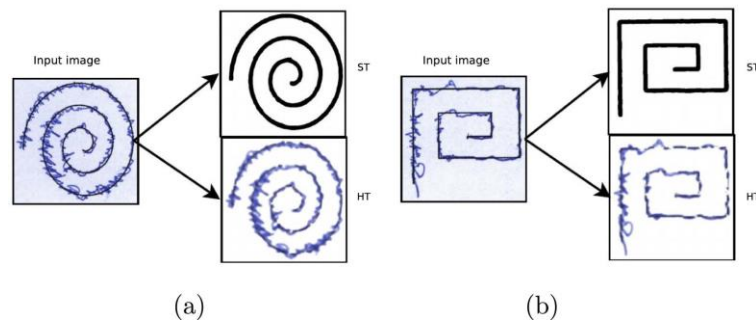


Fig. 2.17 (a) HT and ET for a Spiral image (b) HT and ET for a Meander image.

The deviations were calculated with respect to the radius of the meander or spiral. The radius was defined as the distance between the straight line connecting the center of a meander or spiral and the sample point under consideration as shown in Fig. 2.18. The high deviations point to more impact of the PD resulting in more tremors.

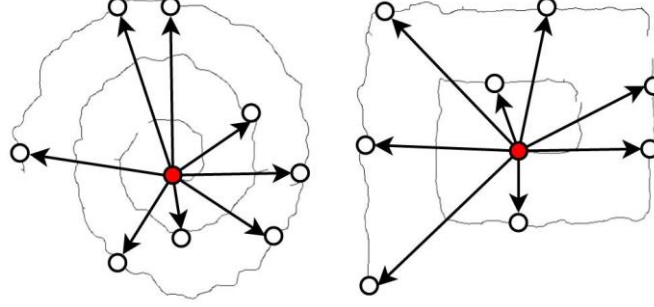


Fig. 2.18 Arbitrary points of spiral and meander images. Every vector starts from the central point of meander or spiral and ends up at the point selected randomly.

The features extracted from the dataset are as follows:

1. Root Mean Square (RMS) – It is calculated as

$$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_{HT}^i - r_{ET}^i)^2} \quad (2.2)$$

where n = sampled points' number, r_{HT}^i = HT radius of i^{th} sample point and r_{ET}^i = ET radius of i^{th} sample point.

2. The largest ET and HT radius difference which is given by

$$d_{max} = \arg \max_i \{|r_{HT}^i - r_{ET}^i|\} \quad (2.3)$$

3. The smallest ET and HT radius difference given by

$$d_{min} = \arg \min_i \{|r_{HT}^i - r_{ET}^i|\} \quad (2.4)$$

4. The standard deviation of ET and HT radius difference.

5. Mean Relative Tremor (MRT) – It measures the amount of tremor is as follows

$$MRT = \frac{1}{n-d} \sum_{i=d}^n |r_{ET}^i - r_{ET}^{i-d+1}| \quad (2.5)$$

where d is the data points' displacement used to calculate the difference in the radius. Three features listed below are computed using relative tremor $|r_{ET}^i - r_{ET}^{i-d+1}|$.

6. Maximum ET.
7. Minimum ET.
8. Standard Deviation of exam template values.

9. The number of instances where the HT and ET radius difference undergoes a change from negative value to positive value or vice versa.

These features are then compiled in a .csv file separately for spiral and meander data of the HandPD dataset along with other details as shown in Fig. 2.19 below.

```
[ ] X_df_M = df_Meander.iloc[:, 6:16]
    print(X_df_M.columns)
    print(X_df_M.shape)

Index(['AGE', 'RMS', 'MAX_BETWEEN_ST_HT', 'MIN_BETWEEN_ST_HT',
      'STD_DEVIATION_ST_HT', 'MRT', 'MAX_HT', 'MIN_HT', 'STD_HT',
      'CHANGES_FROM_NEGATIVE_TO_POSITIVE_BETWEEN_ST_HT'],
      dtype='object')
(368, 10)
```

Fig. 2.19. Columns heads of the .csv file generated after feature extraction (10 features are used for prediction).

c. The nine extracted features along with age, gender and dominant hand information are then scaled using MinMax Scaler. This estimator scales and interprets each feature individually such that it is in the given range on the training set.

d. The data is resampled using SMOTE and ENN.

e. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is used to balance the data while working with imbalanced datasets. It balances class spread by arbitrarily expanding minority class samples by repeating them. It generates new samples in the middle of existing minority samples. It creates the **synthetic training data points** for the minority class **using linear interpolation**. These virtual sample points are created by arbitrarily at least one of the k-nearest neighbors for every sample of the minority class. The steps for the algorithm are as follows:

Algorithm 1: SMOTE

1. *Initialising the minority class set A , for every $x \in A$, the k-nearest neighbours of x are generated by computing the Euclidean distance of x with every other data point in set A .*
-

2. N , the rate of sampling is initialised in accordance with the proportion of imbalance. For every $\mathbf{x} \in \mathbf{A}$, N samples (i.e. $x_1, x_2, \dots x_n$) are selected arbitrarily from its k -nearest neighbours, and they build the set A_1 .

3. For every example $\mathbf{x}_k \in \mathbf{A}_1$ ($k=1, 2, 3 \dots N$), a new sample point is generated using the formula shown below:

$$\mathbf{x}' = \mathbf{x} + \mathbf{rand}(0, 1) * |\mathbf{x} - \mathbf{x}_k| \quad (2.6)$$

SMOTE is preferred over random oversampling. Random oversampling merely increases the training data size set through repetition of the original samples. It does not cause any increase in the variety of training examples, whereas SMOTE creates new (artificial) training examples based on the original ones. For instance, if it sees two examples (of the same class) near each other, it creates a third artificial one in between the original two. Table 2.5 depicts the number of samples after the applying of SMOTE on the HandPD dataset.

Table 2.5 Samples before and after the application of SMOTE on meander and spiral data when 10 features are used for prediction.

CLASS	MEANDER DATA		SPIRAL DATA	
	Before SMOTE	After SMOTE	Before SMOTE	After SMOTE
Class 0 (Parkinson's Patient)	296	296	296	296
Class 1 (Healthy subject)	72	296	72	296

f. Wilson's ENN

Following the SMOTE oversampling process, the data is rebuilt, and Wilson's ENN is applied to it to undersample the data. Wilson developed the ENN algorithm [15] where S (Sample set) is initialised the same as TS (training set), consequently every sample in S is omitted if it does not go along with most of its k nearest neighbors (with $k=3$, generally). ENN erases the noisy samples along with close boundary ones, resulting in smooth decision boundaries. Additionally, it sustains all intermediate points, which keeps it in check from decreasing the need of storage as much as other major reduction methods. The steps involved in ENN algorithm can be summarised as follows:

Algorithm 2: ENN

1. Initialise $S = X$.

2. For every x_i in X do:

-Remove x_i from S if it is miscategorised using the k -NN rule with exemplars in $\mathbf{X} - \{x_i\}$.

Table 2.6 Samples before and after the application of ENN on meander and spiral data when 10 features are used for prediction.

CLASS	MEANDER DATA		SPIRAL DATA	
	After SMOTE	After ENN	After SMOTE	After ENN
Class 0 (Parkinson's Patient)	296	296	296	296
Class 1 (Healthy subject)	296	274	296	278

g. XGBOOST

XGBoost (Extreme Gradient Boosting) [16] is the application of Gradient Boosting Machine (GBM). It is significant among the most optimum performance algorithms used for supervised learning. It can be utilised for both classification and regression issues. It is relevant due to its high execution speed. The working is summarised as follows:

Algorithm 3: XGBoost

1. If dataset DS with n number of samples and m features $DS = \{(x_i, y_i): i = 1 \dots n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$.

2. Let \hat{y}_i be the anticipated output of the ensemble tree model obtained by:

$$\hat{A}_i = \phi(X_i) = \sum_{k=1}^K f_k(X_i), f_k \in \mathcal{F} \quad (2.7)$$

where K is the sum total of trees in the model as seen in Fig.7 below, f_k depicts the k^{th} tree.

3. To solve the above equation, the optimum functions set is computed by cutting down the loss and regularisation objective.

$$(\phi) = \sum_i l(y_i, \hat{A}_i) + \sum_k \Omega(f_k) \quad (2.8)$$

where l is the loss function (predicted output \hat{y}_i and the actual output y_i difference)

-
4. Ω measures model complexity and helps eliminate overfitting. It is computed as:

$$\Omega(\mathbf{f}_k) = \gamma T + \frac{1}{2} \lambda ||\boldsymbol{\omega}||^2 \quad (2.9)$$

where T is the sum total of leaves of the tree and $\boldsymbol{\omega}$ is each leaf's weight.

5. Boosting is employed for training the model to minimise the objective function. The minimisation is done by appending a new function \mathbf{f} in each iteration of training. For the t^{th} iteration, the objective function is given as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(\mathbf{y}_i, \hat{\mathbf{A}}_i^{(t-1)} + \mathbf{f}_t(\mathbf{X}_i)) + \Omega(\mathbf{f}_t) \quad (2.10)$$

$$6. \quad \mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} \right] - \gamma \quad (2.11)$$

$$\mathbf{g}_i = \partial_{\hat{\mathbf{A}}^{(t-1)}} l(\mathbf{y}_i, \hat{\mathbf{A}}_i^{(t-1)}) \quad (2.12)$$

$$\mathbf{h}_i = \partial_{\hat{\mathbf{A}}^{(t-1)}}^2 l(\mathbf{y}_i, \hat{\mathbf{A}}_i^{(t-1)}) \quad (2.13)$$

2.8 Concluding Remarks

Class imbalance is encountered while handling real-world datasets, where one class (i.e., the minority class) contains a lesser number of instances than the other (i.e., the majority class) consists of many instances. It can be solved using algorithmic-level methods, data-level methods, cost-sensitive methods and ensembles of classifiers as discussed in this chapter. There is also emphasis on why SMOTE and ENN are used as resampling techniques in the proposed model. This chapter has also described the impact of age on the manifestation of PD. Furthermore, the entire process of the proposed methodology has been discussed step by step along with the mathematical equations in this chapter.

CHAPTER 3

IMPACT OF GENDER AND DOMINANT HAND ON PARKINSON'S DISEASE DETECTION

3.1 Introduction

This chapter is divided into two parts:

1. How the gender of a person influences the manifestation of PD?
2. What is the role of lateralization/handedness in PD manifestation?

The role of biological sex along with environment, genetics, aging and immunity status, is a significant aspect in the progression of PD has been described in detail. There are comprehensible evidences of gender-associated variances in clinical and epidemiological traits of the disease. The chances of men getting affected by PD is twice more than those of women. Men and women not only do encounter the disorder differently, but also various mechanisms appear to be entailed in the disease pathogenesis.

The second half of the chapter describes how the dominant side of the body is associated with the initial manifestation of the disease visible in the preliminary stages. There is an evident association that is present between the dominant hand and the side of the initial motor symptom in PD. Whether the initial symptom happens on the dominant or non-dominant side, it has implications for the reported first symptom, the time to diagnosis, and the time to dopaminergic treatment initiation. The side of disease onset does not influence the severity of the disease, as evaluated by the Unified Parkinson Disease Rating Scale.

3.2 Impact of gender on PD

Studies reveal that there are shreds of evidence of gender-related variances in clinical and epidemiological characteristics of the disease: PD manifestation in men is double as compared to that in women [53][54]. However, women have a higher mortality rate and the disease spreads faster in them [55]. Additionally, females have distinct symptoms and varied responses to deep brain stimulation and pharmacological therapies methodologies and in the individual assessment of well-being compared to

males [56]. PD-MCT (Parkinson's Disease Multimodal Complex Treatment) which is a posteriori analysis of multi-professional medicaments was performed in Germany in 2010– 2016. It contained pharmacological and non-pharmacological ministrations options like occupational, speech, and physical therapy. This claimed that more men were treated than women patients under this program [57].

Research on 7209 patients at 21 centres in Israel, the US, Canada, and the Netherlands revealed that females are also less vulnerable than males to have informal caregiver support (i.e., support from spouse, family, or friends). Consequently, many women use the assistance of paid caregivers. This is due to the prolonged average lifespan of women and their natural biasedness towards being caregivers rather than receivers of care, despite the fact that their spouse or caregiver is still present in their lives [55].

3.2.1 Clinical Differences

An age-associated increasing situation of PD is seen in both genders, according to a recent analysis. However, a sharp increase is seen in males of age 60– 69 and 70– 79 [58]. There are rising cases of PD in males recorded both for disease with the presence or absence of dementia [59]. Gender-based dissimilarities are key parameters that affect life prognosis in PD. The utilization of the relative survival method revealed that the identification of PD with dementia has a greater influence over life span in women than in men [60]. Park and his colleagues showed that a low body mass index (<18.5) is firmly linked with decreased time of survival, however, this decrease is relevant only in males [61]. In addition to the variances between men and women in PD prevalence and prognosis and studies have proven that there are gender-related differences in the clinical phenotype as seen in Fig. 3.1 below. Patients have varied clinical phenotypes based on their gender. Distinct factors like GLA (galactosidase alpha) contribute to the risks of the disease in men and women. PD diagnosis is largely dependent on the visibility of motor symptoms. The identification of probable gender-related variations in motor symptom plays a critical role in terms of therapeutic strategies and diagnostic accuracy. The impact of gender on the expression and severity of PD motor symptoms has been discussed in detail. Motor symptoms in females have certain characteristics like a decrease in rigidity [62], tremor (as a preliminary symptom) [56], higher propensity to generate instability in posture, and high risk of levodopa-related motor complications [63].

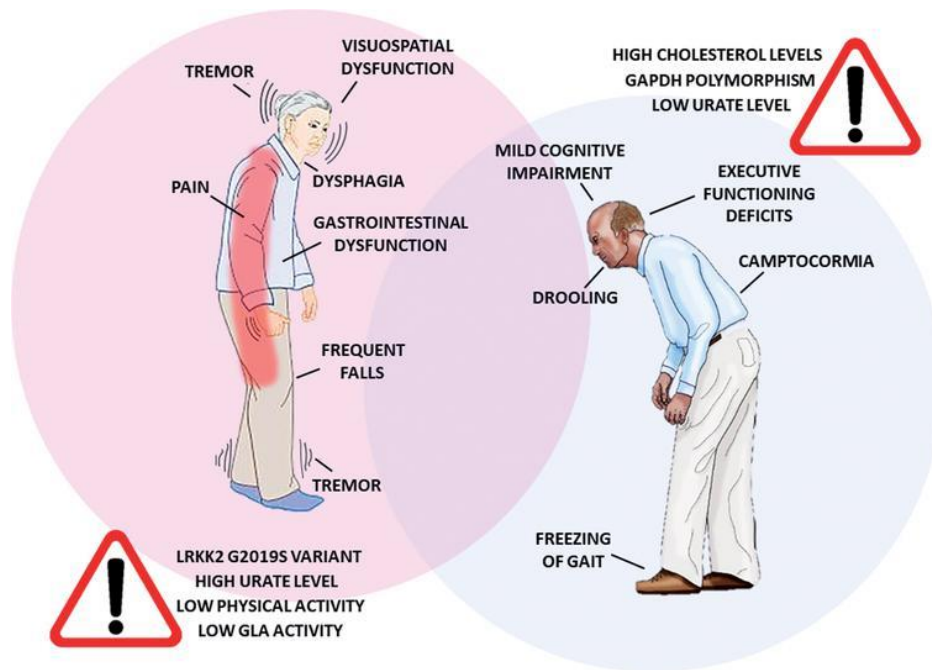


Fig. 3.1 PD risk factors and symptomatology variations in men and women.

On the contrary, the male gender has been linked with the freezing of gait as a symptom in the later stages which is the major incapacitating motor complexity of PD [64]. The female gender was indicated in the list of forecasters of progression of falls due to balance loss in PD [65]. Camptocormia, a motor distortion of PD points to aberrant severe forward bending or folding of the trunk that happens when one tries to stand or walk. This recedes in a supine position. Male patients of PD have a higher probability of developing camptocormia along with the advancement of disease [66].

3.2.2 Non-motor symptoms

A complex study on 951 PD subjects assessed the predominance and severity of non-motor symptoms according to gender. It was deduced that symptoms like depression, exhaustion, fidgety legs, pain, constipation, loss of taste or smell, drastic change in weight, extreme sweating lie among more acute symptoms categories and are frequently seen in females [67].

The link between women and pain was revealed in a wide-range clinical research showing that, along with effective and endogenous symptoms, motor impediments and younger age, female gender forecasts comprehensive severity of pain [68]. Fig. 3.2 shown below illustrates the effect of gender on PD pathophysiology. It summarizes

the major gender-related variations as the major game changers of pathogenesis of PD, concentrating on the defenceless of the dopaminergic mechanism (upper portion), neuroinflammatory cells (central portion) and oxidative stress (lower portion). IP10, interferon-inducible protein 10.

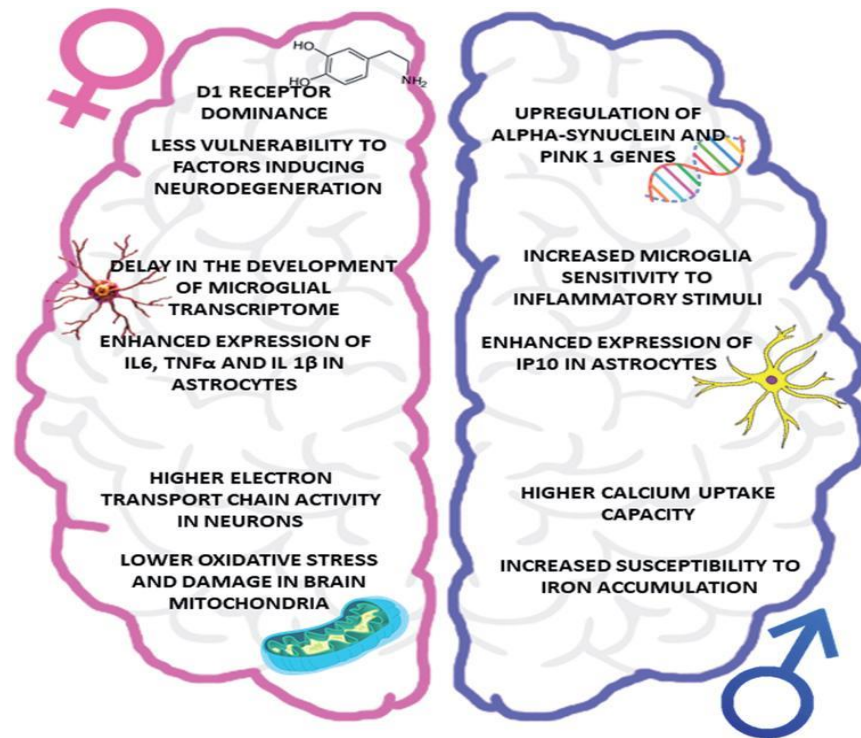


Fig. 3.2 PD pathophysiology variations in both genders.

3.2.3 Impact of gender on PD pathophysiology

The distinguishing clinical characteristics and the influence of various risk factors reveal that PD manifestation involves different pathogenetic procedures (or the same mechanism but in a different manner) in men and women respectively. Oestrogens have a relevant contribution in the gender variations in PD, giving disease prevention as shown by the similar incidents of the disease in males and females after menopause. It is important to note that sex hormones function throughout the brain of both men and women. Gender variations are now underlined in regions of brain and mechanisms that were previously not considered subject to such differences are now taken into consideration for an enhanced interpretation and insight of biological sex-associated behaviour and functions. Clinical and experimental proofs that support the idea that PD varies between males and females. Both sexes encounter the disease distinctly and

distinct mechanisms are contributors in the pathogenesis of the disease. Thus, biological sex is a significant feature in the manifestation and phenotypical expression of PD.

3.3 Impact of handedness on PD

The laterality and lateralization of symptoms often affect disease characteristics in PD. Usually, one side of the body is asymmetric with initial involvement and tends to worsen as the disease progresses. Reports claim of probable lateralized motion of nigrostriatal motor network [69]. Handedness has been associated with the initial side of the body impacted by the symptoms of PD; with the observation that the dominant side is usually the first to be affected, in both right-handed and left-handed persons [27]. Furthermore, studies show some relationship between the side of the body affected at onset of PD and clinical attributes of the disease manifestation; particularly the course of disease progression, with the right-sided onset reported to be associated with faster progression [70]. 44 participants comprising 27 males and 17 females ($p=0.688$); sex of the participants and side of the body presented with motor traits of Parkinson's disease [71]. This resulted in the deductions shown below in Table 3.1 and Table 3.2.

Table 3.1. PD influenced by laterality at onset along with age and duration of motor symptoms.

Variable	Side of the body affected at the onset		p-value
	Dominant side	Non-dominant side	
Mean age in years (SD)	66.8 (11.25)	61.7 (12.60)	0.227
Median age	69	55.5	
Mean age at onset (SD)	63.7 (11.87)	58.1 (14.15)	0.220
Median age at onset	64	50	
Mean duration of symptoms (SD)	3.0 (3.86)	3.4 (2.07)	0.780
Median duration of symptoms	2	3	
Mean time to contralateral Involvement (SD)	2.2 (3.82)	3.1 (2.21)	0.542
Median time to contralateral Involvement	1.33	2.5	

Table 3.2. p-value (level of significance) of link between PD features and laterality.

Clinical features	Dominant body side affected at onset	Dominant body side predominantly affected at presentation
Hoehn and Yahr stages	0.467	0.571
Tremors	1.000	0.795
Rigidity	0.561	0.659
Bradykinesia	0.557	0.222
Postural instability	0.557	0.703
Hypomimia	*0.012	*0.034
Non motor features	0.968	0.883
REM sleep behavior disorder	0.092	#0.029
Constipation	0.817	0.504
Hypo/anosmia	0.118	0.143
Memory deficits	0.606	0.586

* = more common in non-dominant side involvement

= more common in dominant side involvement

Table 3.3 p-value (level of significance) of link between lateralisation of symptom and handedness with PD features.

Clinical features	(p-values of association)		
	Handedness	Affected side at onset	Predominant side at presentation
Hoehn and Yahr stages	0.435	0.571	0.668
Tremors	1.000	1.000	1.000
Rigidity	1.000	0.659	0.771
Bradykinesia	0.551	0.703	0.300
Postural instability	1.000	0.647	0.473
Hypomimia	*0.018	*0.034	0.089
Non motor features	0.363	0.883	0.725
REM sleep behavior disorder	0.721	0.504	0.251
Constipation	0.721	0.401	0.710
Hypo/anosmia	0.587	0.143	0.174
Memory deficits	*0.045	0.181	0.214

* = more common in left-handed patients

= more common in left side involvement

3.4 Proposed Methodology

The methodology used is same as that discussed in chapter 2, section 2.7. The only changes are seen in the features taken as parameters of the model. The features used are age, gender, dominant hand of the subjects along with nine statistical features extracted from the image dataset (root mean square, largest value of radius difference of ET and HT, smallest value of radius difference of ET and HT, standard deviation of ET and HT radius difference, mean relative tremor, maximum ET, minimum HT, standard deviation of exam template values, number of instances where the HT and ET radius difference undergoes a change from negative value to positive value or vice versa) i.e., 12 features as shown by the python code snippet below:

```

X_df_S = df_Spiral.iloc[:, 4:16]
print(X_df_S.columns)
print(X_df_S.shape)

Index(['GENDER', 'DOM_HAND', 'AGE', 'RMS', 'MAX_BETWEEN_ET_HT',
      'MIN_BETWEEN_ET_HT', 'STD_DEVIATION_ET_HT', 'MRT', 'MAX_HT', 'MIN_HT',
      'STD_HT', 'CHANGES_FROM_NEGATIVE_TO_POSITIVE_BETWEEN_ET_HT'],
      dtype='object')
(368, 12)
time: 6.28 ms (started: 2021-03-24 11:03:41 +00:00)

```

Fig. 3.3 Columns heads of the .csv file generated after feature extraction (12 features are used for prediction).

The results of resampling of data when the features mentioned are used for prediction are seen in Table 3.4 below.

Table 3.4. a) Samples before and after the applying SMOTE on meander and spiral data when 12 features are used for prediction.

CLASS	MEANDER DATA		SPIRAL DATA	
	Before SMOTE	After SMOTE	Before SMOTE	After SMOTE
Class 0 (Parkinson's Patient)	296	296	296	296
Class 1 (Healthy subject)	72	296	72	296

Table 3.4. b) Samples before and after the applying ENN on meander and spiral data when 12 features are used for prediction.

CLASS	MEANDER DATA		SPIRAL DATA	
	After SMOTE	After ENN	After SMOTE	After ENN
Class 0 (Parkinson's Patient)	296	296	296	296
Class 1 (Healthy subject)	296	261	296	281

Apart from these variations, the method is same as that described in section 2.7.

3.5 Concluding remarks

This chapter describe the role of a person's gender and his lateral orientation on the manifestation of PD. These factors can thus be used to identify the disease in early diagnosis. It is seen that risk of getting afflicted with PD is double in men as compared to women, however, women have a higher rate of mortality and rapid manifestation of

the disease. Also, irrespective of laterality of body-side dominance, the dominant side of the body is often influenced by PD at onset and remains predominantly affected in the course of the disease. REM sleep behaviour disorder is more common when symptoms are predominant in the dominant limbs. The aim is to construct tailored intercessions and develop innovative methods to cater to the distinct needs of patients.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter broadly describes the performance parameters used to test the validation of the methodology. Once the entire process of data balancing and prediction is completed, the measurement of the validity and efficiency is determined. The performance of the model used here depends on five major parameters. These are sensitivity, F-measure/F-score, accuracy, specificity, Matthew Correlation Coefficient (MCC). Apart from these, ROC and area under the ROC curve are also utilised to test how effective the proposed model is.

4.2 Performance parameters

The performance of the proposed model is evaluated using five evaluation metrics, namely:

- a) Sensitivity
- b) F-measure/F-score
- c) Accuracy
- d) Specificity
- e) Matthew Correlation Coefficient (MCC)

It is seen that the traditional accuracy metric is unable to show the actual behaviour of a model in case of imbalanced data. Therefore, balanced accuracy metric is used for better demonstration of the functioning of the models. The accuracy, balanced accuracy, sensitivity, specificity is calculated as under:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.1)$$

$$Sen = \frac{TP}{TP+FN} \quad (4.2)$$

$$Spec = \frac{TN}{TN+TP} \quad (4.3)$$

$$ACC_{bal} = \frac{Sen+Spec}{2} \quad (4.4)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4.5)$$

$$F = \frac{2TP}{2TP+FP+FN} \quad (4.6)$$

where TP - True Positive

FP - False Positive

TN - True Negative

FN - False Negative

F - F-score/F-measure. It returns a value between 0 and 1 where 0 means worst prediction and 1 indicates perfect prediction.

MCC – It is usually between -1 and 1 where -1 means worst prediction and 1 indicates perfect prediction.

In order to validate its effectiveness, the proposed model is compared with five other earlier models. The first two models [17] and [14] work with balanced data. The effect of in-air movements while writing is used for PD detection [17]. KNN and SVM along with Adb learning model [14] achieved 81.3% accuracy. The spiral drawings were considered to distinguish between healthy and PD subject using NB model. SVM and NB along with OPF were used to obtain 67% accuracy on HandPD dataset [18].

The Chi2-Adaboost ensemble [19] used to remove model biasedness yielded the highest accuracy (balanced accuracy) of 78.04%. Table 4.1 shown below gives the detailed description of the effectiveness of the proposed hybrid resampling along with XGBoost prediction model in terms of the parameters: accuracy, balanced accuracy, F-score, sensitivity, specificity and Mathew's correlation coefficient.

4.2.1 ROC analysis

To compare the efficiency of the proposed method with the existing work [19], Receiver Operator Characteristic (ROC) charts are also utilized along with considered performance measures.

Table 4.1 Comparison of the proposed model with other models. Here IBT: Imbalanced Training, BT:Balanced Training, ACCbal:Balanced Accuracy, MCC: Mathews Correlation Coefficient.

Training	Dataset	Model	ACC	ACC _{bal}	Sen %	Spec %	F score	MCC
BT [17]	20 PD 20 Healthy	MANOVA	97.5	-	95	100	-	-
BT [14]	37 PD 38 Healthy	KNN+Adb+SV M	81.3	-	-	-	-	-
IBT [18]	37 PD 18 Healthy	NB	78.9	-	-	-	-	-
IBT [18]	HandPD	OPF+Nb+SVM	67	-	-	-	-	-
BT [19]	Meander	Chi2-Adb	74.80	78.04	68.58	87.50	0.799	0.450
	Spiral		69.40	72.46	69.96	75.00	0.794	0.365
BT	Meander	Resample with	98.24	98.14	100	96.29	0.984	0.965
	Spiral	Xgb prediction using age	95.37	95.43	94.62	96.25	0.956	0.907
BT	Meander	Resample with	97.02	97.44	94.84	100	0.973	0.941
	Spiral	Xgb prediction using age, gender & dominant hand.	97.12	97.08	97.82	96.34	0.972	0.942

It is fabricated by plotting the true positive rate (TPR) against the false positive rate (FPR). The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers with curves closer to the top-left corner indicate better performance. To compare different classifiers, it can be employed to summarize each classifier's performance into a single measure by calculating the area under the ROC curve, which is abbreviated to AUC. The ROC-AUC score of the XGB model with 10 features (age along with nine statistical features) is 0.9971 for Meander data, and that for Spiral data is 0.9936. The ROC-AUC score of the XGB model with 12 features (age, gender, and dominant hand along with nine statistical features) is 0.9911 for Meander data, and that for Spiral data is 0.9984.

It is evident from the ROC charts in Fig. 4.1 and 4.2 that the proposed model works better than the Chi2-Adaboost method that was used earlier.

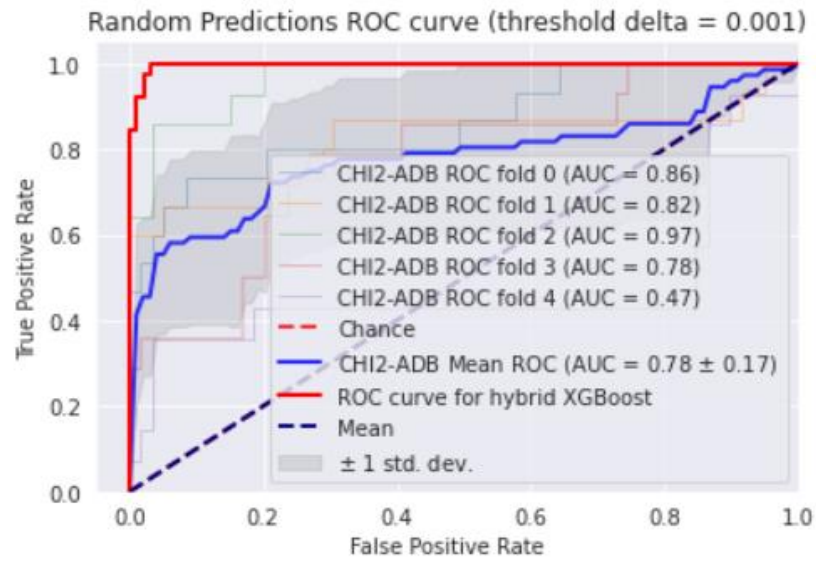


Fig. 4.1 (a) ROC comparison of the proposed method with Chi2-Adboost using Meander data with 10 features

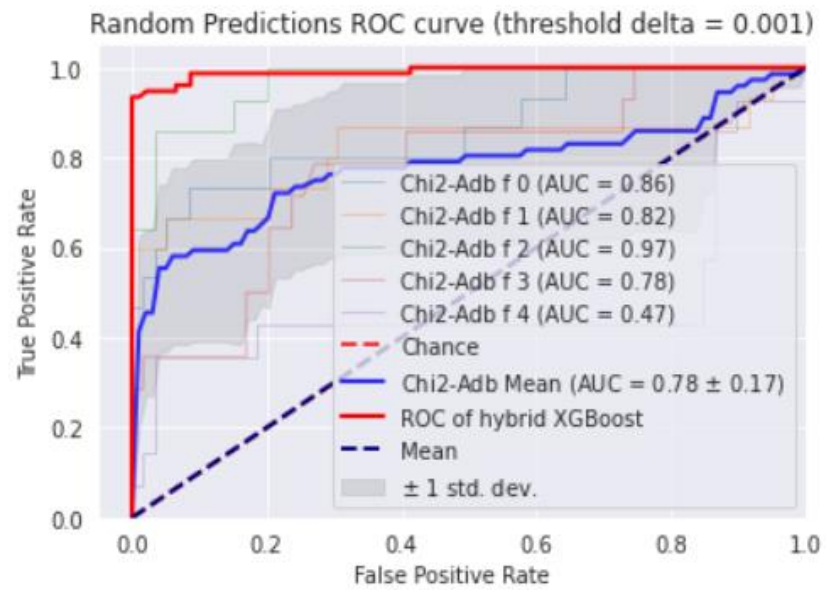


Fig. 4.1 (b) ROC comparison of the proposed method with Chi2-Adboost using Meander data with 12 features

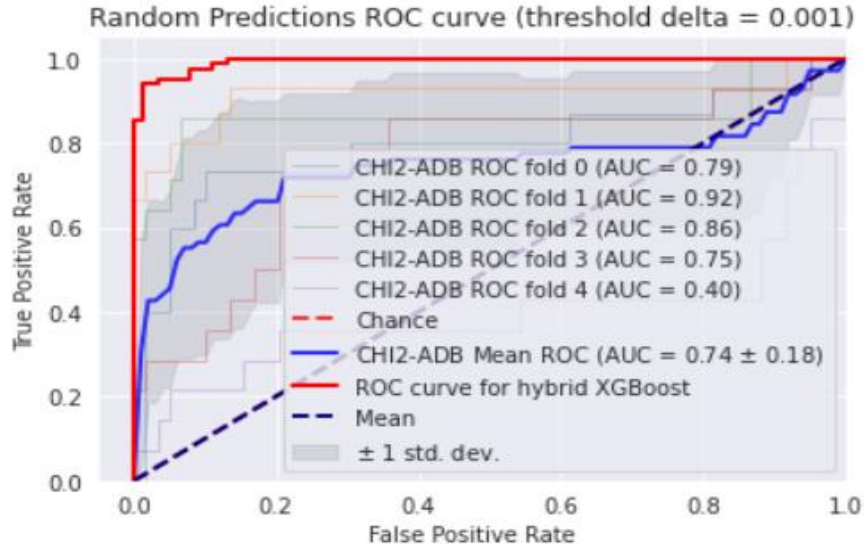


Fig. 4.2 (a) ROC comparison of the proposed method with Chi2-Adboost using Spiral data with 10 features

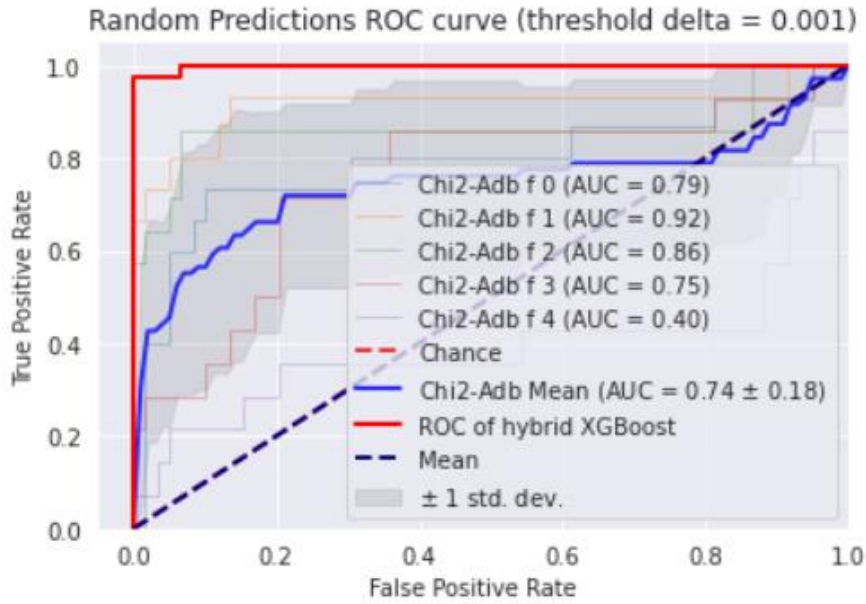


Fig. 4.2 (b) ROC comparison of the proposed method with Chi2-Adboost using Spiral data with 12 features

4.3 Advantages

The proposed method successfully eliminates the problem of data imbalance that comes with using the HandPD dataset. It uses gender, age and dominant hand information along with nine features extracted from the images. Moreover,

categorisation into gender, handedness and age gives a clear perceptive of the distinguishability of a parameter while writing task serves as an indicator of PD. There is no need for computation or instrumentation to access these features and they thus do not add to any further required mechanisms as compared to other pathological techniques. Additionally, the XGBoost prediction model works comparatively faster than the conventional Adaboost model.

4.4 Limitations

The proposed hybrid resampling and XGBoost prediction-based model have been employed to work only with the patients' handwritten samples and healthy subjects. Thus, it only makes use of the fact that the handwriting of the PD patients is affected. However, there are other symptoms like changes in motion [20] and voice impairments [21]. The handwritten data can further include parameters for measuring movement amplitudes, slowness, and rigidity that are not covered in this research. The prediction model only diagnoses the presence and absence of the PD. In future works, it can be used to state the severity of the disease. Furthermore, more parameters of the patients should be considered which impact the body neurologically similar to ageing which can help in early diagnosis.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In this work, resampling is done using SMOTE and Wilson's ENN along with XGBoost prediction. The HandPD dataset was inherently imbalanced, which led to biasedness in the machine learning models. Therefore, a low rate of specificity and a high sensitivity rate was seen due to the Parkinson's patients being in the majority and healthy individuals in the minority. To remove this biasedness, resampling was used. Moreover, an XGBoost model was implemented to improve the accuracy.

The parameters of test subjects like age, gender and dominant hand while writing is also taken as features in order to help the prediction models learn better about the distinct characteristics of the PD patients and healthy individuals. These features play an important role in the manifestation and diagnosis of the disease. It was seen that the proposed model performed better than existing state-of-the-art models. When age was taken as a feature along with nine numerical features namely, root mean square, largest value of radius difference of ET and HT, smallest value of radius difference ET and HT, standard deviation of radius difference between ET and HT, mean relative tremor, maximum ET, minimum HT, standard deviation of exam template values, number of instances where the HT and ET radius difference undergoes a change from negative value to positive value or vice versa, the highest accuracy of 98.24%, sensitivity of 100%, and specificity of 96.29% was achieved for the meander data. Similarly, for the spiral data, the model yielded an accuracy of 95.37%, sensitivity of 94.62%, and specificity of 96.25%.

It was seen that the proposed model yielded an accuracy of 97.12%, sensitivity of 94.84%, and highest specificity of 100% for meander data with age, gender and dominant hand information of the individuals taken into account along with the nine statistical features mentioned above. Similarly, for the spiral data, the model displayed an accuracy of 97.12%, sensitivity of 97.82%, and specificity of 96.34%.

5.2 Scope for future work

The proposed hybrid resampling and XGBoost prediction-based model has been employed to work only with the patients' handwritten samples and those of healthy subjects. Thus, it only makes use of the fact that the handwriting of the PD patients is affected. However, there are other symptoms like changes in motion and voice impairments. The handwritten data can further include parameters for measuring movement amplitudes, slowness, and rigidity that are not covered in this research. The prediction model only diagnoses the presence and absence of the PD. In future works, it can be used to state the severity of the disease.

REFERENCES

- [1] A. Caliskan, H. Badem, A. Baştürk, and M. E. Yüksel, “Diagnosis of the Parkinson disease by using deep neural network classifier,” *Istanbul Univ. - J. Electr. Electron. Eng.*, vol. 17, no. 2, 2017.
- [2] S. Grover, S. Bhartia, Akshama, A. Yadav, and K. R. Seeja, “Predicting Severity of Parkinson’s Disease Using Deep Learning,” in *Procedia Computer Science*, 2018, vol. 132, doi: 10.1016/j.procs.2018.05.154.
- [3] S. Aich, K. Younga, K. L. Hui, A. A. Al-Absi, and M. Sain, “A nonlinear decision tree based classification approach to predict the Parkinson’s disease using different feature sets of voice data,” in *International Conference on Advanced Communication Technology, ICACT*, 2018, vol. 2018-February, doi: 10.23919/ICACT.2018.8323864.
- [4] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, “Refining diagnosis of Parkinson’s disease with deep learning-based interpretation of dopamine transporter imaging,” *NeuroImage Clin.*, vol. 16, 2017, doi: 10.1016/j.nicl.2017.09.010.
- [5] S. L. Oh *et al.*, “A deep learning approach for Parkinson’s disease diagnosis from EEG signals,” *Neural Comput. Appl.*, vol. 32, no. 15, 2020, doi: 10.1007/s00521-018-3689-5.
- [6] K. N. R. Challa, V. S. Pagolu, G. Panda, and B. Majhi, “An improved approach for prediction of Parkinson’s disease using machine learning techniques,” 2017, doi: 10.1109/SCOPE.2016.7955679.
- [7] V. Khare, N. Mehra, and S. Akhter, “Analysis and Identification of Parkinson disease based on fMRI,” no. February, 2017.
- [8] T. Ashish, S. Kapil, and B. Manju, “Parallel bat algorithm-based clustering using mapreduce,” in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 4, 2018.
- [9] A. K. Tripathi, K. Sharma, and M. Bala, “A Novel Clustering Method Using Enhanced Grey Wolf Optimizer and MapReduce,” *Big Data Res.*, vol. 14, 2018, doi: 10.1016/j.bdr.2018.05.002.
- [10] A. K. Tripathi, K. Sharma, and M. Bala, “Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA),” *Int. J. Syst. Assur. Eng. Manag.*, vol. 9, no. 4, 2018, doi: 10.1007/s13198-017-0665-x.
- [11] C. Ornelas-Vences, L. P. Sanchez-Fernandez, L. A. Sanchez-Perez, A. Garza-Rodriguez, and A. Villegas-Bastida, “Fuzzy inference model evaluating turn for Parkinson’s disease patients,” *Comput. Biol. Med.*, vol. 89, pp. 379–388, Oct. 2017, doi: 10.1016/j.combiomed.2017.08.026.
- [12] A. Samà *et al.*, “Estimating bradykinesia severity in Parkinson’s disease by analysing gait through a waist-worn sensor,” *Comput. Biol. Med.*, vol. 84, pp. 114–123, May 2017, doi: 10.1016/j.combiomed.2017.03.020.
- [13] M. MashhadiMalek, F. Towhidkhah, S. Gharibzadeh, V. Daeichin, and M. Ali

- Ahmadi-Pajouh, “Are rigidity and tremor two sides of the same coin in Parkinson’s disease?,” *Comput. Biol. Med.*, vol. 38, no. 11–12, pp. 1133–1139, Nov. 2008, doi: 10.1016/j.compbiomed.2008.08.007.
- [14] G. Rigas *et al.*, “Assessment of tremor activity in the parkinsons disease using a set of wearable sensors,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 478–487, 2012, doi: 10.1109/TITB.2011.2182616.
 - [15] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, “Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease,” *Futur. Gener. Comput. Syst.*, 2018, doi: 10.1016/j.future.2018.02.009.
 - [16] H. B. Kim *et al.*, “Wrist sensor-based tremor severity quantification in Parkinson’s disease using convolutional neural network,” *Comput. Biol. Med.*, vol. 95, pp. 140–146, Apr. 2018, doi: 10.1016/j.compbiomed.2018.02.007.
 - [17] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, “Analysis of in-air movement in handwriting: A novel marker for Parkinson’s disease,” *Comput. Methods Programs Biomed.*, vol. 117, no. 3, 2014, doi: 10.1016/j.cmpb.2014.08.007.
 - [18] S. Rosenblum, M. Samuel, S. Zlotnik, I. Erikh, and I. Schlesinger, “Handwriting as an objective tool for Parkinson’s disease diagnosis,” *J. Neurol.*, vol. 260, no. 9, 2013, doi: 10.1007/s00415-013-6996-x.
 - [19] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, “Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson’s disease,” *Artif. Intell. Med.*, vol. 67, 2016, doi: 10.1016/j.artmed.2016.01.004.
 - [20] C. R. Pereira *et al.*, “A step towards the automated diagnosis of parkinson’s disease: Analyzing handwriting movements,” in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2015, vol. 2015-July, doi: 10.1109/CBMS.2015.34.
 - [21] C. R. Pereira *et al.*, “A new computer vision-based approach to aid the diagnosis of Parkinson’s disease,” *Comput. Methods Programs Biomed.*, vol. 136, 2016, doi: 10.1016/j.cmpb.2016.08.005.
 - [22] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, “Reliable Parkinson’s Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model,” *IEEE Access*, 2019, doi: 10.1109/access.2019.2932037.
 - [23] R. Alejo, J. M. Sotoca, R. M. Valdovinos, and P. Toribio, “Edited nearest neighbor rule for improving neural networks classifications,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6063 LNCS, no. PART 1, pp. 303–310, 2010, doi: 10.1007/978-3-642-13278-0_39.
 - [24] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, “Clustering-based undersampling in class-imbalanced data,” *Inf. Sci. (Ny)*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.

- [25] M. V. Shidha and T. Mahalekshmi, "An Emperical Study on the Effect of Resampling Techniques in Imbalaced Datasets for Improving Consistency of Classifiers," *Appl. Eng. Res.*, vol. 14, no. 7, pp. 1516–1525, 2019.
- [26] U. Gupta, H. Bansal, and D. Joshi, "An improved sex-specific and age-dependent classification model for Parkinson's diagnosis using handwriting measurement," *Comput. Methods Programs Biomed.*, 2020, doi: 10.1016/j.cmpb.2019.105305.
- [27] M. J. Barrett, S. A. Wylie, M. B. Harrison, and G. F. Wooten, "Handedness and motor symptom asymmetry in Parkinson's disease," *J. Neurol. Neurosurg. Psychiatry*, vol. 82, no. 10, 2011, doi: 10.1136/jnnp.2010.209783.
- [28] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, 2004, doi: 10.1145/1007730.1007735.
- [29] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, 2009, doi: 10.1142/S0218001409007326.
- [30] G. M. Weiss, "Mining with Rarity: A Unifying Framework," *SIGKDD Explor.*, vol. 6, no. 1, 2004.
- [31] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 4, 2012, doi: 10.1109/TSMCC.2011.2161285.
- [32] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 2, 2009, doi: 10.1109/TSMCB.2008.2007853.
- [33] T. Kim and H. Ahn, "A Hybrid Under-sampling Approach for Better Bankruptcy Prediction," *J. Intell. Inf. Syst.*, vol. 21, no. 2, 2015, doi: 10.13088/jiis.2015.21.2.173.
- [34] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Syst.*, vol. 41, 2013, doi: 10.1016/j.knosys.2012.12.007.
- [35] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *Science (80-.)*, vol. 30, no. 1, 2006.
- [36] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, 2015, doi: 10.1016/j.patcog.2014.10.032.
- [37] W. W. Cohen, "Fast Effective Rule Induction," in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 115–123.
- [38] B. Raskutti and A. Kowalczyk, "Extreme re-balancing for svms: a case study," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, 2004.
- [39] N. K. Kasabov, "Evolving connectionist systems for adaptive learning and

knowledge discovery: Trends and directions,” *Knowledge-Based Syst.*, vol. 80, pp. 24–33, May 2015, doi: 10.1016/j.knosys.2014.12.032.

- [40] N. K. Kasabov, M. G. Doborjeh, and Z. G. Doborjeh, “Mapping, learning, visualization, classification, and understanding of fMRI Data in the NeuCube evolving spatiotemporal data machine of spiking neural networks,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 4, 2017, doi: 10.1109/TNNLS.2016.2612890.
- [41] N. Kasabov *et al.*, “Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke,” *Neurocomputing*, vol. 134, pp. 269–279, Jun. 2014, doi: 10.1016/j.neucom.2013.09.049.
- [42] A. Mohemmed and N. Kasabov, “Incremental learning algorithm for spatio-temporal spike pattern classification,” 2012, doi: 10.1109/IJCNN.2012.6252533.
- [43] B. Machado, T. Rodrigues, Z. Lopes, R. Lopes, and M. Mesquita, “Paraparesis: A rare presentation of thrombosis of the abdominal aorta,” *Eur. J. Intern. Med.*, vol. 24, no. 1, p. e256, 2013, doi: 10.1016/j.ejim.2013.08.659.
- [44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, 2002, doi: 10.1613/jair.953.
- [45] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” 2008, doi: 10.1109/IJCNN.2008.4633969.
- [46] J. Zhang and I. Mani, “KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction,” 2003.
- [47] P. A. Devijver and J. Kittler, “Pattern recognition: a statistical approach,” *Pattern Recognit. a Stat. approach.*, 1982, doi: 10.1016/0262-8856(85)90018-6.
- [48] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, 1967, doi: 10.1109/TIT.1967.1053964.
- [49] B. D. Ripley, *Pattern recognition and neural networks*. 2014.
- [50] C. S. Penrod and T. J. Wagner, “ANOTHER LOOK AT THE EDITED NEAREST NEIGHBOR RULE,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-7, no. 2, 1977, doi: 10.1109/tsmc.1977.4309660.
- [51] A. S. Buchman *et al.*, “Nigral pathology and parkinsonian signs in elders without Parkinson disease,” *Ann. Neurol.*, vol. 71, no. 2, 2012, doi: 10.1002/ana.22588.
- [52] M. Rodriguez, C. Rodriguez-Sabate, I. Morales, A. Sanchez, and M. Sabate, “Parkinson’s disease as a result of aging,” *Aging Cell*, vol. 14, no. 3. 2015, doi: 10.1111/accel.12312.
- [53] M. Baldereschi *et al.*, “Parkinson’s disease and parkinsonism in a longitudinal study: Two-fold higher incidence in men,” *Neurology*, vol. 55, no. 9, 2000, doi: 10.1212/WNL.55.9.1358.

- [54] P. Solla *et al.*, “Gender differences in motor and non-motor symptoms among Sardinian patients with Parkinson’s disease,” *J. Neurol. Sci.*, vol. 323, no. 1–2, 2012, doi: 10.1016/j.jns.2012.07.026.
- [55] N. Dahodwala *et al.*, “Sex disparities in access to caregiving in Parkinson disease,” *Neurology*, vol. 90, no. 1, 2018, doi: 10.1212/WNL.0000000000004764.
- [56] D. Georgiev, K. Hamberg, M. Hariz, L. Forsgren, and G. M. Hariz, “Gender differences in Parkinson’s disease: A clinical perspective,” *Acta Neurologica Scandinavica*, vol. 136, no. 6. 2017, doi: 10.1111/ane.12796.
- [57] D. Richter *et al.*, “Dynamics of Parkinson’s Disease Multimodal Complex Treatment in Germany from 2010–2016: Patient Characteristics, Access to Treatment, and Formation of Regional Centers,” *Cells*, vol. 8, no. 2, 2019, doi: 10.3390/cells8020151.
- [58] L. Hirsch, N. Jette, A. Frolkis, T. Steeves, and T. Pringsheim, “The Incidence of Parkinson’s Disease: A Systematic Review and Meta-Analysis,” *Neuroepidemiology*, vol. 46, no. 4. 2016, doi: 10.1159/000445751.
- [59] R. Savica, B. R. Grossardt, W. A. Rocca, and J. H. Bower, “Parkinson disease with and without Dementia: A prevalence study and future projections,” *Mov. Disord.*, vol. 33, no. 4, 2018, doi: 10.1002/mds.27277.
- [60] V. Larsson, G. Torisson, and E. Londos, “Relative survival in patients with dementia with Lewy bodies and Parkinson’s disease dementia,” *PLoS One*, vol. 13, no. 8, 2018, doi: 10.1371/journal.pone.0202044.
- [61] K. Park, T. Oeda, M. Kohsaka, S. Tomita, A. Umemura, and H. Sawada, “Low body mass index and life prognosis in Parkinson’s disease,” *Park. Relat. Disord.*, vol. 55, 2018, doi: 10.1016/j.parkreldis.2018.05.011.
- [62] Y. Baba, J. D. Putzke, N. R. Whaley, Z. K. Wszolek, and R. J. Uitti, “Gender and the Parkinson’s disease phenotype,” *J. Neurol.*, vol. 252, no. 10, 2005, doi: 10.1007/s00415-005-0835-7.
- [63] D. Colombo *et al.*, “The ‘gender factor’ in wearing-off among patients with parkinson’s disease: A post hoc analysis of DEEP study,” *Sci. World J.*, vol. 2015, 2015, doi: 10.1155/2015/787451.
- [64] R. Kim *et al.*, “Presynaptic striatal dopaminergic depletion predicts the later development of freezing of gait in de novo Parkinson’s disease: An analysis of the PPMI cohort,” *Park. Relat. Disord.*, vol. 51, 2018, doi: 10.1016/j.parkreldis.2018.02.047.
- [65] S. A. Parashos *et al.*, “What predicts falls in Parkinson disease?: Observations from the Parkinson’s Foundation registry,” *Neurol. Clin. Pract.*, vol. 8, no. 3, 2018, doi: 10.1212/CPJ.0000000000000461.
- [66] R. Ou *et al.*, “Predictors of camptocormia in patients with Parkinson’s disease: A prospective study from southwest China,” *Park. Relat. Disord.*, vol. 52, 2018, doi: 10.1016/j.parkreldis.2018.03.020.
- [67] P. Martinez-Martin *et al.*, “Gender-related differences in the burden of non-motor symptoms in Parkinson’s disease,” *J. Neurol.*, vol. 259, no. 8, 2012, doi:

10.1007/s00415-011-6392-3.

- [68] M. A. Silverdale *et al.*, “A detailed clinical study of pain in 1957 participants with early/moderate Parkinson’s disease,” *Park. Relat. Disord.*, vol. 56, 2018, doi: 10.1016/j.parkreldis.2018.06.001.
- [69] J. Jankovic, “Parkinson’s disease: Clinical features and diagnosis,” *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 79, no. 4. 2008, doi: 10.1136/jnnp.2007.131045.
- [70] C. R. Baumann, U. Held, P. O. Valko, M. Wienecke, and D. Waldvogel, “Body side and predominant motor features at the onset of Parkinson’s disease are linked to motor and nonmotor progression,” *Mov. Disord.*, vol. 29, no. 2, 2014, doi: 10.1002/mds.25650.
- [71] S. K. Oparah and S. I. Ozomma, “Impact of Handedness and Symptom Lateralization on Disease Characteristics in Parkinson ’ s Disease in Calabar , Southern Nigeria,” vol. 17, no. 2, pp. 43–51, 2020, doi: 10.5829/idosi.wjms.2020.43.51.
- [72] A. Ibrahim Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, “Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia,” *Ain Shams Eng. J.*, no. xxxx, 2021, doi: 10.1016/j.asej.2020.11.011.
- [73] Shivangi, A. Johri, and A. Tripathi, “Parkinson Disease Detection Using Deep Neural Networks,” *2019 12th Int. Conf. Contemp. Comput. IC3 2019*, pp. 1–4, 2019, doi: 10.1109/IC3.2019.8844941.

LIST OF PUBLICATIONS

Paper	Author list . Title. Conference/Journal	Status
[1]	Aishwarya Keller, Anukul Pandey, “SMOTE and ENN based XGBoost prediction model for Parkinson’s disease detection” International conference on Smart Electronics and Communication (ICOSEC 2021)	Accepted
[2]	Aishwarya Keller, Anukul Pandey, “Hybrid resampling and XGBoost Prediction using patient’s details as features for Parkinson’s Disease Detection” International Conference on Innovative Computing, Intelligent Communication and Smart Electrical systems (ICSSES -2021)	Accepted
[3]	Aishwarya Keller, Anukul Pandey, “Resampling, Bagboost and mutual information techniques for optimization of Parkinson’s Disease Prediction Model” Journal of medical systems, Springer	Communicated