Machine Learning Approaches for Early Diagnosis of Parkinson's Disease: A Comparative Study and Model Optimization

A Dissertation Submitted

In Partial Fulfillment of the Requirements

for the Degree of

MASTERS OF SCIENCE

in

BIOTECHNOLOGY

by

Tannu Yadav

2K23/MSCBIO/62

Under the Supervision of

Prof. Yasha Hasija

Department of Biotechnology

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-110042, India

June, 2025

ACKNOWLEDGEMENT

I’d like to offer my earnest appreciation to everyone, without whom, this project would have been impossible to cover upon.

Chiefly and primarily, I am indebted to my mentor, Prof. Yasha Hasija, Head of the Department, Department of Biotechnology, Delhi Technological University, for her unfailing leadership, critical insights, and support throughout this journey. Their experience and careful counsel have been invaluable at every level of this project.

I want to convey my heartfelt gratitude to Ms. Khushi Yadav and Ms. Akansha Bisht for their incisive insights and persistent mentoring. Their insightful remarks on the project's issues were critical to its successful completion.

I would like to thank Mr. Jitender Singh and Mr. C.B. Singh, the technical personnel, for their backing whenever necessary.

Special thanks are also owed to my peers, whose ponderings and suggestions helped me refine my notions and approach. Their companionship and intellectual support have been truly invaluable.

I am also obliged to my family and friends, who made me walk easily on this grim path of thesis progression.

Finally, I would like to acknowledge the researchers whose introductory work on Parkinson’s disease, UPDRS based voice analysis, and ML techniques laid the groundwork upon which this thesis has been built. Their contributions have been a continuous source of encouragement.

This study lays a big milestone in my academic career, and I am very grateful to everyone who has contributed towards the same.

Tannu Yadav

23/MSCBIO/62

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad, Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **TANNU YADAV**, hereby certify that the work which is being presented as the major project in the thesis entitled **"**Machine Learning Approaches for Early Diagnosis of Disease: A Comparative Study and Model Optimization**"** in partial fulfilment of the requirements for the award of the Degree of Masters of Science in Biotechnology, and submitted to the Department of Biotechnology, Delhi Technological University, Delhi is an authentic record of my work carried out during the period from January 2025 to May 2025 under the supervision of **Prof. Yasha Hasija.**

I have not submitted the matter presented in the thesis for the award of any other degree from this or any other institute.

Place: Delhi — Tannu Yadav

Date: 5th June, 2025 — 23/MSCBIO/62

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad, Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **Tannu Yadav (23/MSCBIO/62)** has carried out her research work presented in this thesis entitled **"Machine Learning Approaches for Early Diagnosis of Disease: A Comparative Study and Model Optimization"** for the award of **Degree of Master of Science in Biotechnology** and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. This thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date : 5^{TH} June, 2025

Prof. Yasha Hasija
Supervisor, Head of Department
Department of Biotechnology
Delhi Technological University

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad, Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of thesis

____

____

Total Pages _______Name of the Scholar_______________________________________

Supervisor(s)

____

Department____

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: __________________ Similarity Index: _______

Total Word Count: _________

Date:

Candidate's Signature **Signature of Supervisor**

Machine Learning Approaches for Early Diagnosis of Parkinson's Disease: A Comparative Study and Model Optimization

Tannu Yadav

ABSTRACT

Parkinson's disease is the neurological illness that is comprises of both motor and non-motor symptoms which affects the population worldwide. Symptoms like Bradykinesia, postural instability, muscle stiffness, cognitive dysfunction and speech impairments are observed in the patients having PD. As early diagnosis is important for disease management but it is quite difficult to achieve when the PD symptoms are mild that causes a delay in clinical surveillance.

This study offers a novel approach to improve diagnostic accuracy by using different ML algorithms on these acoustic features, extracted from Parkinson's dataset would help in the early disease prediction. The 'Clinical Parkinson's dataset' extracted from Kaggle, comprises of various vocal parameters like jitter, shimmer, nhr etc. which is used to predict Parkinson's status by optimizing the UPDRS scores. Different Classification ML algorithms including Naïve Bayes, Logistic Regression, Random Forest, XG Boost, KNN and Deep Learning model i.e. ANN are implemented on the Parkinson's dataset for PD detection and progression. Data preprocessing , feature selection and dataset splitting are crucial steps before the application of ML models. Splitting of dataset into 80/20 ratio for the training and testing, respectively, to check the model performance. This study reveals that the Deep Learning Model, ANN, shows the highest accuracy up to 97%, followed by XG Boost with 96%. This approach also helps in minimizing prediction errors. In addition to accuracy, there are certain other metric parameters like precision, recall and F-1 score which are used for model evaluation mainly in case of class imbalance.

Incorporating voice-based data with effective ML models will facilitate the non-intrusive, and effective treatment of PD. This method holds the potential for remote precise and interpretable outcomes, resulting in early detection and enhanced patient outcomes.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PD	Parkinson's Disease
AD	Alzeihmer's Disease
ML	Machine Learning
AI	Artificial Intelligence
UPDRS	Unified Parkinson's Disease Rating Scale
(MDS-)UPDRS	Movement Disorder Society- (MDS-)UPDRS
NDD	Neurodegenerative Disorder
NB	Naïve Bayes
RF	Random Forest
XG Boost	Extreme Gradient Boosting
KNN	K-Nearest Neighbor
ANN	Artificial Neural Network
HNR	Harmonic to noise ratio
NHR	Noise to harmonic ratio
LRRK2	Leucine-rich repeat kinase 2
SNCA	α-synuclein
VPS35	Vacuolar protein sorting 35
PINK1	PTEN-induced kinase 1
PRKN	Parkin
DJ1	Parkinsonism associated deglycase
GBA 1	Glucosylceramidase beta 1
GCase	Glucocerebrosidase
GD	Gaucher's Disease
H-Y scale	Hoehn and Yahr scale

fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional near-infrared spectroscopy
SVM	Support Vector Machine
DT	Decision tree
BC	Bayes Classifier
CNN	Convolutional Neural Networks
LSTM	Long Short-term Memory Network
RPDE	Recurrence Period Density Entropy
DFA	Detrended Fluctuation Analysis
DaTSCAN	Dopamine Transporter Scan

Chapter 1
INTRODUCTION

Parkinson's disease(PD) is known to be the most frequent neurodegenerative disorder(NDD) amidst several others, including AD, Brain cancer, and Epilepsy[1]. It is a prominent neurological condition that alters the movement of muscles in the body. It hampers speech, posture, and flexibility, which causes muscle stiffness, tremors, and bradykinesia[2]This comprises two types of symptoms: Motor and non-motor symptoms. Motor symptoms, including speech disorders, are frequently experienced by PD patients. Also, imbalanced postures, muscular stiffness, and slow movement are the major motor symptoms, whereas the non-motor symptoms primarily include cognitive and sensory dysfunction, dysautonomia, and mood-related disorders[3]. Even though PD is simply and precisely recognized at an advanced stage, it is challenging to treat it effectively. As medications might be less successful in regulating the progression of PD during the advanced stage[1]. Patients with Parkinson's disorder frequently exhibit motor speech problems. Over half of the patients have speech abnormalities, especially silent and stuttered speech. This interpretation of linguistic signals is recognized as a vital non-surgical approach for assessing Parkinson's condition[3]. Observable alterations in the vocal tract include monotony, dysphonia, and a higher frequency of speech disruptions, along with diminished speech clarity. Vocal fold closure often appears partially weak in PD patients, which makes the modulation complicated[4].

Machine Learning(ML) is a branch of AI that trains from previously collected data samples and formulates predictions about novel data employing computational algorithms to complete a given task without any explicit programming. Prediction models based on ML are being constructed to monitor the outcome of rehabilitation and boost decision-making, along with advanced disease indications[5]. Integration of ML into healthcare has significantly improved early disease detection, overcoming the delays and limitations of traditional diagnostic methods. ML could shift healthcare into a more customize, preventive field via enhancing the patient outcomes, permitting more potent use of medical resources[6]

A prevalent diagnostic approach named Unified Parkinson's Disease Rating Scale(UPDRS) is employed to examine the combination of motor and sensory symptoms linked to Parkinson's illness. This helps neuroscience experts by providing a uniform method to estimate the degree of severity and progression of PD with time.

UPDRS came into existence in 1987 for examining numerous aspects of PD, comprising mental failure, motor dysfunction, and non-motor implications. In 2008, UPDRS was further updated and modified into the Movement Disorder Society- (MDS-)UPDRS[7] It comprises various key characteristics, among which the important ones are Motor and Total UPDRS, which involve major symptoms such as language, tremors, and motor abilities, along with the degree of severity. These assessment scores are used as diagnostic criteria for identifying the progression of Parkinson's illness. Apart from this, there are certain acoustic parameters like jitter, shimmer, HNR, etc. can assist in the early identification of PD by assessing the voice impairments, which makes it a feasible non-intrusive approach.

Traditional diagnostic approaches in PD diagnosis are often time-consuming and tedious to perform, because of the obscure and progressive nature of disease. Enhancing the accuracy of prevailing ML and Deep learning models, via using the different voice and speech features, retrieved from the UPDRS dataset.
The employment of the classification ML algorithm provides discrete binary classification of the Parkinson's status or severity by using the motor or total UPDRS values, present inside the corresponding dataset. These Classification ML models like Logistic regression, Naïve Bayes, etc, along with a Deep Learning network named ANN help in classifying PD by extracting the complex, undefined designs and patterns from the unprocessed data.

In the provided thesis and Literature review, we've mainly talked about different kinds of Classification Machine Learning Algorithms like Naïve Bayes, Logistic regression, Random Forest, XG Boost, KNN, and ANN that can aid in categorizing the severity of Parkinson's status using voice parameters. This study emphasizes the remodelling of continuous UPDRS scores into distinct values to accommodate the classification model because continuous values are used for regression analysis.

The implementation of an advanced classification technique into a specific dataset of voice recordings would overcome the limitations of the traditional diagnosis approach by boosting medical outcomes. This ensures the robustness of the ML algorithm by comparing and evaluating each model using various parameters like accuracy, F-1 score, precision, recall, and confusing matrix. It not only examines the accuracy but also assists in analyzing which voice trait accurately demonstrates the severity of Parkinson's diagnosis.

Chapter 2

LITERATURE REVIEW

2.1 Parkinson's disease: An Overview

PD is a degenerative neurological condition that adversely hampered movement. Symptoms of this disease emerge slowly, persist, and worsen over time. Over a million people globally suffer from Parkinson's disorder, although its precise cause is rarely acknowledged[8]. One of the main causes of PD is the gradual loss of dopaminergic neurons in the mid-brain region named substantia nigra, which is considered as "movement control area of the brain". This dopamine loss results in the unregulated release of neurons called a Hyperkinetic neurological condition[1]. This decrease in dopaminergic nerve cells along with accumulation of Lewy bodies(proteinaceous particles i.e. alpha-synuclein), present in remaining survived neurons acts as a biological hallmark for the disease[9]As a result of this, when about 80% of the nerve cells get destroyed in the substantia nigra, it may start showing indications of Parkinson's disease. Men with symptoms of PD may suffer from mating and uttering or speech-related problems[10]

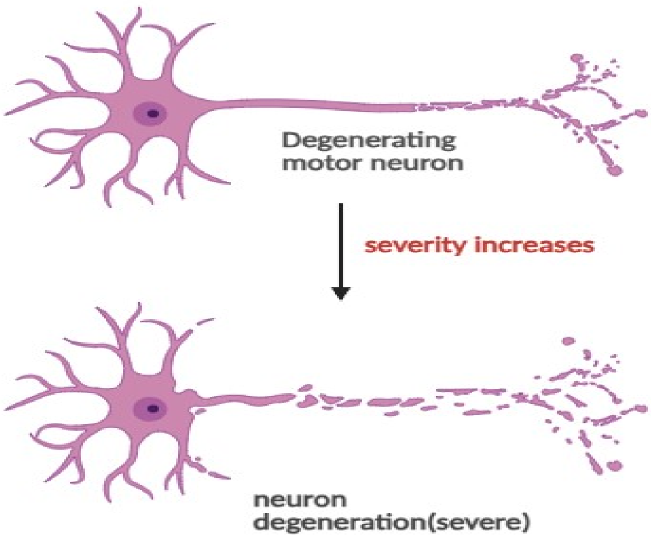


Fig: Degeneration of neuron

There are certain well known autosomal dominant and autosomal recessive genes(LRRK2, SNCA, and VPS35; and PINK1, PRKN, and DJ1), causes Parkinson's disease along with a number of genes which have been detected in few cases. Moreover, the GBA1 gene codes for glucocerebrosidase which causes Gaucher's disease, and has both prominent and unusual variations that are linked to PD, yet they frequently do not distribute in autosomal dominant families[11]

Apart from genetic factors, there are certain environmental factors comprises air pollutants like ozone, CO, NO2, NOx etc. that mainly results in brain inflammation and oxidative degradation. This will acts as a risk factor and contributes to progression of Parkinson's disease[12]

Other possible symptoms include limb stiffness, Bradykinesia, tremor, and difficulty in mobility and coordination. This diverse nature of disease makes the symptoms frequently vary from person to person. Shoulder and hip stiffness can serve as a potential indicator of the disorder. Along with this, non-motor symptoms are also linked to Parkinson's disease like anxiety, anosmia, and dementia[13]. A majority of PD cases are seen in adults above 60, among which some are caused by several underlying reasons like neural inflammation, oxidative damage, genetic mutation, protein degradation and accumulation, and adverse environmental conditions[14]. About ~90% of the PD population is assumed to be affected by vocal impairments which are usually referred to as dysarthria and dysphonia. This speech analysis is considered an effective method to support healthcare professionals because of its high prevalence, and the possibility of quick, non-intrusive, and affordable collection of voice signals[15]

The acoustic analysis comprises various parameters like jitter, shimmer, fundamental frequency, etc. which have been frequently used to detect voice peculiarities. These sound parameters have demonstrated potential for determining the voice traits across a variety of circumstances, by assessing the vocal quality in patients with speech disorders[16] According to study findings, PD could potentially be identified by cognitive signs earlier than the emergence of motor symptoms. In order to diagnose PD and ensure early diagnosis, certain clinical examinations and evaluations are needed[17]

Pathophysiological causes underlying PD are different in the preliminary and acute phases from those of later stages, so the sort of treatment must be precisely planned and implemented for the underlying disorders. The possibility of apparent Parkinsonism rises with the severity of prodromal symptoms. This phase may begin as early as age 20, before emerging motor symptoms in Parkinson's disease[18]

2.2 Traditional and Existing Diagnostic Techniques:

The prognosis of the disease is the primary stage in the treatment process. Physicians need to gain insight into the diagnosis of PD which is quite challenging for patients. It is beneficial to have a thorough and beneficial approach to diagnosis foundation along with the overall therapy. The quality of life associated with health is determined by diagnosis approval, even years after an accurate diagnosis is made[19] At the late stage of PD, the use of a particular biomarker would not create a great impact on the patient's life, even while biomarkers aid in the clinical assessment of the disease. Longevity of the disease is one of the factors that is used as a traditional method, but not considered as a reliable indicator to check PD severity[20] During Early stage PD, Clinical experts suggests that patients perform various aerobic exercises and make them familiar with the PD-specific workout initiatives available in their localities. Some medication therapies that came into existence for treating motor symptoms include dopaminergic and Levodopa therapy but during the initial stage, when it is used as a monotherapy, doesn't succeed in resulting in dementia modifying effect[21]

To determine the severity of Parkinson's disease and its associated neuron loss, distinct stages of PD have been formulated. Each exhibits a different range and severity of PD symptoms. There are majorly two rating scales namely UPDRS and Hoehn and Yahr(H-Y) scale which are used to evaluate the PD progression. During the first stage, the individual is affected with Parkinson's symptoms on one side of the body, whereas in the second stage, it spreads to both sides. Movement is mainly affected in the third stage of Parkinson's disease. Patients with the final two stages are not capable of managing routine tasks without any assistance[22]

As Traditional diagnostic approaches may have certain limitations such as late-stage prognosis, use of various Neuroimaging techniques like DaTscan, and Functional Magnetic Resonance Imaging (fMRI) which are expensive in nature.

Functional near-infrared spectroscopy (fNIRS) tracks and records deviations in cerebral blood circulation along with fluctuation in neuron activity. It is a robust approach which provides higher spatial resolution and provide resistance to movement artifacts over fMRI in determining early cognitive disorders in PD patients[23]

2.3 Voice analysis using Acoustic features and the UPDRS dataset:

The acoustic analysis of voice signals has drawn an extra attention to diagnose PD by emerging as a non- invasive approach. One kind of speech disorder named Dysarthritis arises due to disruption in central and peripheral nervous system along with affecting muscles of speech mechanism. This condition may impact production, speech, amplification and breathing, that leads to monotonous and unbearable voice[24]

The UPDRS(Unified Parkinson's Disease Rating Scale) was initially employed in the 1980's and widely utilized as a diagnostic tool for identifying and assessing the progression of PD. The updated version for speech signals is referred as (MDS-) UPDRS scale which examines the modulation, volume and clarity of voice[25] A group of experts examined and revised the scale, which comprised of four parts (Part I: Mood, Mentation and behavior; Part II: Daily routine activities; Part III: Motor; Part IV: Implications)[26]

The phonation comprises of various parameters like jitter, shimmer, harmonic to noise ratio(HNR) and noise to harmonic ratio(NHR) which results in vibration of speech sounds. An examination of fundamental frequency between periodic cycles is considered as jitter. Whereas shimmer is a variation in amplitude of sound waves [27] The periodic and non-periodic features of voice signals named as HNR, that is used to identify varying types of dysarthria and basic voice. These measurements are relying on the concept that oscillating signal and its mean remains uniform over time[28]

UPDRS is used in in the identification of motor and non-motor indications by determining the severity of PD by allotting UPDRS scores. Extracting UPDRS dataset from the platform named Kaggle and employing ML algorithms on acoustic parameters extracted from the voice recordings helps in discriminating the healthy individuals from the ones having Parkinson's disorder.

It is important to first determine the dataset quality to promote feature selection. However this is not feasible, because of background noise, resulting in acoustic features that fail to accurately present the actual condition. This noise can be eliminated by computing a threshold below the predefined frequency level[29]

2.4 Machine Learning:

ML is an AI subset which analyses the use of computers to mimic human intelligence by detecting patterns in the given data, to boost learning tasks[30] ML models are applied to multiple datasets like voice patterns and helps in recognizing appropriate features that are not employed clinical evaluation of PD [31].

ML based prediction models are being constructed to monitor rehabilitation efficacy, facilitate decision making, and to detect early symptoms. These models are capable of producing deeper knowledge about patients from massive, pre-existing datasets and provides more accessibility of databases along with code libraries[32]

In Healthcare sector, ML is used to evaluate information from several sources, and helps in assisting disease management, monitoring and results prediction. This assess the disease severity and record patient’s response to medication[33]

Machine learning is further classified into:

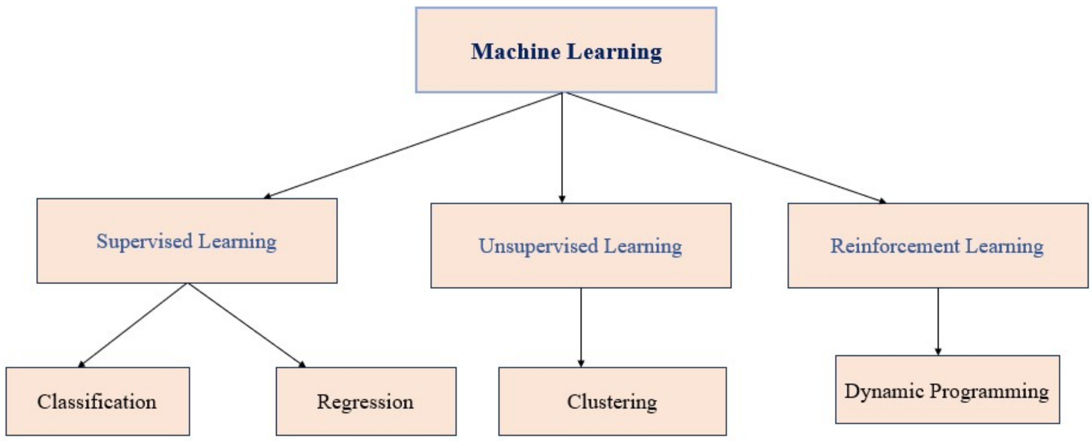


Fig: Machine learning classification

2.4.1 Supervised Learning:-

This Learning model generally require a dataset for training to train the model using already existing data. For training purpose, labelled dataset is considered which comprise raw input data and intended output. This approach is sometimes not feasible due to unavailability of database i.e. no public access[34]. Several types of Supervised ML algorithms are employed for motion analysis to estimate the extent of PD progression. Examples under this Learning are- Support Vector Machine(SVM), Decision tree(DT), Bayes Classifier(BC) etc. These algorithms are used to demonstrate the precise ranking to categorize level of severity for PD using Hohen and Yahr (H-Y) scale[35].

2.4.2 Unsupervised Learning:-

The primary objective of this method is to grouping the data according to their similarity and it does not require any training data[34].

This approach is used to reduce data dimensionality by enhancing disease detection. This method also helps in noise reduction, similarity estimation and data splitting[36] This method is based on clustering approach to organize multidimensional data according to similarity or correlation metrices[37].

2.4.3 Reinforcement Learning:-

This Learning approach usually makes the decision based on the current situation after receiving response from the surroundings for the new condition**.** It's main objective is to establish the best path that maximize the aggregate benefit, and accumulates it over the time. This discovers which actions are best by relying on trial and error method instead of providing specific instructions[38].

2.5 Applications of ML in Disease Diagnosis:

Utilizing the ideal ML practices would assist experts to predict more accurate and accessible PD algorithms, which in turn would enhance their effect on the quality of life and patient outcomes. Incorporating voice-based data with effective ML models will facilitate non-intrusive, effective treatment of PD. This method holds the potential for remote precise and interpretable outcomes, resulting in early detection and enhanced outcomes for patients.

By utilizing deeper topologies, DL models extend the potential of ANN and making it feasible to recognize abstract designs from the unprocessed data. By employing Convolutional Neural Networks (CNN) with Long Short-term Memory(LSTM) networks obtained a high accuracy score in determining voice recordings[39].

AI is utilized to diagnose the disease by employing various ML models which allow healthcare organizations to develop medical treatments for patient outcomes[40].

Machine learning is playing an increasingly vital role in medical science, especially in analyzing visual, audio, and language data. ML evaluates the weight of given feature and enhances its final prediction[41].

Their integration into healthcare has significantly improved early disease detection, overcoming the delays and limitations of traditional diagnostic methods. Research highlights how ML is transforming diagnostics across various medical fields[42].

CHAPTER 3:
METHODOLOGY

3.1 Data Collection:

Data gathering is one of the primary steps in order to build a Machine Learning based identification framework. The dataset used here for model training and evaluation is obtained from the source named Kaggle, which is a widely used platform that organizes various competitions using freely available datasets. The “Clinical Parkinson's Dataset” is taken into consideration, which contains voice measurements and clinical information from individuals who have Parkinson’s disease or not. Researchers and data experts implementing machine learning algorithms to aid in early detection of Parkinson's symptoms and progression surveillance may gain insight from the dataset.

3.1.1 Description of Dataset-

This dataset uses voice features for categorizing the Parkinson’s illness and UPDRS Scores for the disease progression. It contains cleaned and processed data related to PD. Comprising an aggregate of 23,841 records and 30 columns, which consist mainly of 10 features, i.e., voice measures, clinical evaluations, and statistical information from multiple individuals.

Table 1 :PD Dataset Description

recording_id	Unique identifier assigned to each voice sample.
fundamental_freq_hz, max_freq_hz, min_freq_hz	Frequency-related speech features.
jitter (various types)	Measures changes in fundamental frequency
shimmer (various types)	amplitude perturbation indicators
NHR, HNR	Noise-to-harmonics and harmonics-to-noise ratios, determining speech quality.
parkinson_status	Binary indicator (1 = Parkinson’s, 0 = Healthy).
rpde, dfa, spread_1, spread_2, detrended_fluctuation, ppe	Nonlinear and dynamic voice signals.
subject_id, age, gender	Demographic attributes of the subject.
test_time	Time elapsed since the first recording test for a subject.
motor_updrs_score, total_updrs_score	Clinical severity scores based on motor and total UPDRS

A1 | f_x recording_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	recording_i	fundamenta	max_freq_h	min_freq_h	jitter_perce	jitter_abs	jitter_rap	jitter_ppq	jitter_ddp	shimmer	shimmer_dł	shimmer_aı	shimmer_aı	shimmer_aı	shimmer_dc	nhr	hnr	parkinson_s	rpde	dfa
2	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
3	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
4	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
5	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
6	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
7	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
8	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
9	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
10	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
11	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
12	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
13	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
14	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
15	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
16	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
17	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
18	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
19	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
20	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
21	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
22	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
23	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
24	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
25	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
26	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
27	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
28	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528
29	phon_R01_!	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.414783	0.81528

Fig: Overview of dataset

3.2 Data Preprocessing:

3.2.1 Handling of missing data or Data Augmentation-

To make the model generalized, there is need to add the data at the place of missing values, by which it helps in reducing biasedness, and prevents overfitting during model training.

If there is no missing data available- no need of imputing is required.

While if there is presence of missing data- imputing is done or elimination of affected rows or outliers is performed.

3.2.2 Label Imbalance-

As original UPDRS scores (including motor and total UPDRS) present in continuous manner, usually for regression tasks, so there is need to convert it into discrete labels to accommodate classification models. This can be achieved by establishing a threshold for each level of severity.

In this dataset, when an Individual has Parkinson's disease, the dataset displays instances with the label '1', whereas the person without Parkinson's address the instances as label '0'. This imbalance is because of uneven distribution of these labels, which results in biased prediction by misleading the ML algorithm. To resolve this imbalance, certain python libraries like Seaborn and Matplotlib are introduced.

3.3 Feature Selection:

Selection of feature is a kind of approach use to remove redundant, irrelevant or noisy characteristics from the dataset of original features. Under Parkinson's dataset several features like jitter, shimmer, nhr or hnr and certain non-linear signal characteristics derived from audio recordings. This may carry some useful data that need to be selected and processed to reduce risk of overfitting and redundancy, and further results in feature extraction.

3.3.1 Correlation analysis

Correlation is performed to generate heat maps which determines that how a particular dataset affects other set. It is utilized to find the robust relationship among dependent(output) and independent variables, by eliminating high redundancy between independent dataset.

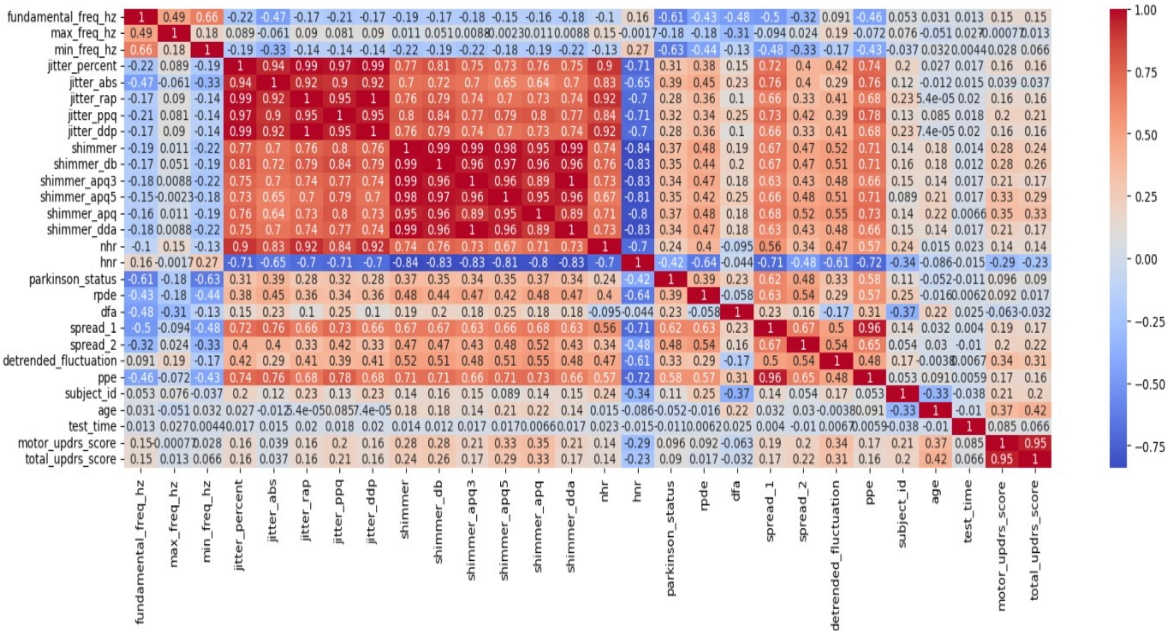


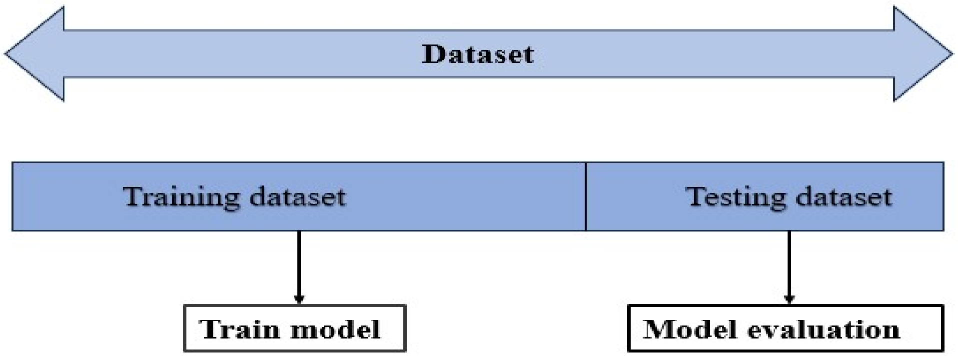
Fig: Heat map- representing correlation between voice features

3.4 Train- Test splitting:

In machine learning, Train- testing on data is a key step to ensure efficacy of model by assessing the non-reviewed data. It means that model is constructed on the previous data and later applied to perform the estimation on new data.

Thus, in order to assessing the effectiveness of selected classification algorithms, the Parkinson's dataset is split into subset for train and test. The training subset is employed to train the ML models and conducts cross-validation on the existing data, whereas training subset is used for final estimation to gauge the algorithm's reliability on undefined variables.

The 80% of data subset is utilized for training part, while 20% of the data is employed for testing purpose. By implementing train test split in python, splitting is performed.

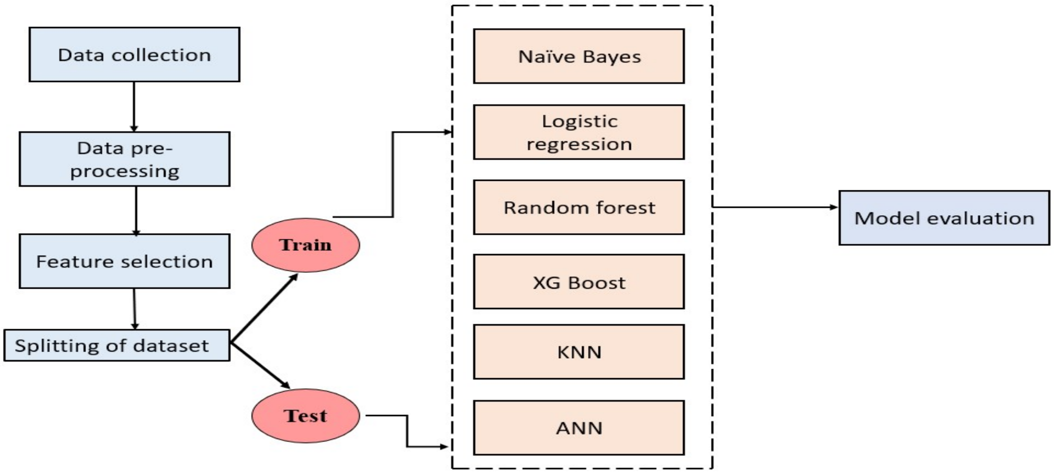


Fig; splitting of raw dataset

3.5 Model Identification:

It is one of the important steps which is followed by feature selection and correlation analysis. Here, various Classification ML algorithms like Naïve Bayes, Logistic regression, Random Forest, XG Boost, K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) are executed on the given dataset to estimate the severity level of Parkinson's disease.

The selection of ML based algorithm is depend on its capacity to manage multi-dimensional information along with managing non-linear signal characteristics.



Fig; Proposed methodology for early diagnosis of PD

3.6 Implementation of classification algorithms:

3.6.1 Naïve Bayes:

Naïve Bayes predicts the possibility of an occurrence based on its circumstances. It provides the framework for calculating probability of target event in a simple manner[43]. This is one of the widely used classification method that play an essential role in probabilistic classification. It reveals remarkable accuracy when applied to enormous datasets. It can also be used as statistical technique for categorization and Supervised ML approach[44]. NB classifiers under Bayes' theorem state that every component contributes equally and independently to the target class. While meeting the independent criteria is often challenging, the NB classifier works well in practical situations[45].

Baye's Hypothesis given below:

$$\mathbf{P(Y/X)} = \frac{\mathbf{P(X/Y)} \times \mathbf{P(Y)}}{\mathbf{P(X)}}$$

where,

P(Y/X): represents the posterior likelihood of class Y based on attributes X

P(X/Y): Conditional probability of attributes X given class Y

P(Y): prior Likelihood of class Y

P(X): Likelihood of evidence(features)

There are different types of NB classifier present, among which we used Gaussian NB classifier to apply this ML algorithm. This classifier distributes the continuous values of each voice features such as jitter, shimmer, nhr etc. according to normal distribution.

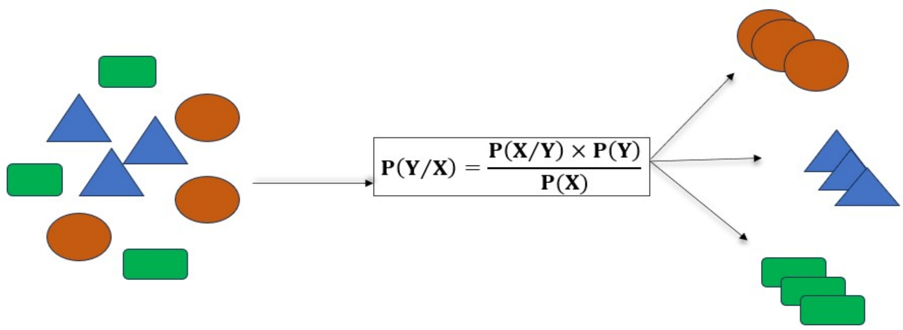


Fig: Workflow of Naïve bayes Classifier

3.6.2 Logistic Regression:

It is a classification approach which is used to estimate the probability of categorical dependent labels. The target value in dataset is binary variable which consists of two alternative values 1 and 0. The likelihood of target value 1 in dataset is predicted as function of independent variable. These both binary dependent and independent variables are either unrelated to one another or have very low correlation[46].

It utilizes the sigmoid function to estimate the chances, that a given input variable lying in which class by converting it between the value 0 and 1.

By predicting the classification probabilities of given voice features as input, it sets a boundary between different classes.

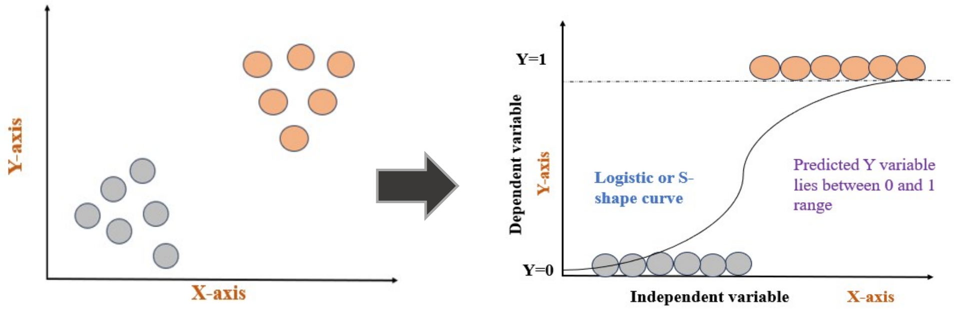


Fig: Logistic Regression- representing S-shaped curve

3.6.3 Random Forest:

Random forest classifier is a supervised ML algorithm that is applicable to both classification and regression. It educate multiple decision trees on a subset of dataset and concluding the outputs to improve the predictive accuracy of the outcomes. An average output prediction is obtained by considering the majority vote of predictions from all the models. This indicates that it does not treat any single decision tree model as superior[47]. RF classifier is used in several fields because of its adaptability, and its ability to manage complex datasets[48]. It built many decision trees based on various data subsets, instead of relying on a single one.

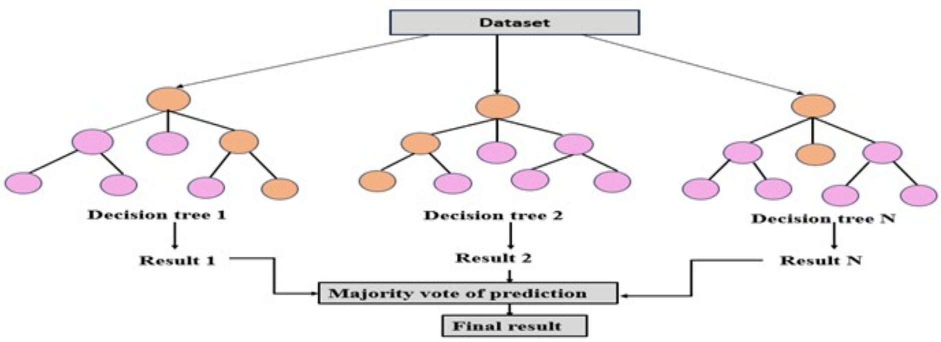


Fig: Workflow of Random Forest Classifier

For the Classification ML algorithm, a majority vote of the predicted trees was considered. This RF classifier helps to reduce the overfitting problem which can be seen under decision trees.

3.6.4 XG Boost:

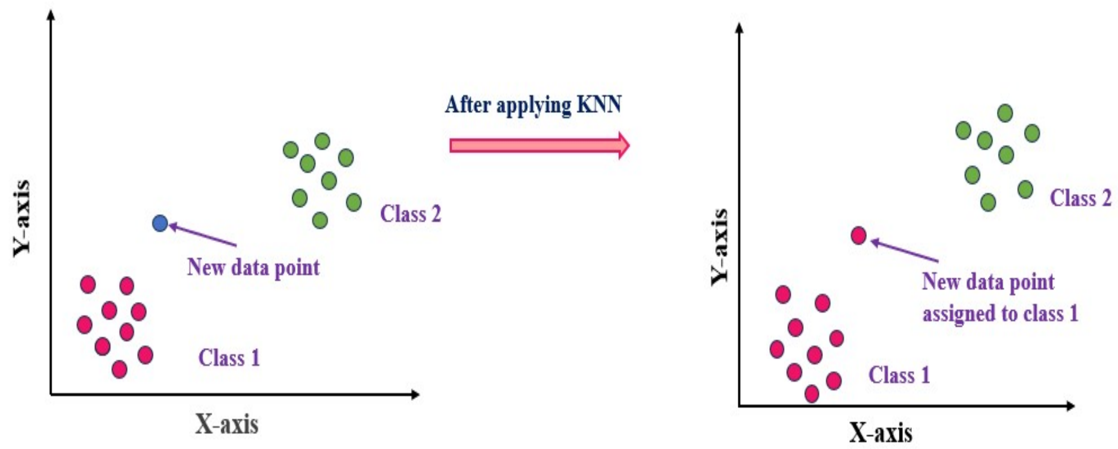
XG Boost ML algorithm is well suited for different classification and regression tasks.Tianqi Chen, a data scientist, introduced the extreme Gradient Boosting method in 2014. It is a decision tree-based technique that stands for extreme Gradient Boosting. This algorithm generates billions of outcomes very quickly[49]. This approach counts the instances a feature is required for splitting the data among all the model's trees to determine its relevance. All relevant features are thus found, and eliminating the less significant features of the dataset further helps the model boost its efficiency and is computationally simpler[50].

It is widely used gradient boosting algorithm which provides high accuracy and scalability. It reduces the overfitting problems by improving model predictions.

3.6.5 K-Nearest Neighbor:

It is known to be the most simplified algorithms among various ML models which is used to perform classification tasks. This made the predictions by relying on outcomes of k-neighbors, present nearest to that data point[51].

By altering KNN with numerous modifications, it results in different KNN variants. These KNN variants vary in a number of algorithmic ways, including truncating training datasets, assigning weight to different datapoints, k-value optimization, and enhancing distance calculations [52]



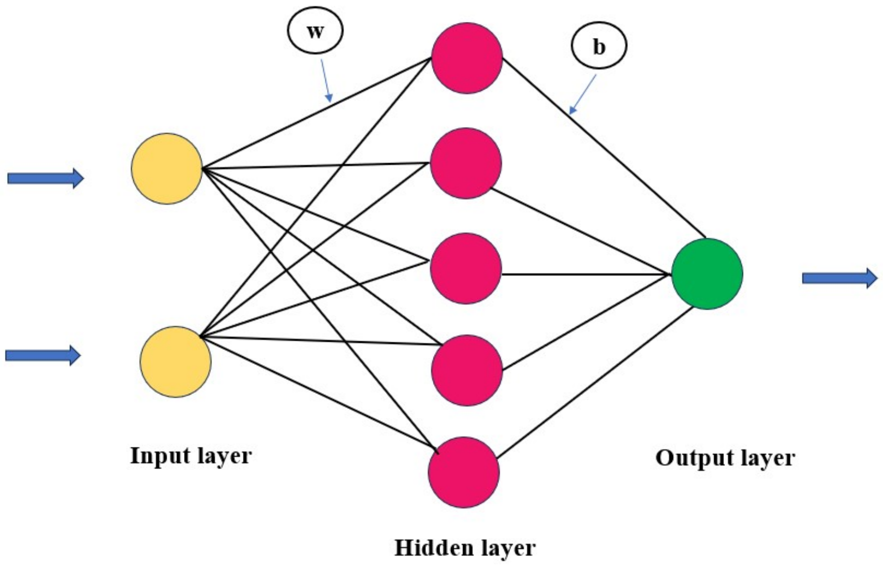
Fig; KNN- Identifying nearest neighbors in feature space

In feature space, it easily predicts the common data point under the class that is close to it. It shows greater importance in solving classification and regression problems of noise levels and performs computation of training data set during prediction.

3.6.6 Artificial Neural Network(ANN):

ANN is a branch of Deep Neural Network(Deep Learning algorithm) which makes the prediction about functioning of human brain. ANN and human brain significantly differ from one another as machine only have certain number of processors, whereas brain possess 'n' number of parallel neurons[43]. ANN predicts a function coupled to multiple outputs, inspired by biological neural networks that are made up of a vast network of interconnected neurons. These connections are weighted according to knowledge of prior experience, making it an adaptable network along with learning ability[53].

It represents a parallel design which is influenced by the functioning of biological neural network. Despite the presence of many of numerous ANN design variants, the widely used architecture is MLP(Multilayer feed-forward neural network)[51]. This neural network is made up of three distinct layers: input, hidden and output. A significant number of input nodes contributes to individual layer. Typically, the elements in the dataset are used to define these values. A neuron is integrated into the input layer for every value of dataset. Output layer is used to illustrate the disease classification, by applying trial and error method to the hidden layer's neuronal count[54].



Fig; Workflow of ANN

3.7 Model Optimization:

Hyperparameters are assigned through training to determine the model construction and are adjusted to obtain the best performance among all the models. There are two ways namely grid search or random search that are used for tuning followed by parameter selection to

reduce the model error. This method is quite expensive as it involves computational power and the use of complex dataset[55].

There is another way to evalute model's performance by performing cross-validation that helps to reduce the overfitting problems produced by an imbalance in dataset.

3.8 Model Evaluation:

Machine learning algorithm usually works in two main steps when dealing with data. First, we split the dataset— about 80% is employed for model training so it can recognizes the patterns, and remaining 20% is saved for testing how well a model performs on new, unseen data.

To assess how well the model performed, several commonly used evaluation metrics were optimized i.e. accuracy, precision, recall, and F1-score. In this context, TP stands for true positives, FP examines false positives, TN stands for true negatives, and FN refers to false negatives.

i) Accuracy: reflects the model's overall correctness by estimating the proportion of total predictions it got right:

$$\frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

ii) Precision: shows the number of the instances the model identified as being positive were actually accurate:

$$\frac{\mathbf{TP}}{\mathbf{TP + FP}}$$

iii) Recall (or sensitivity): focuses on the algorithm's ability to recognize actual positive cases, showing how many true positives it captured:

$$\frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

iv) F1 Score: offers an optimal value between precision and recall, especially valuable when both types of classification errors — false positives and false negatives — are important to minimize.

$$\frac{\mathbf{2 \times Precision \times Recall}}{\mathbf{Precision + Recall}}$$

CHAPTER 4:
RESULTS AND DISCUSSION

4.1 Results and Evaluation of Machine Learning Algorithms

Before the Implementation of ML models, Exploratory Data Analysis(EDA) is performed to recognize the patterns and variations in the voice dataset. Correlation analysis among the dependent and independent variables is performed to find the strong correlation among them and by recognizing the feature distribution to estimate the severity of disease.
This section examines the results of various ML models by comparing them with one another by using various performance metrics. The optimization of each model is done by using accuracy, precision, recall, and F1 score, along with a confusion matrix that helps in determining variation between classification values.

Using various Classification ML models on the voice dataset provides actual healthcare solutions. The result of each model is shown below along with the corresponding confusion matrix.

4.1.1 Naïve Bayes Algorithm:

```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()
nb.fit(X_train_scaled, y_train)
y_pred_nb = nb.predict(X_test_scaled)

print("Naive Bayes Accuracy:", accuracy_score(y_test, y_pred_nb))
print("Classification Report:\n", classification_report(y_test, y_pred_nb))
ConfusionMatrixDisplay.from_estimator(nb, X_test_scaled, y_test)
plt.title("Naive Bayes Confusion Matrix")
plt.show()
```

```
Naive Bayes Accuracy: 0.8366606170598911
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.96      0.70       441
           1       0.99      0.80      0.89      1763

    accuracy                           0.84      2204
   macro avg       0.77      0.88      0.79      2204
weighted avg       0.90      0.84      0.85      2204
```

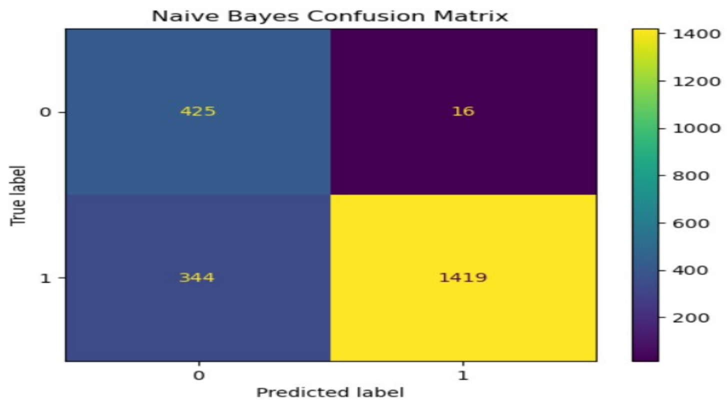
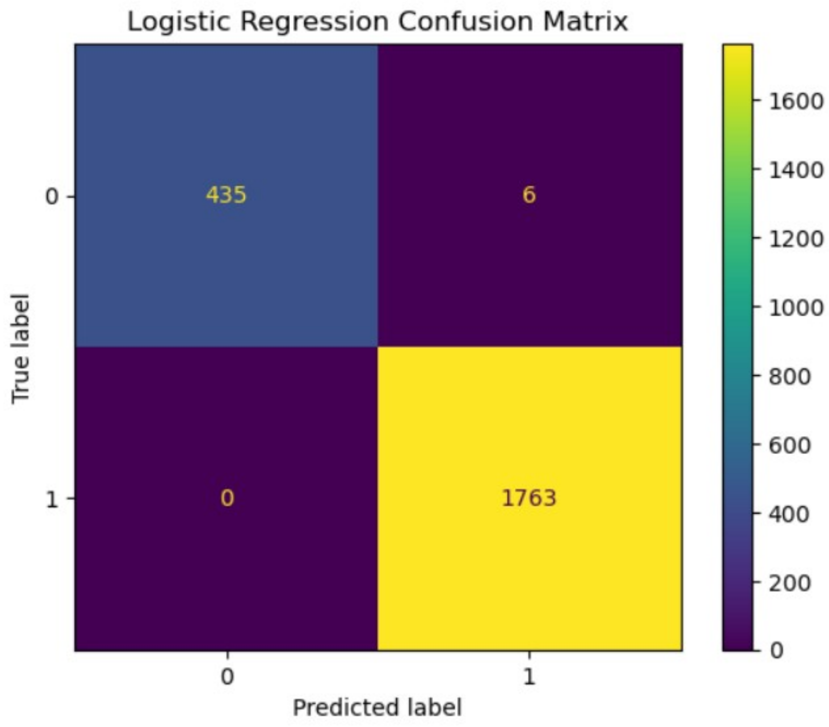


Fig: Naïve Bayes Result

4.1.2 Logistic Regression:

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()
lr.fit(X_train_scaled, y_train)
y_pred_lr = lr.predict(X_test_scaled)

print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_lr))
print("Classification Report:\n", classification_report(y_test, y_pred_lr))
ConfusionMatrixDisplay.from_estimator(lr, X_test_scaled, y_test)
plt.title("Logistic Regression Confusion Matrix")
plt.show()
```

```
Logistic Regression Accuracy: 0.9972776769509982
Classification Report:
               precision    recall  f1-score   support

           0       1.00      0.99      0.99       441
           1       1.00      1.00      1.00      1763

    accuracy                           1.00      2204
   macro avg       1.00      0.99      1.00      2204
weighted avg       1.00      1.00      1.00      2204
```



Fig; Logistic Regression Result

4.1.3 Randon Forest Classifier:

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()
rf.fit(X_train_scaled, y_train)
y_pred_rf = rf.predict(X_test_scaled)

print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Classification Report:\n", classification_report(y_test, y_pred_rf))
ConfusionMatrixDisplay.from_estimator(rf, X_test_scaled, y_test)
plt.title("Random Forest Confusion Matrix")
plt.show()
```

```
Random Forest Accuracy: 1.0
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       849
           1       1.00      1.00      1.00      3920

    accuracy                           1.00      4769
   macro avg       1.00      1.00      1.00      4769
weighted avg       1.00      1.00      1.00      4769
```

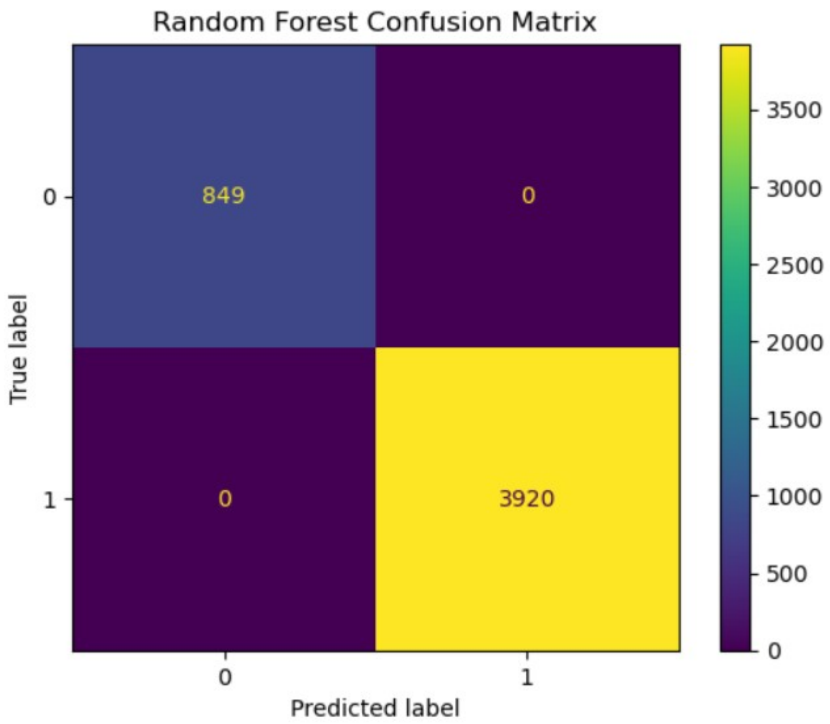


Fig Random Forest Result

4.1.4 XG Boost:

```
from xgboost import XGBClassifier

xgb = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
xgb.fit(X_train_scaled, y_train)
y_pred_xgb = xgb.predict(X_test_scaled)

print("XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("Classification Report:\n", classification_report(y_test, y_pred_xgb))
ConfusionMatrixDisplay.from_estimator(xgb, X_test_scaled, y_test)
plt.title("XGBoost Confusion Matrix")
plt.show()
```

```
XGBoost Accuracy: 1.0
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       441
           1       1.00      1.00      1.00      1763

    accuracy                           1.00      2204
   macro avg       1.00      1.00      1.00      2204
weighted avg       1.00      1.00      1.00      2204
```

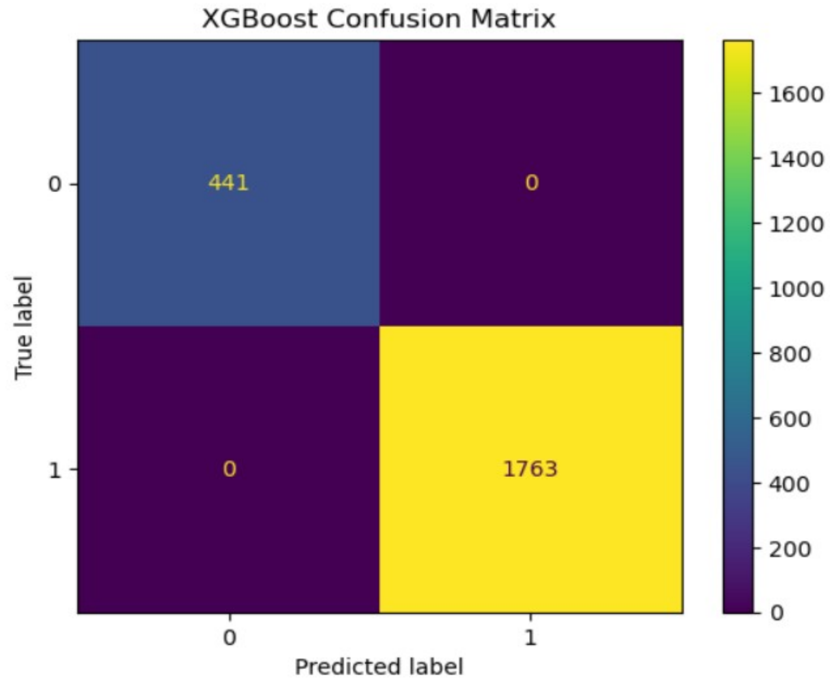
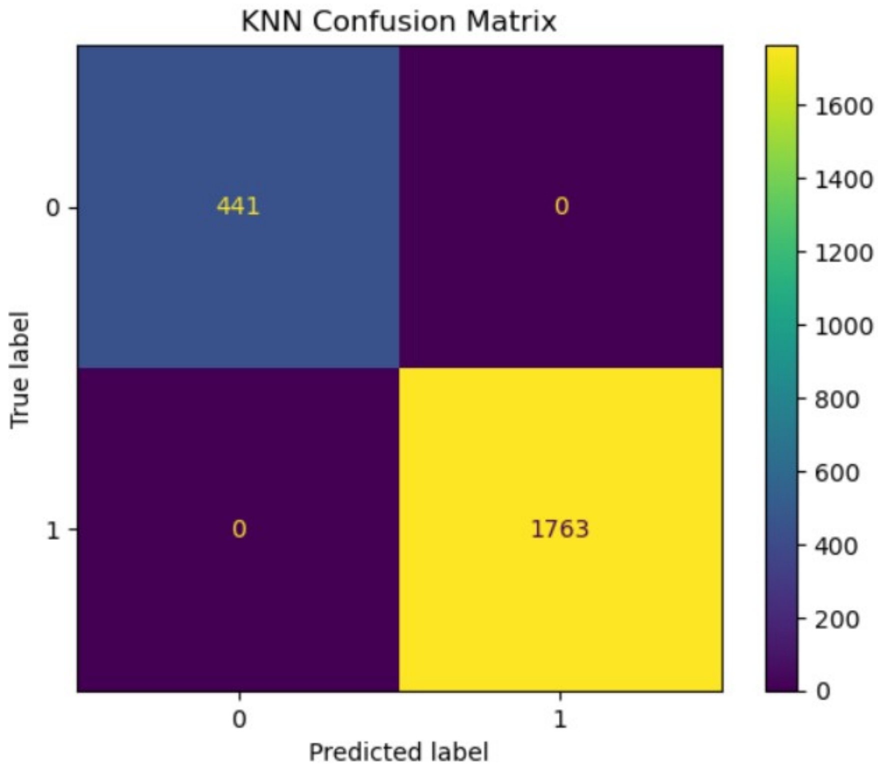


FIG : XG Boost Result

4.1.5 K-Nearest Neighbor:

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()
knn.fit(X_train_scaled, y_train)
y_pred_knn = knn.predict(X_test_scaled)

print("KNN Accuracy:", accuracy_score(y_test, y_pred_knn))
print("Classification Report:\n", classification_report(y_test, y_pred_knn))
ConfusionMatrixDisplay.from_estimator(knn, X_test_scaled, y_test)
plt.title("KNN Confusion Matrix")
plt.show()
```

```
KNN Accuracy: 1.0
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       441
           1       1.00      1.00      1.00      1763

    accuracy                           1.00      2204
   macro avg       1.00      1.00      1.00      2204
weighted avg       1.00      1.00      1.00      2204
```



Fig; Result of KNN

4.1.6 Artificial Neural Network (ANN):

```python
# Import libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# TensorFlow / Keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scale features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Build ANN model
model = Sequential()
model.add(Dense(units=64, activation='relu', input_dim=X_train_scaled.shape[1]))
model.add(Dense(units=32, activation='relu'))
model.add(Dense(units=1, activation='sigmoid'))  # Use softmax for multi-class

# Compile model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Train model
history = model.fit(X_train_scaled, y_train, epochs=50, batch_size=32, validation_split=0.2, verbose=1)
```

```python
# Predict
y_pred_prob = model.predict(X_test_scaled)
y_pred = (y_pred_prob > 0.5).astype(int).flatten()

# Evaluate
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))

# Plot confusion matrix
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Blues')
plt.title("ANN Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

```
Epoch 1/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 2s 3ms/step - accuracy: 0.8532 - loss: 0.3351 - val_accuracy: 0.9875 - val_loss: 0.0514
Epoch 2/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 0.9959 - loss: 0.0309 - val_accuracy: 1.0000 - val_loss: 0.0082
Epoch 3/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 0.0059 - val_accuracy: 1.0000 - val_loss: 0.0026
Epoch 4/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 0.0021 - val_accuracy: 1.0000 - val_loss: 0.0012
Epoch 5/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 9.6201e-04 - val_accuracy: 1.0000 - val_loss: 7.0183e-04
Epoch 6/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 5.9557e-04 - val_accuracy: 1.0000 - val_loss: 4.4634e-04
Epoch 7/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 3.9763e-04 - val_accuracy: 1.0000 - val_loss: 3.0357e-04
Epoch 8/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 2.9030e-04 - val_accuracy: 1.0000 - val_loss: 2.1909e-04
Epoch 9/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 1.8590e-04 - val_accuracy: 1.0000 - val_loss: 1.6602e-04
Epoch 10/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 1.5013e-04 - val_accuracy: 1.0000 - val_loss: 1.2442e-04
Epoch 11/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 1.1573e-04 - val_accuracy: 1.0000 - val_loss: 9.7921e-05
Epoch 12/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 8.3504e-05 - val_accuracy: 1.0000 - val_loss: 7.7994e-05
Epoch 13/50
```

```
Epoch 13/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 7.1389e-05 - val_accuracy: 1.0000 - val_loss: 6.4124e-05
Epoch 14/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 5.6629e-05 - val_accuracy: 1.0000 - val_loss: 5.1739e-05
Epoch 15/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 4.6448e-05 - val_accuracy: 1.0000 - val_loss: 4.1958e-05
Epoch 16/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 3.8496e-05 - val_accuracy: 1.0000 - val_loss: 3.4738e-05
Epoch 17/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 3.2390e-05 - val_accuracy: 1.0000 - val_loss: 2.9067e-05
Epoch 18/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 2.6440e-05 - val_accuracy: 1.0000 - val_loss: 2.4556e-05
```

```
Epoch 19/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 2.0979e-05 - val_accuracy: 1.0000 - val_loss: 2.1108e-05
Epoch 20/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 1.0000 - loss: 1.8998e-05 - val_accuracy: 1.0000 - val_loss: 1.7978e-05
Epoch 21/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 1.0000 - loss: 1.5885e-05 - val_accuracy: 1.0000 - val_loss: 1.4983e-05
Epoch 22/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 1.3146e-05 - val_accuracy: 1.0000 - val_loss: 1.2801e-05
Epoch 23/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 1.1998e-05 - val_accuracy: 1.0000 - val_loss: 1.1080e-05
Epoch 24/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 1.0008e-05 - val_accuracy: 1.0000 - val_loss: 9.4824e-06
Epoch 25/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 1.0000 - loss: 8.3208e-06 - val_accuracy: 1.0000 - val_loss: 8.2317e-06
Epoch 26/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 7.5947e-06 - val_accuracy: 1.0000 - val_loss: 7.1003e-06
Epoch 27/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 6.5025e-06 - val_accuracy: 1.0000 - val_loss: 6.1833e-06
Epoch 28/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 5.6250e-06 - val_accuracy: 1.0000 - val_loss: 5.3800e-06
Epoch 29/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 1.0000 - loss: 5.1478e-06 - val_accuracy: 1.0000 - val_loss: 4.7876e-06
Epoch 30/50
221/221 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: 1.0000 - loss: 3.9204e-06 - val_accuracy: 1.0000 - val_loss: 4.1265e-06
```

```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       441
           1       1.00      1.00      1.00      1763

    accuracy                           1.00      2204
   macro avg       1.00      1.00      1.00      2204
weighted avg       1.00      1.00      1.00      2204

Confusion Matrix:
 [[ 441    0]
 [   0 1763]]
```

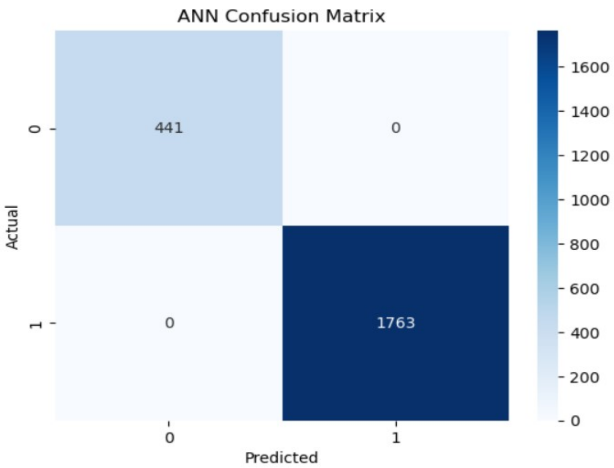
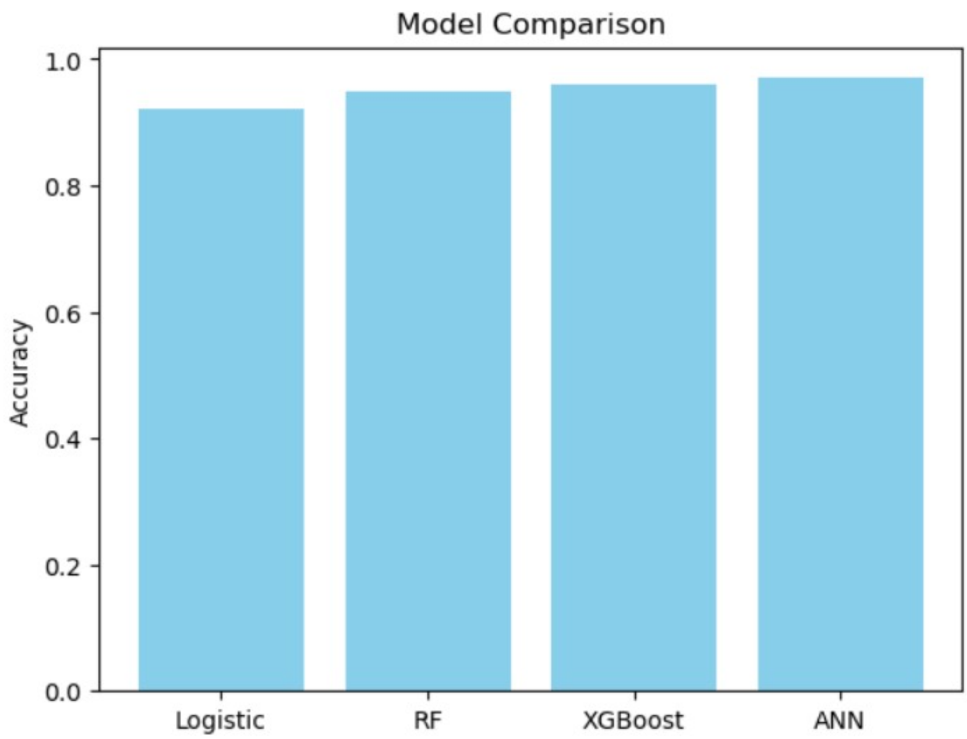


Fig ANN Result

Table : Comparison of different Classification ML algorithms

Metric	Naïve Bayes	Logisitc Regression	Random Forest	XG boost	KNN	ANN
Accuracy	83.67%	92%	95%	96%	95%	97%
Precision	0.98	0.99	1.0	1.0	1.0	1.0
Recall	0.80	1.0	1.0	1.0	1.0	1.0
F1 score	0.88	0.99	1.0	1.0	1.0	1.0

```
import matplotlib.pyplot as plt

models = ['Logistic', 'RF', 'XGBoost', 'ANN']
accuracies = [0.92, 0.95, 0.96, 0.97]

plt.bar(models, accuracies, color='skyblue')
plt.ylabel("Accuracy")
plt.title("Model Comparison")
plt.show()
```



Fig; Comparison of Different ML models

Among the different ML model tested, ANN achieved the highest accuracy on the UPDRS dataset, around 97%, followed by XG Boost with 96%. These results suggests that advanced ML models under classification algorithm are particularly effective in determining distinct patterns within complex datasets, and are used to predict PD detection and progression.

CHAPTER 5:
CONCLUSION

Voice features are one of the parameters that help in the diagnosis of PD by determining the disease severity. In this given study, we have applied several ML algorithms on the acoustic features extracted from the Clinical Parkinson's Dataset, to evaluate their performance in the identification of Parkinson's disorder.

Among all the tested model, ANN emerged as the most effective algorithm, with an accuracy of 97%, followed by the XG Boost performance of around 96%. This determines that the ANN model which is a deep learning approach is highly effective in handling the complex PD dataset. The comparative results of different ML models revealed that the both KNN and Random-forest classifier achieved an accuracy of 95%. This exhibits closely related metric values among tested algorithms that shows the reliability of acoustic features in detection of PD. A specialized architecture is followed in order to implement the classification models that requires a proper processing of raw data followed by feature selection. As a result, speech recordings are considered as robust and non-invasive diagnostic approach in monitoring neurodegenerative disorders.

This study emphasizes the future scope by integration of various data attributes such as gait and handwritten pattern in order to examine the accurate prediction of the Parkinson's disease. This multimodal solution improves the patient monitoring in healthcare sectors.

Although there are certain challenges of existing clinical methods like diagnosis at advanced level, data privacy, and limited scalability etc, that need to be addressed by implementing these ML algorithms to result in a cost-effective treatment.

REFERENCES

[1] G. Pahuja and T. N. Nagabhushan, "A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection," 2021, *Taylor and Francis Ltd.* doi: 10.1080/03772063.2018.1531730.

[2] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 249–261. doi: 10.1016/j.procs.2023.01.007.

[3] C. Quan, K. Ren, and Z. Luo, "A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech," *IEEE Access*, vol. 9, pp. 10239–10252, 2021, doi: 10.1109/ACCESS.2021.3051432.

[4] M. Dudek *et al.*, "Analysis of Voice, Speech, and Language Biomarkers of Parkinson's Disease Collected in a Mixed Reality Setting," *Sensors*, vol. 25, no. 8, Apr. 2025, doi: 10.3390/s25082405.

[5] T. Tabashum, R. C. Snyder, M. K. O'Brien, and M. V. Albert, "Machine Learning Models for Parkinson Disease: Systematic Review," 2024, *JMIR Publications Inc.* doi: 10.2196/50117.

[6] M. 'Hider, M. 'Nasiruddin, and A. 'Al Mukaddim, "Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems," 2024.

[7] R. M. Hendricks and M. T. Khasawneh, "An Investigation into the Use and Meaning of Parkinson's Disease Clinical Scale Scores," *Parkinsons Dis*, vol. 2021, 2021, doi: 10.1155/2021/1765220.

[8] S. M., M. Martin, and A. Tripathi, "ANN based Data Mining Analysis of the Parkinson's Disease," *Int J Comput Appl*, vol. 168, no. 1, pp. 1–7, Jun. 2017, doi: 10.5120/ijca2017914254.

[9] K. Kalinderi, S. Bostantjopoulou, and L. Fidani, "The genetic background of Parkinson's disease: current progress and future prospects," Nov. 01, 2016, *Blackwell Publishing Ltd*. doi: 10.1111/ane.12563.

[10] I. Ahmed, S. Aljahdali, M. S. Khan, and S. Kaddoura, "Classification of parkinson disease based on patient's voice signal using machine learning," *Intelligent Automation and Soft Computing*, vol. 32, no. 2, pp. 705–722, 2022, doi: 10.32604/iasc.2022.022037.

[11] H. R. Morris FRCP *et al.*, "The pathogenesis of Parkinson's disease."

[12] N. Palacios, "Air pollution and Parkinson's disease - Evidence and future directions," Dec. 20, 2017, *Walter de Gruyter GmbH*. doi: 10.1515/reveh-2017-0009.

[13] K. P. Swain *et al.*, "Towards Early Intervention: Detecting Parkinson's Disease through Voice Analysis with Machine Learning," *Open Biomed Eng J*, vol. 18, no. 1, Apr. 2024, doi: 10.2174/0118741207294056240322075602.

[14] F. C. Church, "Review treatment options for motor and non-motor symptoms of parkinson's disease," Apr. 01, 2021, *MDPI*. doi: 10.3390/biom11040612.

[15] S. Scimeca *et al.*, "Robust and language-independent acoustic features in Parkinson's disease," *Front Neurol*, vol. 14, 2023, doi: 10.3389/fneur.2023.1198058.

[16] G. Li, Q. Hou, C. Zhang, Z. Jiang, and S. Gong, "Acoustic parameters for the evaluation of voice quality in patients with voice disorders," *Ann Palliat Med*, vol. 10, no. 1, pp. 130–136, Jan. 2021, doi: 10.21037/apm-20-2102.

[17] T. Tabashum, R. C. Snyder, M. K. O'Brien, and M. V. Albert, "Machine Learning Models for Parkinson Disease: Systematic Review," 2024, *JMIR Publications Inc.* doi: 10.2196/50117.

[18] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," Jun. 12, 2021, *Elsevier B.V.* doi: 10.1016/S0140-6736(21)00218-X.

[19] T. E. Kimber, "Approach to the patient with early Parkinson disease: diagnosis and management," *Intern Med J*, vol. 51, no. 1, pp. 20–26, Jan. 2021, doi: 10.1111/imj.15148.

[20] K. Rukavina, L. Batzu, A. Boogers, A. Abundes-Corona, V. Bruno, and K. R. Chaudhuri, "Non-motor complications in late stage Parkinson's disease: recognition, management and unmet needs," 2021, *Taylor and Francis Ltd.* doi: 10.1080/14737175.2021.1883428.

[21] T. E. Kimber, "Approach to the patient with early Parkinson disease: diagnosis and management," *Intern Med J*, vol. 51, no. 1, pp. 20–26, Jan. 2021, doi: 10.1111/imj.15148.

[22] I. Kamran, S. Naz, I. Razzak, and M. Imran, "Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease," *Future Generation Computer Systems*, vol. 117, pp. 234–244, Apr. 2021, doi: 10.1016/j.future.2020.11.020.

[23] P. Hui *et al.*, "Exploring the application and challenges of fNIRS technology in early detection of Parkinson's disease," *Front Aging Neurosci*, vol. 16, 2024, doi: 10.3389/fnagi.2024.1354147.

[24] H. Azadi, M. R. T. Akbarzadeh, A. Shoeibi, and H. R. Kobravi, "Evaluating the Effect of Parkinson's Disease on Jitter and Shimmer Speech Features," *Adv Biomed Res*, vol. 10, no. 1, p. 54, Jan. 2021, doi: 10.4103/abr.abr_254_21.

[25] E. Majda-Zdancewicz, A. Potulska-Chromik, M. Nojszewska, and A. Kostera-Pruszczyk, "Speech Signal Analysis in Patients with Parkinson's Disease, Taking into Account Phonation, Articulation, and Prosody of Speech," *Applied Sciences (Switzerland)*, vol. 14, no. 23, Dec. 2024, doi: 10.3390/app142311085.

[26] J. Setter and J. Jenkins, "State-of-the-Art Review Article," *Language Teaching*, vol. 38, no. 1, pp. 1–17, Jan. 2005, doi: 10.1017/s026144480500251x.

[27] *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2017.

[28] S. Disease, "ACOUSTIC SPEECH MARKERS FOR TRACKING CHANGES IN HYPOKINETIC DYSARTHRIA ASSOCIATED WITH," 2023.

[29] M. Dudek *et al.*, "Analysis of Voice, Speech, and Language Biomarkers of Parkinson's Disease Collected in a Mixed Reality Setting," *Sensors*, vol. 25, no. 8, Apr. 2025, doi: 10.3390/s25082405.

[30] N. Auslander, A. B. Gussow, and E. V. Koonin, "Incorporating machine learning into established bioinformatics frameworks," Mar. 02, 2021, *MDPI AG*. doi: 10.3390/ijms22062903.

[31] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature," May 06, 2021, *Frontiers Media S.A.* doi: 10.3389/fnagi.2021.633752.

[32] T. Tabashum, R. C. Snyder, M. K. O'Brien, and M. V. Albert, "Machine Learning Models for Parkinson Disease: Systematic Review," 2024, *JMIR Publications Inc.* doi: 10.2196/50117.

[33] F. Khaliq, J. Oberhauser, D. Wakhloo, and S. Mahajani, “Decoding degeneration: The implementation of machine learning for clinical detection of neurodegenerative disorders,” Jun. 01, 2023, *Wolters Kluwer Medknow Publications*. doi: 10.4103/1673-5374.355982.

[34] G. Mirzaei and H. Adeli, “Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia,” Feb. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.bspc.2021.103293.

[35] B. E., B. D., and B. R., “Supervised machine learning based gait classification system for early detection and stage classification of Parkinson’s disease,” *Applied Soft Computing Journal*, vol. 94, Sep. 2020, doi: 10.1016/j.asoc.2020.106494.

[36] M. Nilashi *et al.*, “Predicting Parkinson’s Disease Progression: Evaluation of Ensemble Methods in Machine Learning,” *J Healthc Eng*, vol. 2022, 2022, doi: 10.1155/2022/2793361.

[37] M. R. Salmanpour, M. Shamsaei, G. Hajianfar, H. Soltanian-Zadeh, and A. Rahmim, “Longitudinal clustering analysis and prediction of Parkinson’s disease progression using radiomics and hybrid machine learning,” *Quant Imaging Med Surg*, vol. 12, no. 2, pp. 906–919, Feb. 2022, doi: 10.21037/qims-21-425.

[38] “Reinforcement learning”.

[39] M. Dudek *et al.*, “Analysis of Voice, Speech, and Language Biomarkers of Parkinson’s Disease Collected in a Mixed Reality Setting,” *Sensors*, vol. 25, no. 8, Apr. 2025, doi: 10.3390/s25082405.

[40] A. Rana, A. Dumka, R. Singh, M. Rashid, N. Ahmad, and M. K. Panda, “An Efficient Machine Learning Approach for Diagnosing Parkinson’s Disease by Utilizing Voice Features,” *Electronics (Switzerland)*, vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223782.

[41] D. J. Park, M. W. Park, H. Lee, Y. J. Kim, Y. Kim, and Y. H. Park, “Development of machine learning model for diagnostic disease prediction based on laboratory tests,” *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-87171-5.

[42] “1010-1042redc2024 (2)”.

[43] Rana, A., et al., *An efficient machine learning approach for diagnosing Parkinson’s disease by utilizing voice features.* Electronics, 2022. **11**(22): p. 3782.

[44] Bhatia, A. and R. Sulekh, *Predictive Model for Parkinson’s Disease through Naive Bayes Classification.* International Journal of Computer Science & Communication, 2017. **9**(1): p. 194-202.

[45 Farida, Y., et al., *Comparing support vector machine and naïve bayes methods with a selection of fast correlation based filter features in detecting Parkinson’s disease.* Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 2023. **14**(2): p. 80.

[46] Kamble, M., P. Shrivastava, and M. Jain, *Digitized spiral drawing classification for Parkinson's disease diagnosis.* Measurement: Sensors, 2021. **16**: p. 100047.

[47] Govindu, A. and S. Palwe, *Early detection of Parkinson's disease using machine learning.* Procedia Computer Science, 2023. **218**: p. 249-261.

[48] Srinivasan, S., et al., *Detection of Parkinson disease using multiclass machine learning approach.* Scientific Reports, 2024. **14**(1): p. 13813.

[49] Nasif, F.W., et al., *Parkinson Disease Detection: Using XGBoost Algorithm to Detect Early Onset Parkinson Disease.* ResearchGate, 2020 Available, 2020.

[50] Shyamala, K. and T. Navamani, *Design of an efficient prediction model for early parkinson’s disease diagnosis.* IEEE Access, 2024.

[51] Pahuja, G. and T. Nagabhushan, *A comparative study of existing machine learning approaches for Parkinson's disease detection.* IETE Journal of Research, 2021. **67**(1): p. 4-14.

[52] Uddin, S., et al., *Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction.* Scientific Reports, 2022. **12**(1): p. 6256.

[53] Yasar, A., et al. *Classification of Parkinson disease data with artificial neural networks*. in *IOP conference series: materials science and engineering*. 2019. IOP Publishing.

[54] Srinivasan, S.M., M. Martin, and A. Tripathi, *ANN based data mining analysis of the Parkinson's disease.* International Journal of Computer Applications, 2017. **168**(1): p. 56-60.

[55] Tabashum, T., et al., *Machine learning models for parkinson disease: Systematic review.* JMIR medical informatics, 2024. **12**(1): p. e50117.

Tanu thesis (2) (2).docx

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:99311032

Submission Date

Jun 4, 2025, 4:15 PM GMT+5:30

Download Date

Jun 4, 2025, 4:18 PM GMT+5:30

File Name

Tanu thesis (2) (2).docx

File Size

2.0 MB

30 Pages

6,492 Words

38,203 Characters

5% Overall Similarity

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

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 10 words)

Match Groups

23 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 23 Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
4% Submitted works (Student Papers)

Top Sources

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

1	Submitted works	
	Middlesex University on 2023-04-15	<1%
2	Submitted works	
	University of Monastir on 2022-08-01	<1%
3	Publication	
	Chengyi Qian, Yuanjun Wang. "MMANet: A multi-task residual network for Alzhei...	<1%
4	Submitted works	
	The University of the West of Scotland on 2023-08-23	<1%
5	Publication	
	Łukasz Piotr Pawelec, Katarzyna Graja, Anna Lipowicz, Jagoda Marchewczyk. "Wo...	<1%
6	Internet	
	dspace.vut.cz	<1%
7	Internet	
	phys.org	<1%
8	Publication	
	"Poster Presentations", Movement Disorders, 2012	<1%
9	Submitted works	
	Crown Institute of Business and Technology on 2024-09-15	<1%
10	Submitted works	
	Dire-Dawa University on 2025-05-26	<1%

11 Submitted works

Polytechnic of Turin on 2024-02-11 <1%

12 Submitted works

University of Hull on 2022-08-21 <1%

13 Submitted works

University of Wales Swansea on 2024-09-28 <1%

14 Internet

dspace.uiu.ac.bd <1%

15 Publication

Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%

16 Submitted works

The University of the West of Scotland on 2023-12-15 <1%

17 Submitted works

University of Hull on 2023-12-10 <1%

18 Submitted works

University of Technology on 2025-05-21 <1%

19 Submitted works

University of Westminster on 2024-04-10 <1%

20 Submitted works

Uttar Pradesh Technical University on 2020-05-26 <1%

21 Internet

ijisrt.com <1%

22 Internet

mdpi-res.com <1%

23 Internet

www.coursehero.com <1%

Tannu-1.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:99269486

Submission Date

Jun 4, 2025, 11:14 AM GMT+5:30

Download Date

Jun 4, 2025, 11:17 AM GMT+5:30

File Name

Tannu-1.pdf

File Size

1.8 MB

36 Pages

7,643 Words

43,020 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.