**Classification of Alzheimer’s Disease using Handwriting Analysis**

*Abstract*— The project aims to address the urgent need for accurate early diagnosis of neurodegenerative diseases, focusing on Alzheimer's disease. Traditional diagnostic methods have proven inadequate for timely identification, hindering effective treatment and intervention. Leveraging machine learning and handwriting analysis, this project seeks to bridge this diagnostic gap. By utilizing a secondary dataset from the DARWIN repository, the project develops machine learning models to distinguish between healthy individuals and Alzheimer's patients. Results from the LGBM classifier have shown promising performances on the default configuration of the model. But, in our tuned model out performed for the hyperparameter have been tuned. The main reason of the decrease in accuracy is because of the use of the subset which has a smaller number of features.

# Introduction

The contemporary healthcare arena confronts significant challenges presented by neurodegenerative diseases (NDs) such as Alzheimer's and Parkinson's. These illnesses progressively impair nerve cells, resulting in profound motor and cognitive deficits. Alzheimer's and Parkinson's stand as the most prevalent NDs, imposing substantial hardships on individuals and their families. Traditional diagnostic methods often grapple with providing timely and accurate identification due to the intricate and subtle nature of ND symptoms. Parkinson's primarily affects motor functions, leading to tremors and mobility challenges, while Alzheimer's leads to memory loss and cognitive decline. With an aging population, the instances of ND cases are projected to rise, necessitating innovative strategies for timely detection. Recent progress in machine learning and data analysis presents novel solutions to these diagnostic complexities. Python, a versatile programming language, empowers the creation of advanced machine learning models. By harnessing accessible data, researchers explore avenues to employ machine learning for the diagnosis of neurodegenerative diseases. An intriguing approach involves analyzing handwriting dynamics, a non-invasive and cost-effective method to assess disease progression. Through the application of machine learning, our aim is to capitalize on handwriting patterns, creating a predictive tool for the diagnosis of not only Alzheimer's but also other neurodegenerative conditions. This project aspires to construct a Python-based framework that utilizes machine learning to scrutinize handwriting data. This project propels machine learning into an integral role within healthcare, paving the way for a more accurate and accessible method of early ND diagnosis. Subsequent sections will delve into the intricacies of our methodology, experimental design, and results, highlighting handwriting analysis as a pivotal tool for the early diagnosis of Alzheimer's and other neurodegenerative conditions. By combining Python programming and machine learning, we aspire to create meaningful contributions to healthcare, ultimately fostering enhanced ND diagnosis and management protocols.

# Motivation of the Project

The decision to undertake this project arises from the pressing need to address the challenges posed by neurodegenerative diseases, specifically Alzheimer's disease. Neurodegenerative diseases have a profound impact on individuals and their families, leading to severe cognitive and motor impairments. With the prevalence of these diseases rising due to increased lifespans, there's a critical gap in early and accurate diagnosis. Conventional diagnostic methods have proven inadequate in providing timely identification of Alzheimer's disease, which greatly hampers effective treatment and intervention.

The motivation behind this project is rooted in the potential of leveraging machine learning techniques and handwriting analysis to bridge this diagnostic gap. Handwriting, being a complex integration of motor and cognitive processes, has the potential to serve as a valuable indicator of neurodegenerative disease progression. By harnessing the power of machine learning, we aim to create a predictive tool that enhances the identification and management of Alzheimer's disease. The lack of effective diagnostic methods, coupled with the rising impact of Alzheimer's, underscores the necessity of pursuing innovative approaches like ours to make a meaningful contribution to the field of healthcare.

**Benefits to Project Stakeholders and Society:**

**Benefits to the Project Team:** Engaging in this project offers numerous benefits to the project team. Firstly, it provides an opportunity to contribute to cutting-edge research in the intersection of machine learning and healthcare. It offers a chance to develop expertise in a highly relevant and impactful field, positioning team members as specialists in a growing domain. The project also nurtures collaboration between multidisciplinary experts, fostering skill development and a deeper understanding of neurodegenerative diseases, machine learning, and data analysis. Additionally, successful completion of the project can lead to publication opportunities and recognition within the scientific community.

**Benefits to Society:** The societal benefits of this project are far-reaching. Firstly, the development of an effective predictive tool for Alzheimer's disease aids medical practitioners in early and accurate diagnosis. This, in turn, enables timely interventions that can slow disease progression and enhance the quality of life for patients. The project also contributes to advancing the field of neurodegenerative disease research, potentially leading to breakthroughs in understanding disease mechanisms and treatment strategies. Furthermore, the open-source nature of the Darwin dataset and the methodologies developed have the potential to empower other researchers and institutions to make similar advancements. Collectively, this project stands to improve healthcare outcomes for individuals affected by Alzheimer's disease while fostering progress in the broader field of medical research.

# Objective of the project

The primary objective of this project is to develop a robust and innovative framework that leverages machine learning techniques and handwriting analysis to facilitate early and accurate diagnosis of neurodegenerative diseases, with a specific focus on Alzheimer's disease. The project aims to address the pressing challenges posed by the increasing prevalence of neurodegenerative diseases and the limitations of traditional diagnostic methods. By harnessing the power of Python programming and machine learning, the project endeavors to achieve the following objectives:

**Machine Learning Model Development:** Utilizing the DARWIN dataset and the extracted features, the project aims to design and train machine learning classifiers. These classifiers will learn to recognize patterns indicative of neurodegenerative diseases within handwriting dynamics. The goal is to create accurate and reliable models capable of distinguishing between healthy individuals and Alzheimer's patients.

**Validation and Performance Assessment:** The project will rigorously validate the effectiveness of the developed machine learning models using appropriate performance metrics. By assessing the models' sensitivity, specificity, accuracy, and other relevant measures, the project aims to demonstrate the reliability of the predictive tool in diagnosing Alzheimer's disease.

**Contribution to Healthcare and Research:** Through successful implementation of the project's methodology, the ultimate objective is to contribute to the field of healthcare and medical research. The project strives to provide medical practitioners with an accessible, non-invasive, and timely diagnostic tool for neurodegenerative diseases. Additionally, the open-source nature of the DARWIN dataset and the methodologies employed can empower other researchers to advance similar diagnostic and research efforts in the realm of neurodegenerative diseases.

**Advancement of Expertise and Collaboration:** The project aims to nurture expertise in machine learning, data analysis, and neurodegenerative diseases within the project team. Collaboration with multidisciplinary experts will facilitate the development of specialized skills and a deeper understanding of the complex interplay between medical science and technology.

By achieving these objectives, the project aspires to bridge the diagnostic gap for neurodegenerative diseases, particularly Alzheimer's, offering a transformative tool that aids in early detection, improved patient care, and enhanced medical research.

# Methodology



Figure-1: Working procedures of the project

## Data Collection

The Data for this project is secondary dataset and it is collected from Uc Irvine machine learning repository. Which is a popular repository to access dataset. Here is the URL link to download the dataset.

URL: <https://archive.ics.uci.edu/dataset/732/darwin>

Data Information: The DARWIN dataset contains handwriting data collected according to the acquisition protocol described in which is composed of 25 handwriting tasks. The protocol was specifically designed for the early detection of Alzheimer’s disease (AD). The dataset includes data from 174 participants (89 AD patients and 85 healthy people). The file “DARWIN.csv” contains the acquired data. The file consists of one row for each participant plus an additional header row. The first row is the header row, the next 89 rows collect patients’ data, whereas the remaining 84 rows collect information from healthy people. The file consists of 452 columns. The first column shows participants’ identifiers, whereas the last column shows the class to which each participant belongs

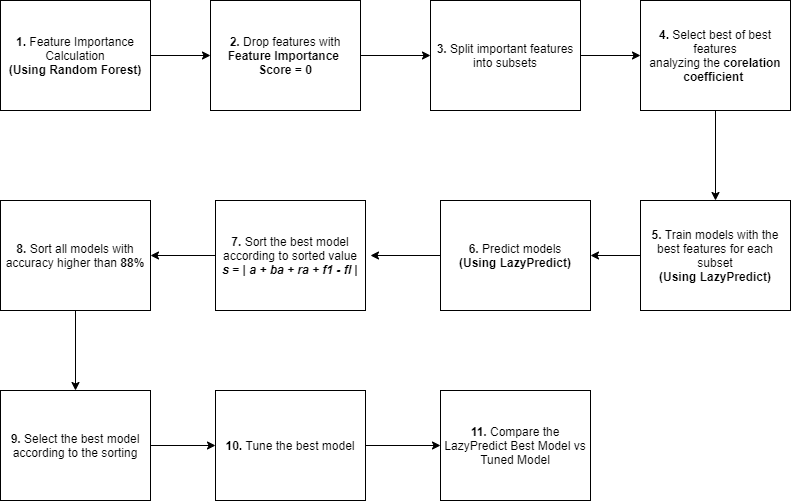
## Data processing

In the data preprocessing phase, the provided dataset is loaded into a Pandas DataFrame. The 'ID' column is dropped, resulting in a DataFrame named 'df' containing the remaining data. Missing values are checked and found to be absent in the dataset. The structure of the data is then summarized using the ‘info()’ function, revealing that the dataset consists of 174 entries and 451 columns with various data types including float64, int32, and int64. The independent variables (X) are extracted by dropping the 'class\_encoded' column, and the dependent variable (y) is assigned the 'class\_encoded' column. The output showcases the initial rows of the scaled data, highlighting the transformation achieved through scaling. The preprocessing steps involve data loading, removal of unnecessary columns, scaling of features, and preparation of independent and dependent variables for subsequent machine learning analysis.

## Dataset description

The dataset consists of 174 instances (rows) and 451 columns (features). The columns represent various attributes related to handwriting dynamics, with the following types of data: 300 columns of float64 data type, 150 columns of int64 data type, and 1 column of int32 data type. Each row corresponds to a unique instance, which in this context likely represents handwriting samples from individuals. The dataset includes features such as 'air\_time1', 'disp\_index1', 'gmrt\_in\_air1', 'gmrt\_on\_paper1', 'max\_x\_extension1', 'max\_y\_extension1', 'mean\_acc\_in\_air1', 'mean\_acc\_on\_paper1', 'mean\_gmrt1', 'mean\_jerk\_in\_air1', and so on, for a total of 451 features. Additionally, the dataset has a column labeled 'class\_encoded', which is the transformed numerical representation of the categorical 'class' variable. This dataset is intended for use in machine learning tasks, such as classification, to predict the 'class\_encoded' values based on the provided features.

## Machine Learning model development and evaluation



**Figure-2**: Model Development Diagram

In Figure-2, the model suggests that the first step is the feature importance calculation following an embedded feature selection method of Random Forest Classifier to get the best features according to the feature importance score. In step-2 all the features with the feature importance score zero (0) has been dropped.

In step-3 all the remaining important features have been split into subsets to find out the best features. Subsets have been created in terms of the following formula:

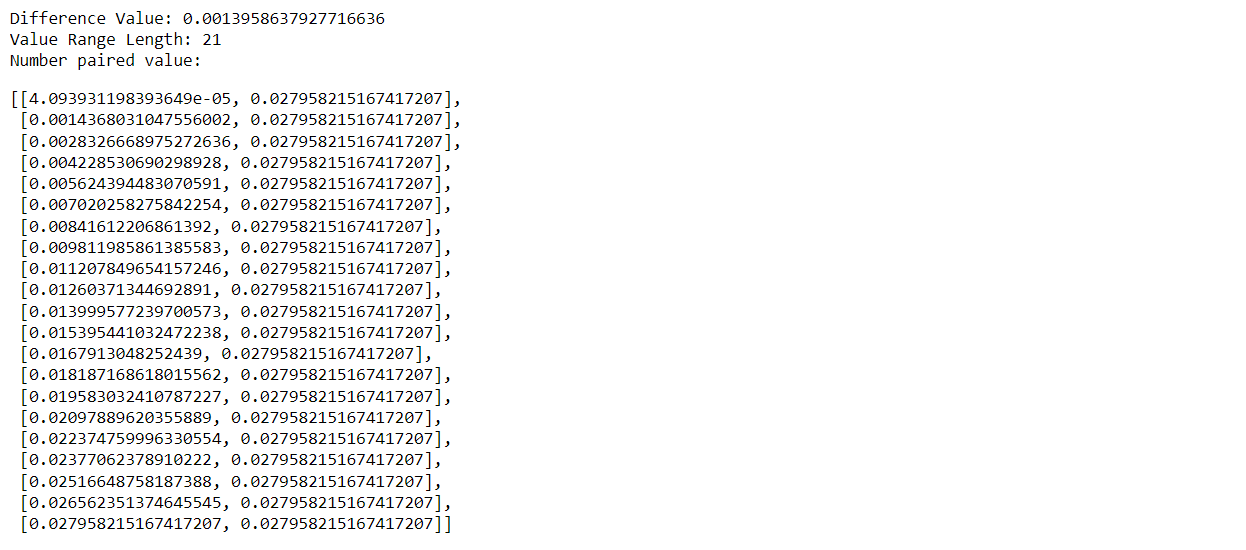
*for each iteration:  
 (min\_imp\_score - increment) to max\_imp\_score*

**Here,**

*min\_imp\_score =* minimum importance score

*max\_imp\_score =* maximum importance score

*increment* = (*max\_imp\_score - min\_imp\_score)/20*



**Figure-3**: Splitting important features into subset

Selecting the best of best features analyzing the correlation coefficient took place in step-3 and feature columns with correlation coefficient higher than 9.0 between them one column have been dropped as high coefficient means two columns with the same impact on the output.

In step-5 models have been trained using ‘LazyPredict’ and reduce the list if feature have less than 2 feature and in step-6 prediction accuracies have been tested for those models. Among all the models the best models have been sorted according to the following formula in step-7-

*s = | a + ba + ra + f1 - fl |*

**Here,**

***s =*** sorted value

***a =*** accuracy score of the model

***ba =*** balance accuracy score of the model

***ra =*** roc auc score of the model

***f1 =*** f1-score of the model

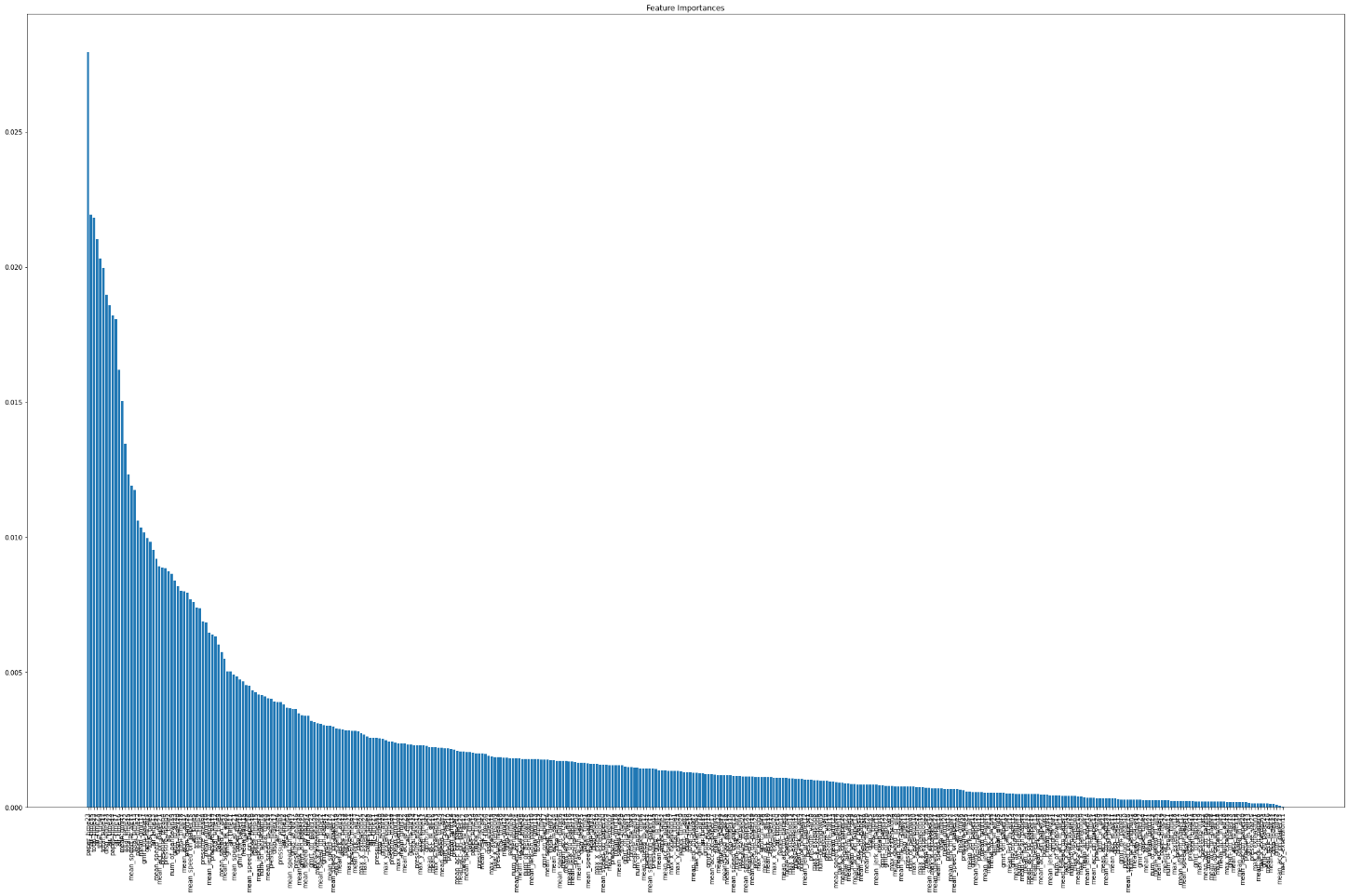
***fl =*** feature length of the subset

According to the sorted value, we get the best model in terms of best classification report for minimum number of features.

In step-8, the best models have been filtered by which models with accuracy higher than 88%, have been granted for comparison. In this step we get the sorted result of models with best accuracy and less features.

Step-9 is to select the best model according to the sorting and in step-10 the best model has been tuned to get higher accuracy. In the last step a comparison between the best model and tuned best model have been performed to evaluate the model.

## Results



**Figure-4:** Important features

**Feature importance calculation:**

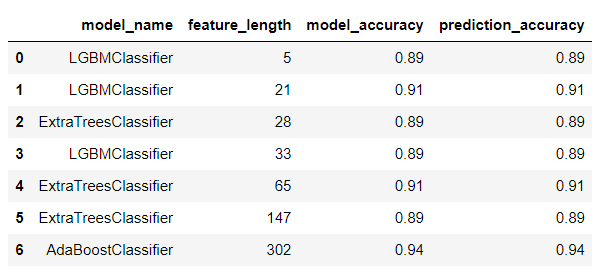
Feature selection using ‘Random Forest’ built in feature selection method have given a range of importance scores from 0.027958215167417207 to 0.00. Features with important score 0 have dropped and we got maximum important score 0.027958215167417207 and minimum importance score 4.093931198393649e-05.

**Splitting the important features into subsets:**

After following the formula of the proposed model to spilt the feature set we got 21 subsets after dropping the subsets less than two (2) features.

**Sorting the best models:**

According to the sorting formula and the condition of being the model accuracy to be higher than 88% we get six (6) models. Among them LGBM have been performed the best accuracy with less feature length. So, we select the LGBM as the best model.



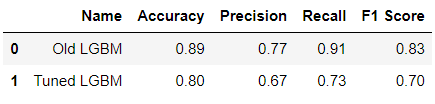
**Figure-5:** Best sorted model

**Tuning the best model:**

The LGBM model has been tuned using the grid search and tuning the hyper parameters: ‘n\_estimator’ and ‘max\_depth’ which is respectively 3 and 40

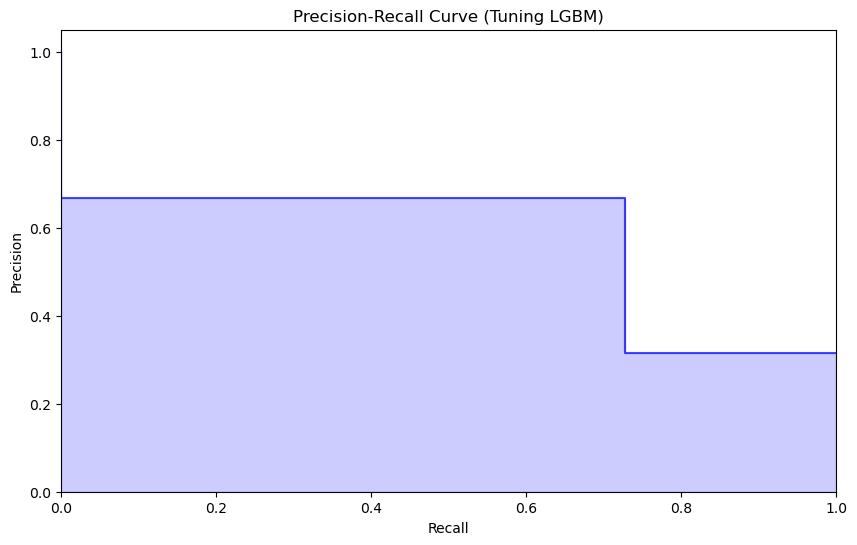
**Comparison of the Old LGBM and Tuned LGBM:**

Our findings after the comparison of the ‘Old LGBM’ that we predicted first without any tuning, with the ‘Tuned LGBM’ is that the accuracy has been dropped in our tuned model while testing the model. This is normal as ‘GridSearchCV’ do cross validation under the hood. So, the accuracy compare to old one is good.



**Figure-6:** Comparison result of old vs tuned model

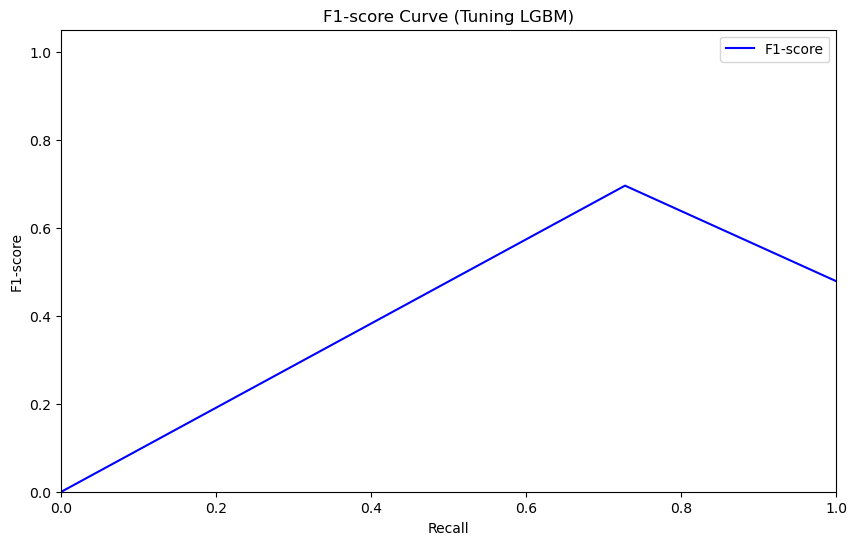
**Classification report of the tuned model:**



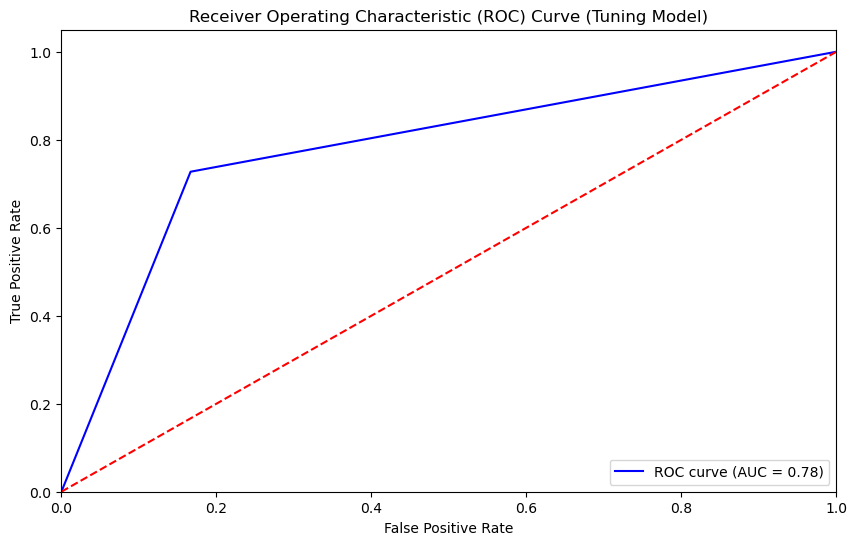
**Figure-7:** Precision-recall curve of tuned LGBM model

From Figure-7 we observe that the precision for class 1 (positive class) is approximately 0.67, indicating that when the classifier predicts a positive class, it is correct around 67% of the time. This means that there are some false positives, where the classifier predicts a positive outcome when it's actually negative.

The recall for class 1 is approximately 0.73, which means that the classifier is able to identify around 73% of the actual positive instances. This is also known as the true positive rate.

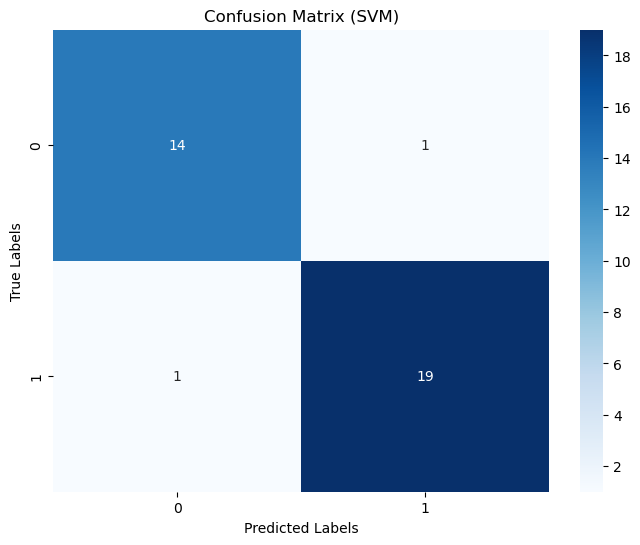


**Figure-8:** F1-score curve of tuned LGBM model



**Figure-9:** ROC curve of tuned LGBM model

In **Figure-9,** ROC curve (AUC) of 0.78 suggests that the tuned model exhibits average discriminatory power, effectively distinguishing between positive and negative instances. A value close to 1 indicates a strong ability to differentiate classes, and in this case, it implies that tuned model performs lower than the ‘Old LGBM’ in terms of separating the two classes.



In Figure-10, the confusion matrix depicts the model's classification results for healthy (0) and patient (1). In the context of this matrix, rows correspond to the actual classes, while columns represent the predicted classes. The LGBM correctly classified 20 instances as belonging to class 0 (true negatives) and 8 instances as class 1 (true positives), while misclassifying 4 instances of class 0 as class 1 (false positive) and 3 instances of class 1 as class 0 (false negative).