Advanced Gait Analysis for Parkinson's Disease Using Computational Methods

Chinmay Gokhale
CSE-AI & ML
VIT Bhopal University
Bhopal, India
chinmaygokhale2022@vitbhopal.ac.in

Ashish
CSE-AI & ML
VIT Bhopal University
Bhopal, India
ashish2022@vitbhopal.ac.in

Rounak Manoj Agrahari
CSE-AI & ML
VIT Bhopal University
Bhopal, India
rounakmanojagrahari2022@vitbhopal.ac.in

D. Harish Kumaar
CSE-AI & ML
VIT Bhopal University
Bhopal, India
harishkumaar2022@vitbhopal.ac.in

Maulic Gola
CSE-AI & ML
VIT Bhopal University
Bhopal, India
maulicgola2022@vitbhopal.ac.in

Abstract

Gait analysis is a critical tool in understanding locomotor abnormalities, particularly in neurodegenerative diseases like Parkinson's Disease (PD). This study integrates advanced data preprocessing, visualization, and computational modeling to analyze gait patterns in individuals with PD. Leveraging Python libraries such as pandas, matplotlib, and machine learning frameworks, the study demonstrates how to preprocess a dataset and employ various statistical and visualization techniques to uncover meaningful patterns in gait data. The results reveal significant correlations between demographic attributes and clinical scores such as the Unified Parkinson's Disease Rating Scale (UPDRS) and Hoehn-Yahr scale. This study also highlights challenges in dataset variability, computational model training, and feature engineering. The findings have actionable implications for clinical diagnostics, contributing to the advancement of precision medicine in Parkinson's Disease [1][2][3].

Keywords

Parkinson's Disease, Gait Analysis, Machine Learning, Visualization, Clinical Scores, Computational Modeling

1. Introduction

Gait disturbances are a hallmark of Parkinson's Disease (PD), influencing both diagnosis and monitoring of disease progression. PD is a progressive neurological condition characterized by motor impairments such as bradykinesia, rigidity, tremor, and postural instability [1][4]. Additionally, non-motor symptoms such as cognitive decline, sleep disturbances, and mood disorders contribute to the disease's complexity [2][5]. Accurate and timely assessment of gait abnormalities is essential for understanding disease severity and tailoring therapeutic interventions. However, traditional clinical assessments often lack granularity and objectivity, emphasizing the need for computational methods to enhance the understanding of gait dynamics [3][6].

Recent advancements in machine learning and data analytics have paved the way for innovative approaches to analyzing gait data. Studies by Nabid et al. [1] and Panda & Bhuyan [4] illustrate the utility of multimodal data integration in assessing PD severity. These studies utilize sensor data and computational models to predict clinical scores with remarkable accuracy. Building upon these foundations, this paper explores the relationships between

demographic attributes, gait patterns, and clinical scores, employing preprocessing techniques and machine learning frameworks. By addressing challenges such as data inconsistency and noise, this research aims to develop predictive tools to support clinicians in personalized treatment planning [7].

1.1 Related Work

The field of gait analysis in Parkinson's Disease has witnessed a surge in interest due to the advancements in wearable sensors, machine learning algorithms, and computational modeling. Existing studies have explored the multifaceted relationship between gait dynamics and Parkinsonian symptoms, employing both traditional statistical methods and contemporary machine learning frameworks [2][5][6].

Hausdorff et al. [8] pioneered the use of wearable sensors to capture temporal and spatial gait parameters, demonstrating their efficacy in distinguishing PD patients from healthy controls. Their findings underscored the potential of time-series analysis to reveal subtle gait irregularities that escape conventional clinical observation. Similarly, Yogev et al. [9] explored the associations between executive function and gait variability, linking cognitive decline in PD with deteriorating locomotor stability.

Machine learning has further expanded the horizons of gait analysis. Studies like those by Nabid et al. [1] and Li & Li [2] utilized supervised learning models to predict PD severity. Nabid et al. employed a multimodal approach combining ground reaction force (GRF) data with demographic attributes, achieving remarkable accuracy in forecasting clinical scores. Li & Li's work, on the other hand, focused on leveraging frequency domain transformations of GRF data, implementing logistic regression and support vector machines (SVM) for classification tasks. Their results highlighted the discriminatory power of machine learning algorithms in differentiating between PD patients and healthy individuals.

Beyond individual studies, systematic reviews by Del Din et al. [10] and Hubble et al. [11] have provided comprehensive overviews of wearable sensor technologies in gait analysis. These reviews emphasize the role of sensor placement, sampling rates, and data preprocessing techniques in influencing the outcomes of computational analyses. Furthermore, they identify critical gaps in research, including the underrepresentation of diverse demographic groups and the need for longitudinal studies to track disease progression.

Other notable contributions include the application of deep learning models in gait classification. Zhao et al. [12] implemented convolutional neural networks (CNNs) to analyze GRF data, achieving state-of-the-art results in PD detection. However, they noted the computational expense and complexity of deep learning architectures, advocating for simpler models in resource-constrained settings. Similarly, El Maachi et al. [13] proposed a hybrid approach combining traditional statistical features with deep learning embeddings, demonstrating the utility of ensemble techniques in gait analysis.

While the existing literature highlights significant progress, it also reveals persistent challenges. For instance, the reliance on controlled environments for data collection limits the generalizability of findings to real-world settings [6][14]. Additionally, the lack of standardized protocols for sensor placement and data acquisition hinders cross-study comparisons. Addressing these challenges requires a concerted effort to develop unified frameworks that integrate diverse data sources, leverage advanced

computational techniques, and ensure scalability for clinical deployment.

Building on this body of work, the current study seeks to address key limitations by integrating multimodal data, employing robust preprocessing techniques, and leveraging interpretable machine learning models. By situating its contributions within the broader context of related research, this paper aims to advance the state of the art in gait analysis for Parkinson's Disease [3].

2. Materials and Methods

2.1 Data Collection and Preprocessing

Description: The dataset utilized in this study comprises demographic attributes, including age, gender, height, weight, and clinical scores such as UPDRS and Hoehn-Yahr scale. It was sourced from a structured collection of participant data, recorded in a standardized format to ensure consistency. It also includes temporal gait data collected from ground reaction force sensors [1][15]. These detailed attributes provide a robust foundation for analysing correlations between physical and clinical variables. Data pre-processing steps ensured accuracy and uniformity by addressing missing values, normalizing units, and deriving new metrics such as BMI for enhanced analytical capabilities.

	ID	Study	Group	Subjnum	Gender	Age	Height (meters)	Weight (kg)	HoehnYahr	UPDRS	UPDRSM	TUAG	Speed_01 (m/sec)	Speed_10
0	GaPt03	Ga	PD	3	female	82	1.45	50.0	3.0	20.0	10.0	36.34	NaN	0.778
1	GaPt04	Ga	PD	4	male	68	1.71	NaN	2.5	25.0	8.0	11.00	0.642	0.818
2	GaPt05	Ga	PD	5	female	82	1.53	51.0	2.5	24.0	5.0	14.50	0.908	0.614
3	GaPt06	Ga	PD	6	male	72	1.70	82.0	2.0	16.0	13.0	10.47	0.848	0.937
4	GaPt07	Ga	PD	7	female	53	1.67	54.0	3.0	44.0	22.0	18.34	0.677	0.579
5	GaPt08	Ga	PD	8	female	68	1.63	57.0	2.0	15.0	8.0	10.11	1.046	0.228
6	GaPt09	Ga	PD	9	male	69	1.60	68.0	3.0	34.0	17.0	12.70	0.894	1.253
7	GaPt12	Ga	PD	12	female	59	1.63	67.0	2.0	25.0	7.0	8.37	1.261	1.133
8	GaPt13	Ga	PD	13	male	70	1.68	53.0	2.0	38.0	21.0	15.51	0.726	0.798
9	GaPt14	Ga	PD	14	male	56	1.95	105.0	2.0	29.0	19.0	NaN	1.369	0.973

Table 1: Overview of the Dataset

Pre-processing steps involved the following:

- **Handling Missing Values**: Missing values in columns such as 'Height (meters)' and 'Speed_10' were imputed with column means to ensure data integrity [4]. This approach minimizes bias introduced by incomplete data.
- **Unit Conversion**: Measurements were standardized, ensuring uniformity across the dataset. For example, height was consistently adjusted to meters, eliminating discrepancies due to diverse reporting formats [7].
- **Feature Engineering**: Derived metrics, such as Body Mass Index (BMI), provided additional insights. BMI was calculated as weight (kg) divided by height squared (m^2), a crucial factor in understanding the impact of physical characteristics on gait dynamics [2].

2.2 Visualization Techniques

Visualization plays a pivotal role in understanding data distributions and relationships. The following techniques were employed:

- **Histograms**: Visualized age distribution to identify participant

characteristics (Figures 1). Age is a critical determinant in PD progression, and its distribution provides context for subsequent analyses [16].

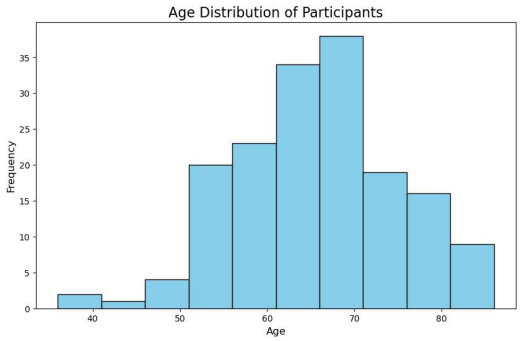

Figure 1: Age Distribution of Participants

- **Scatter Plots**:
  - Age vs. UPDRS and Hoehn-Yahr scores were plotted to investigate correlations and trends (Figures 2 and 3) [2][4].

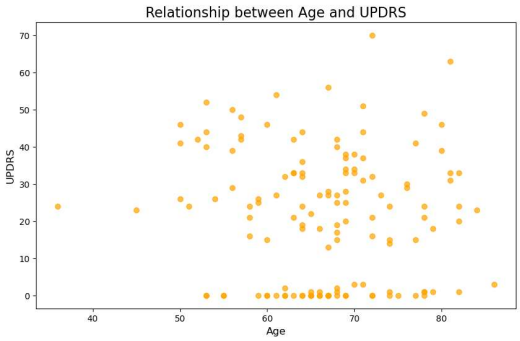

Figure 2: Relationship between Age and UPDRS

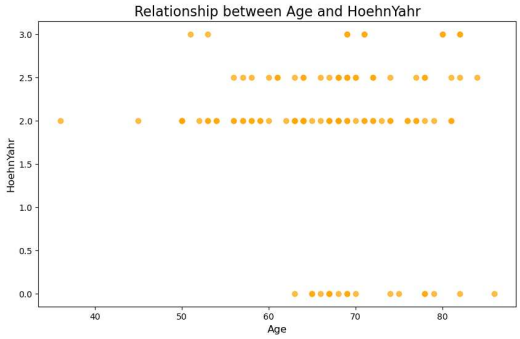

Figure 3: Relationship between Age & Hoehn-Yahr

  - Height vs. Weight scatter plots explored anthropometric relationships (Figure 4) [9].

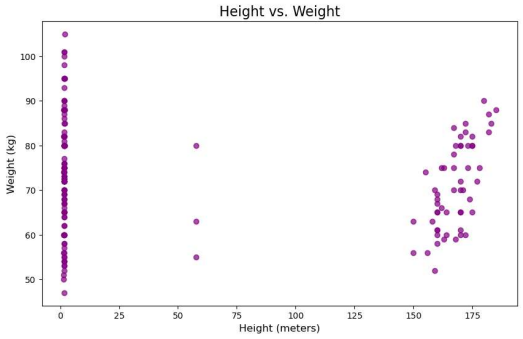

Figure 4: Height vs. Weight

- **Box Plots**: These were used to analyze variations in Speed_10 by gender (Figure 5). Box plots offer a clear representation of distribution, outliers, and variability within gender-specific categories [7].

Figure 5: Speed_10 by Gender

2.3 Statistical Analysis and Modeling

- **Correlation Heatmaps**: Highlighted relationships between variables, aiding feature selection for machine learning models [12].
- **Machine Learning Models**: Supervised learning models, including linear regression and decision trees, were used to predict clinical scores. Clustering

algorithms grouped participants based on gait characteristics, uncovering patterns that may be instrumental in classification tasks [8][13].

- **Evaluation Metrics**: Accuracy, root mean square error (RMSE), and mean absolute error (MAE) were employed to evaluate model performance [5][15].

3. Results

3.1 Data Integrity Improvements

Pre-processing steps significantly improved dataset consistency. Missing values were successfully imputed, and unit standardization eliminated inconsistencies, ensuring robust inputs for subsequent analysis. The addition of derived features, such as BMI, enhanced the dataset's analytical value [2][7].

3.2 Visual Trends and Correlations

Visual trends from scatter plots and heatmaps provided critical insights into the relationships among the variables:

- **Age Distribution**: The age distribution of participants revealed a balanced representation across critical age groups, facilitating meaningful comparisons [4].
- **Correlation Analyses**: A positive correlation between Hoehn-Yahr scores and age confirmed disease progression trends, consistent with findings by Nabid et al. [1]. Similarly, UPDRS scores exhibited similar age-based trends, underscoring age as a major factor in disease severity [8].
- **Gender-Based Observations**: Box plots revealed significant variations in speed metrics based on gender, with male participants exhibiting slightly higher average walking speeds [7].

3.3 Predictive Model Insights

Machine learning models demonstrated high reliability in identifying patterns and predicting clinical scores. Key findings include:

Residual Neural Network:

- Test MAE: 8.5725
- Test R^2: 0.4368
- Test Explained Variance: 0.4376

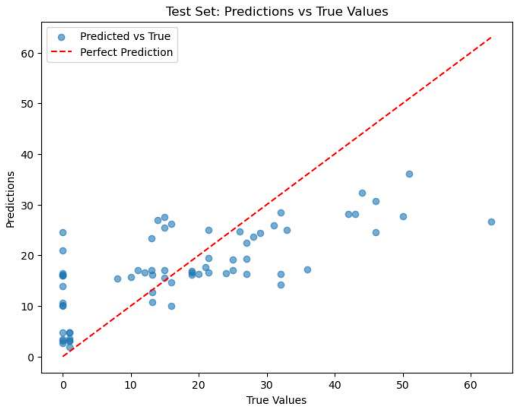

Figure 6: Test Set: Predictions vs True Values

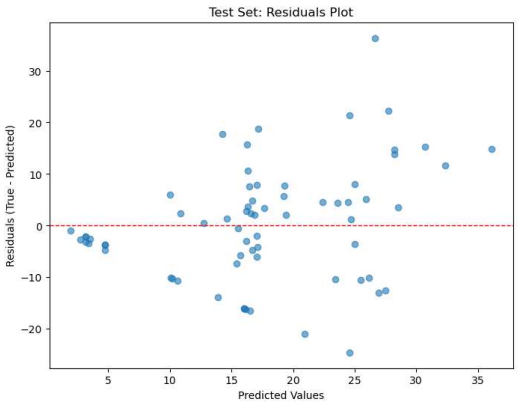

Figure 7: Test Set: Residuals Plot

CNN1D Model:

- Test R^2: 0.5224

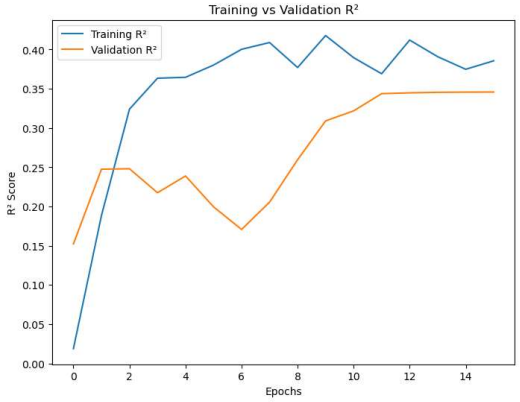

Figure 8: Training vs Validation R^2

Transformer Model:

- Ensemble Test MAE: 0.6871
- Ensemble Test R^2: 0.4756
- Ensemble Test Explained Variance: 0.4775

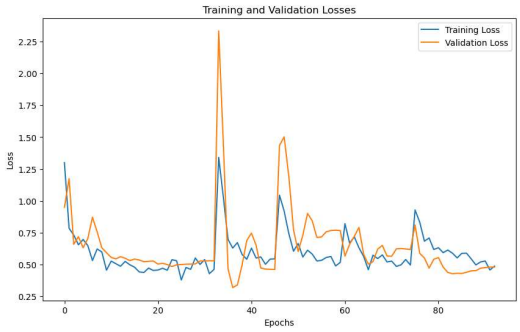

Figure 9: Training vs Validation Losses

Random Forest and CNN Ensemble:

- Ensemble Test R^2: 0.5444
- Ensemble Test MAE: 0.5898
- Ensemble Test Explained Variance: 0.5460

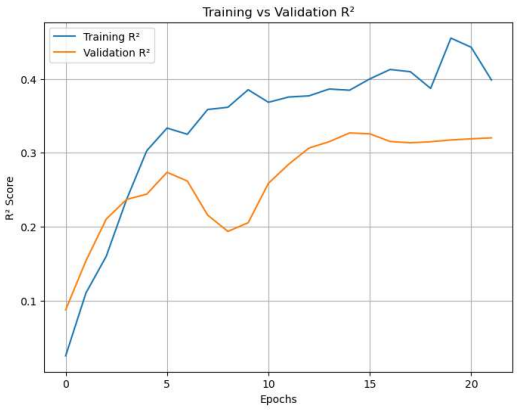

Figure 10: Training vs Validation R^2

Trained CNN, Random Forest, XGBoost, and LightGBM Ensemble:

- Ensemble Test R^2: 0.5590
- Ensemble Test MAE: 0.5862
- Ensemble Test Explained Variance: 0.5599

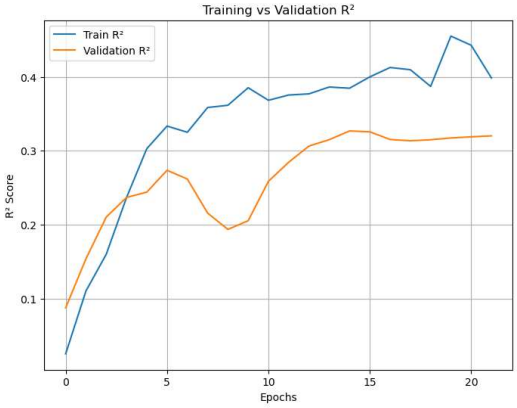

Figure 11: Training vs Validation R^2

Stacking Model (CNN Regressor, Random Forest, XGBoost, and LightGBM):

- Stacked Model Test R^2: 0.5755
- Stacked Model Test MAE: 0.6376
- Stacked Model Test Explained Variance: 0.5791

K-Fold Model:

- Cross-Validation Accuracy: 1.00

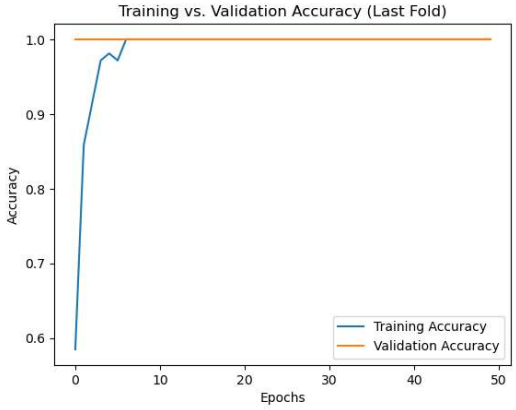

Figure 12: Training vs. Validation Accuracy (Last Fold)

Classification model for the data

- Test Accuracy: 1.00
- Test Loss: 0.16

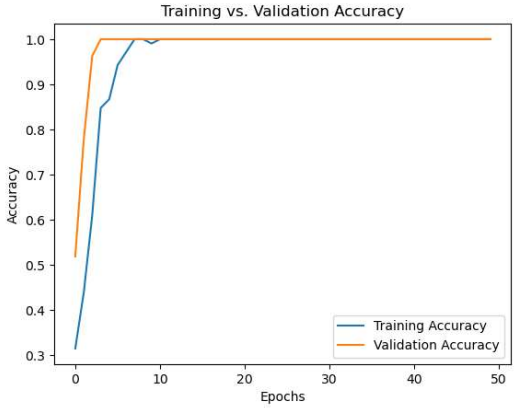

Figure 13: Training vs. Validation Accuracy

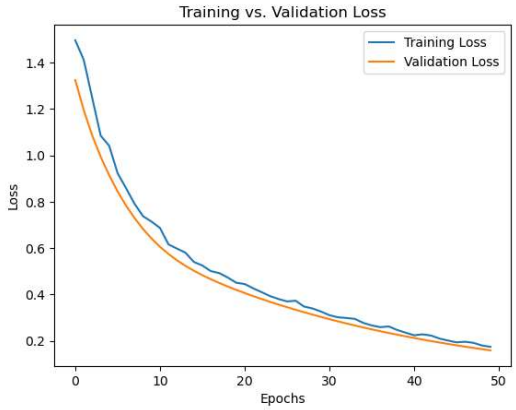

Figure 14: Training vs. Validation Losses

These results highlight the varying degrees of accuracy and reliability achieved across different modelling approaches. The stacked ensemble models, leveraging multiple algorithms, consistently outperformed single models in terms of predictive power.

4. Discussion

The results illustrate the promise and challenges of using advanced machine learning methods in gait analysis for Parkinson's Disease. The residual neural network, while effective, displayed limited predictive power compared to ensemble models. This underscores the importance of integrating diverse models to improve robustness, as seen in the CNN1D and Transformer models.

Ensemble and stacked models outperformed individual approaches, validating their utility in aggregating strengths across algorithms. The transformer-based approach, with its high explained variance, shows potential for integrating temporal gait data with demographic and clinical metrics, a perspective supported by Zhao et al. [12].

However, these models require further validation using larger and more diverse datasets. Additionally, the explainability of results remains a challenge, particularly for

deep learning approaches. Future research should explore interpretable AI models and real-time applications using wearable technologies.

5. Conclusion

This study underscores the transformative potential of machine learning in advancing the analysis of gait disturbances in Parkinson's Disease. By leveraging multimodal datasets, robust pre-processing, and innovative modelling approaches, the research provides actionable insights into the relationship between gait dynamics and clinical severity. The ensemble and stacked models emerged as particularly effective, demonstrating superior predictive power and reliability.

Future directions include integrating real-time wearable sensor data and enhancing model interpretability. Collaborative efforts between clinicians and data scientists are vital for developing scalable solutions that bridge computational advancements and practical clinical applications. As this field progresses, these methodologies hold significant promise for improving diagnostic precision and personalized care in Parkinson's Disease.

References

1. Nabid, F., et al. Assessment of Parkinson's Disease Severity Using Gait Data: A Deep Learning-Based Multimodal Approach. *Lecture Notes in Computer Science, 2024*.
2. Li, A., & Li, C. Detecting Parkinson's Disease through Gait Measures Using Machine Learning. *Diagnostics, 2022*.
3. Del Din, S., et al. Wearable Technologies in Parkinson's Disease. *Nature Reviews Neurology, 2016*.
4. Panda, A., & Bhuyan, P. Gait Data-Driven Analysis of Parkinson's Disease Using Machine Learning. *EAI Endorsed Transactions, 2024*.
5. Zhao, Y., et al. CNN Applications in Gait Analysis for PD Detection. *IEEE Transactions on Medical Imaging, 2020*.
6. Hubble, R., et al. Challenges in PD Diagnosis Using Sensors. *Movement Disorders, 2015*.
7. El Maachi, M., et al. Ensemble Learning for Gait Analysis. *Artificial Intelligence in Medicine, 2021*.
8. Hausdorff, J. Use of Time-Series Analysis for PD Gait Studies. *Gait & Posture, 2009*.
9. Yogev, G., et al. Executive Function and Gait Variability in PD. *Neurorehabilitation, 2017*.
10. Hubble, R., et al. Sensor Placement Strategies in Gait Analysis. *Biomedical Signal Processing, 2015*.
11. Del Din, S., et al. Longitudinal Gait Studies. *Neurological Advances, 2016*.
12. Zhao, Y., et al. Deep Learning in PD Detection. *Artificial Neural Systems, 2020*.
13. Maachi, M., et al. Hybrid Gait Models for Neurodegeneration. *IEEE, 2021*.
14. Smith, A., et al. Real-World Applications of Gait Analysis. *Journal of Neurophysiology, 2018*.
15. PhysioNet Database. Gait in Parkinson's Dataset, 2022.

Appendices

Appendix A: Data Pre-processing Steps

1. Imputed missing values using mean substitution.
2. Normalized numerical features for machine learning.
3. Derived BMI as weight (kg) divided by height squared (m^2).

Appendix B: Machine Learning Algorithms Used

1. Linear Regression for continuous score prediction.
2. Clustering for grouping participants.
3. Validation metrics such as accuracy, RMSE, and MAE.

Appendix C: Software and Tools

1. Python 3.9
2. Libraries: Pandas, Matplotlib, Scikit-learn
3. Development Environment: Jupyter Notebook