# Methodology

## Introduction

Briefly introduce the methodology section, explaining its importance in the research. Mention what the section will cover, such as the approaches, techniques, and procedures used to collect and analyze data.

## Research Design

Outline the overall structure and plan of the research. Explain whether the research is qualitative, quantitative, or mixed methods. Specify whether it is experimental, descriptive, exploratory, or correlational in nature.

## Data Collection Methods

Describe how the data was collected for the study.

-Primary Data: Interviews, surveys, experiments, or direct observations.

-Secondary Data: Pre-existing datasets, academic literature, reports, or online sources.

-Tools and Instruments: Mention tools like questionnaires, sensors, or web scraping techniques.

## Data Sources

Detail where the data comes from. Discuss whether the data is sourced from public databases (e.g., Kaggle, UCI Machine Learning Repository), private company data, or user-generated data.

## Data Preprocessing

Explain how the raw data is cleaned and prepared for analysis.

- Data cleaning (handling missing data, removing outliers).

- Data transformation (normalization, standardization).

- Feature selection and feature extraction.

## Model Selection/Algorithm Choice

Justify the choice of models or algorithms used in the research.

- Explain why certain machine learning or statistical models were chosen (e.g., Random Forest, SVM, CNN).

- Mention relevant metrics for evaluation (accuracy, precision, recall).

- Justify hyperparameter choices if applicable.

## Experimental Setup

Describe how the experiments or simulations are conducted.

- Computational environment (e.g., hardware, software, libraries).

- Experimental steps and parameters.

- How many trials or repetitions were conducted to ensure robustness.

## Evaluation Metrics

Define the metrics used to evaluate the performance of models or approaches.

- Accuracy, precision, recall, F1-score for classification.

- MSE (Mean Squared Error), RMSE for regression.

- AUC-ROC for binary classification.

## Validation Strategy

Explain how you ensured the results are valid and generalizable.

- Cross-validation: k-fold cross-validation, holdout methods.

- Split methods: Training and test dataset split ratio (e.g., 80-20 split).

- External validation: Comparison with baseline models or external data sources.

## Ethical Considerations

Address any ethical concerns related to the research.

- Consent and privacy (especially for human data).

- Bias mitigation.

- Compliance with data protection laws (e.g., GDPR).

## Limitations

Acknowledge any limitations in the methodology. Discuss potential weaknesses or constraints in data, algorithms, or experimental design that might affect the results.

## Research Design

This study aims to develop a deep learning model to detect ADHD in children based on handwriting features. The project follows an experimental design, where several machine learning models are tested and optimized to determine the most accurate and interpretable model for ADHD detection. Deep learning-based handwriting analysis have been used to identify specific patterns associated with ADHD and Explainable AI (XAI) techniques are applied to understand the model’s decision-making process.

## Data Collection and Description

The dataset used in this study consists of handwriting samples from two groups: children diagnosed with ADHD and a control group without ADHD. The dataset was collected from a publicly available repository of children's handwriting patterns in various tasks, such as writing letters and numbers. Each sample is represented by a set of features extracted from the writing, such as pen pressure, speed, curvature, and stroke order.

### Dataset Description

* Number of samples: 1,500 handwriting instances
* Features: Pen pressure, writing speed, pen-lift occurrences, stroke duration, and various geometric features (e.g., letter size, slant).
* Target: Binary classification (ADHD or non-ADHD)
* Data source: Public repository of children’s handwriting data (collected with parental consent).

## Data Preprocessing

To ensure the data was ready for model training, several preprocessing steps were taken

### Handling Missing Data

Handwriting samples with missing or incomplete feature data were removed. Approximately 2% of the samples were excluded.

### Normalization

All continuous features, such as writing speed and pen pressure, were normalized to a 0-1 scale using min-max normalization. This ensured that no feature disproportionately influenced the model.

### Feature Selection

Feature correlation was performed to remove redundant features. Features with a Pearson correlation above 0.90 were excluded to avoid multicollinearity.

### Label Encoding

The target variable (ADHD diagnosis) was label-encoded to 0 for non-ADHD and 1 for ADHD.

## Model Selection

Three machine learning models are evaluated based on their performance for the classification task

* 1. Convolutional Neural Network (CNN)

CNN was selected for its ability to capture spatial and temporal handwriting features. It was particularly effective in identifying visual patterns in the letter shapes, pressure points, and stroke sequences.

* 1. Support Vector Machine (SVM)

SVM was chosen due to its success in previous studies involving handwriting analysis and ADHD detection. It was used with a radial basis function (RBF) kernel.

* 1. Random Forest (RF)

Random Forest was selected to explore an interpretable model for feature importance analysis. As an ensemble learning method, it was helpful in handling complex decision boundaries while providing insight into feature contributions.

## Model Training and Evaluation

The dataset was split into training and testing sets using an 80/20 split. A 5-fold cross-validation was conducted on the training data to ensure that the model generalized well.

### Evaluation Metrics

The models were evaluated using accuracy, precision, recall, and F1-score to address any imbalance between the ADHD and non-ADHD classes.

### Training Process

CNN: The CNN model was trained using a batch size of 32, a learning rate of 0.001, and 50 epochs. Adam optimizer and categorical cross-entropy loss has been used.

SVM and Random Forest: These models were optimized using grid search to find the best hyperparameters.

## Explainable AI (XAI) Techniques

To interpret the predictions made by the CNN, the following XAI methods were used:

### Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM was used to visualize the parts of the handwriting samples that contributed most to the CNN's decision-making process. This allowed us to highlight areas of the handwriting that might indicate ADHD tendencies, such as specific stroke formations or inconsistencies in letter shapes.

### SHAP (SHapley Additive Explanations)

SHAP values were calculated for the Random Forest model to quantify the contribution of each feature (e.g., writing speed, pen pressure) to the final prediction. This helped in identifying the most important handwriting characteristics associated with ADHD.

## Deployment (Optional)

The final model, a CNN optimized for both performance and interpretability, was deployed using a Flask API. The model was integrated into a web-based application where clinicians can upload new handwriting samples for real-time ADHD prediction. The web app provides a diagnosis with a confidence score, and visual explanations (Grad-CAM heatmaps) are generated to show which parts of the handwriting contributed most to the model’s decision.