# **Methodology**

## 1. Data Collection

The dataset used for fine-tuning AyurGenix AI was sourced from a combination of **ancient Ayurvedic texts, government health websites, research papers, and synthetic data generation**. The primary sources include:

* **Books**: Charaka Samhita, Madhava Nidana, Ayurvedic Pediatrics, and other classical texts.
* **Government and Medical Websites**: Ayush.gov.in, Mayo Clinic, NHS Inform, MedicineNet, MedlinePlus, etc.
* **Public Datasets**: UCI Machine Learning Repository, Papers with Code.
* **Synthetic Data**: Created to enhance underrepresented cases, ensuring data diversity.

### *Figures & References*

* A figure showcasing the **data collection pipeline** (scraping, preprocessing, augmentation) can be added.
* References to Ayurvedic and modern medicine papers can be included to support data authenticity.

## 2. Data Preprocessing

Given the diversity of sources, multiple preprocessing steps were undertaken:

* **Text Cleaning**: Removal of non-standard characters, stopwords, and irrelevant text.
* **Normalization**: Standardizing terminology across different sources.
* **Feature Engineering**: Extracting structured information like **disease symptoms, risk factors, Ayurvedic herbs, treatments, and dosha classifications**.
* **Handling Missing Data**: Imputation techniques were used to maintain data integrity.

### *Figures & References*

* A sample **table** displaying key dataset attributes can be included.
* A flowchart illustrating the preprocessing steps would be helpful.

## 3. Model Selection & Fine-Tuning

Initially, simple models like **Random Forest and XGBoost** were tested, but due to data limitations, they resulted in overfitting and poor generalization. To improve accuracy and scalability, we shifted to **Large Language Models (LLMs)**. After evaluating various models, we selected **LLaMA 3.1 1B** based on the following criteria:

1. **Pre-trained knowledge**: Improved accuracy due to extensive prior training.
2. **Built-in APIs**: Facilitated **Natural Language Processing (NLP) tasks** like question answering.
3. **Computational feasibility**: Balanced between performance and available resources.

Fine-tuning was performed on our Ayurvedic dataset using **transfer learning**, optimizing it for domain-specific knowledge retention.

### *Figures & References*

* A diagram illustrating the **model selection process** (comparing XGBoost, Random Forest, LLaMA).
* Citations from papers discussing **LLM fine-tuning techniques** in healthcare AI.

## 4. System Architecture & Backend Integration

#### **System Architecture**

The AyurGenix AI platform is built on a three-tier architecture comprising a **frontend**, **backend**, and **machine learning model**. The frontend, developed using **Django**, **HTML**, and **CSS**, provides a user-friendly interface for inputting symptoms and viewing recommendations. The backend, powered by **PostgreSQL**, stores user profiles, symptom data, and remedy details. The **LLaMA 3.1 1B** model, fine-tuned for symptom analysis and severity assessment, serves as the core AI component.

* **Figure Suggestion**: Include a **system architecture diagram** showing the interaction between the frontend, backend, and ML model.
* **Reference Suggestion**: Cite papers or resources on Django, PostgreSQL, and LLaMA 3.1 to support your choice of technologies.

## 5. Evaluation Metrics

To assess model performance, the following metrics were used:

* **Accuracy & F1 Score**: To measure classification performance.
* **BLEU & ROUGE Scores**: Evaluating text generation quality for chatbot responses.
* **User Feedback & Expert Validation**: Ensuring Ayurvedic recommendations align with traditional knowledge.

### *Figures & References*

* A **table of performance metrics** comparing different models.
* References on **evaluating AI models in healthcare applications**.