Data Cleaning in Data Science

Data Cleaning in Data Science: A Comprehensive Guide

01 **GIGO effect**

02 **What is Data Cleaning?**

03 **Step-by-Step Data Cleaning Process**

04 **Common Data Issues**

05 **Data Cleaning Techniques**

06 **Data Cleaning Tools**

About Me

WORK EXPERIENCE

Current

HoD, M.Sc. Data Science(VITOL) , Link
Associate Professor, SCOPE
Vellore Institute of Technology, Chennai.

Past

Senior Assistant Professor,
SCSE, SASTRA University.

ACHIEVEMENTS

NASSCOM

- Certified Master Trainer, Associate Analytics

EDUCATION

- PhD, CSE, VIT Chennai.
- M.Tech, CSE, JNTU, Hyderabad.
- B.Tech, CSE, JNTU, Hyderabad.

Areas of Interest

- Natural Language Processing
- Machine Learning
- Data Analytics

CONTACT INFO

- email: tulasiprasad.sariki@vit.ac.in
- linkedin: www.linkedin.com/in/iamstp

GIGO effect

What is Data Cleaning?

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in datasets.

It involves:

- Removing duplicate or irrelevant observations
- Fixing structural errors
- Handling missing data
- Filtering out outliers
- Standardizing data formats
- Correcting typos or formatting issues

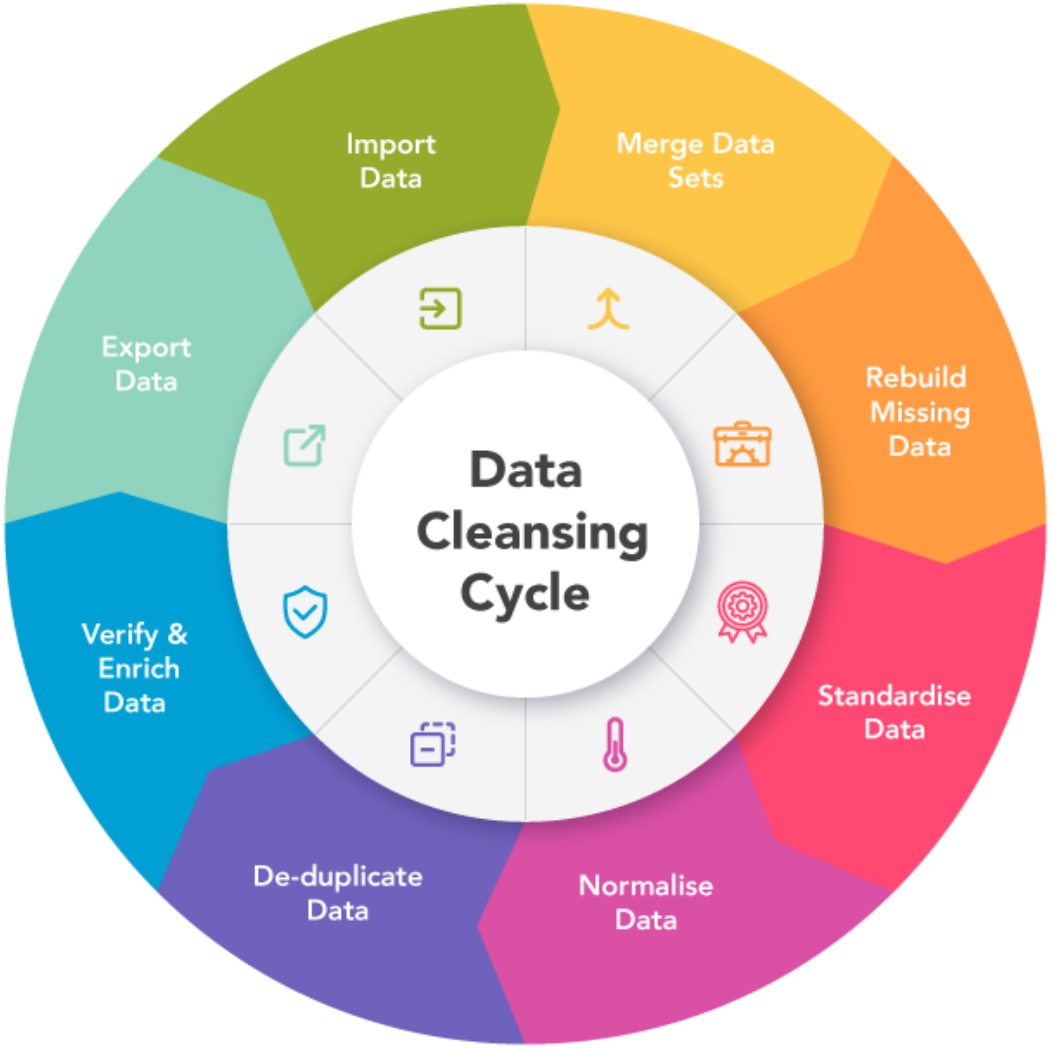
Step-by-Step Data Cleaning Process

- Understand the Data
- Make a Copy of the Raw Data
- Perform Initial Data Exploration
- Check Data Types and Structures
- Handle Missing Data
- Remove Duplicates
- Handle Outliers
- Standardize and Normalize

Step-by-Step Data Cleaning Process

- Correct Invalid Values
- Handle Structural Errors
- Validate and Cross-Check
- Handle Special Cases
- Document the Cleaning Process
- Create Automated Cleaning Scripts
- Perform a Final Review

Step-by-Step Data Cleaning Process



Common Data Issues

Missing Data

Duplicate Data

Inconsistent Formatting

Typos and Spelling Errors

Outliers

Inconsistent Units

Structural Errors

Encoded or Garbled Data

Inconsistent Naming Conventions

Data Type Mismatches

Truncated Data

Unnecessary Metadata

Inconsistent Aggregation

Data Cleaning Techniques

Handling Missing Data

- Deletion
- Imputation
- Using a Dedicated Category
- Advanced Techniques

Data Cleaning Techniques

Dealing with Duplicates

- Exact Matching
- Fuzzy Matching
- Composite Key Matching

Data Cleaning Techniques

Standardizing Formats

- Regular Expressions
- Parsing Libraries
- Lookup Tables

Data Cleaning Techniques

Correcting Typos and Misspellings

- Spell-Checking Algorithms
- Fuzzy String Matching
- Manual Correction

Data Cleaning Techniques

Handling Outliers

- Statistical Methods
- Visualization
- Domain Expertise
- Winsorization

Data Cleaning Tools

Python Libraries

R Packages

SQL

OpenRefine

Trifacta Wrangler

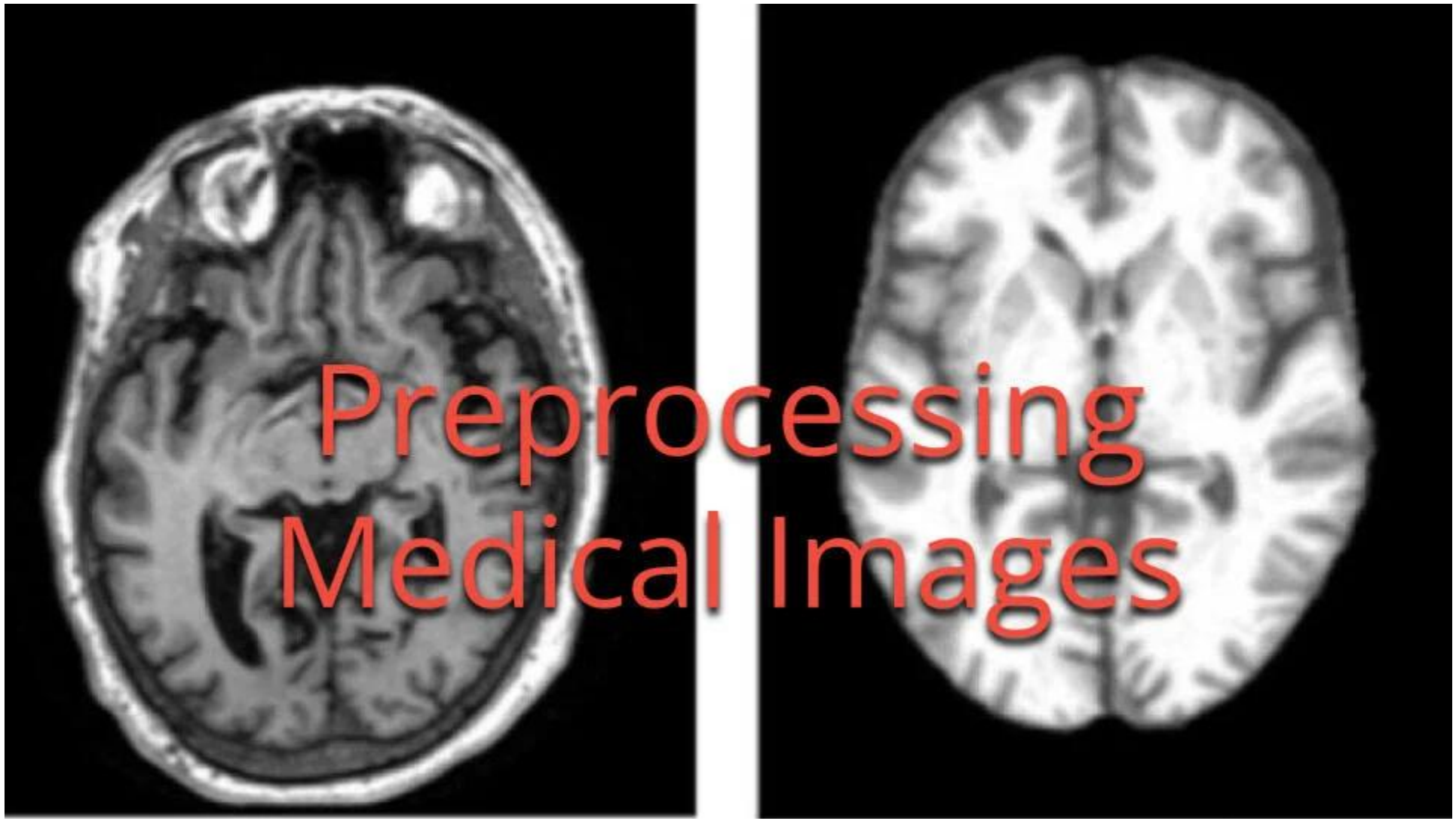
Talend Data Preparation

Excel

Google Cloud Dataprep

Alteryx

Databricks


Preprocessing
Medical Images

Medical Image Preprocessing

What is Medical Image Preprocessing?

Why is Preprocessing Essential in Medical Imaging?

Common Preprocessing Techniques

Advanced Preprocessing Workflows

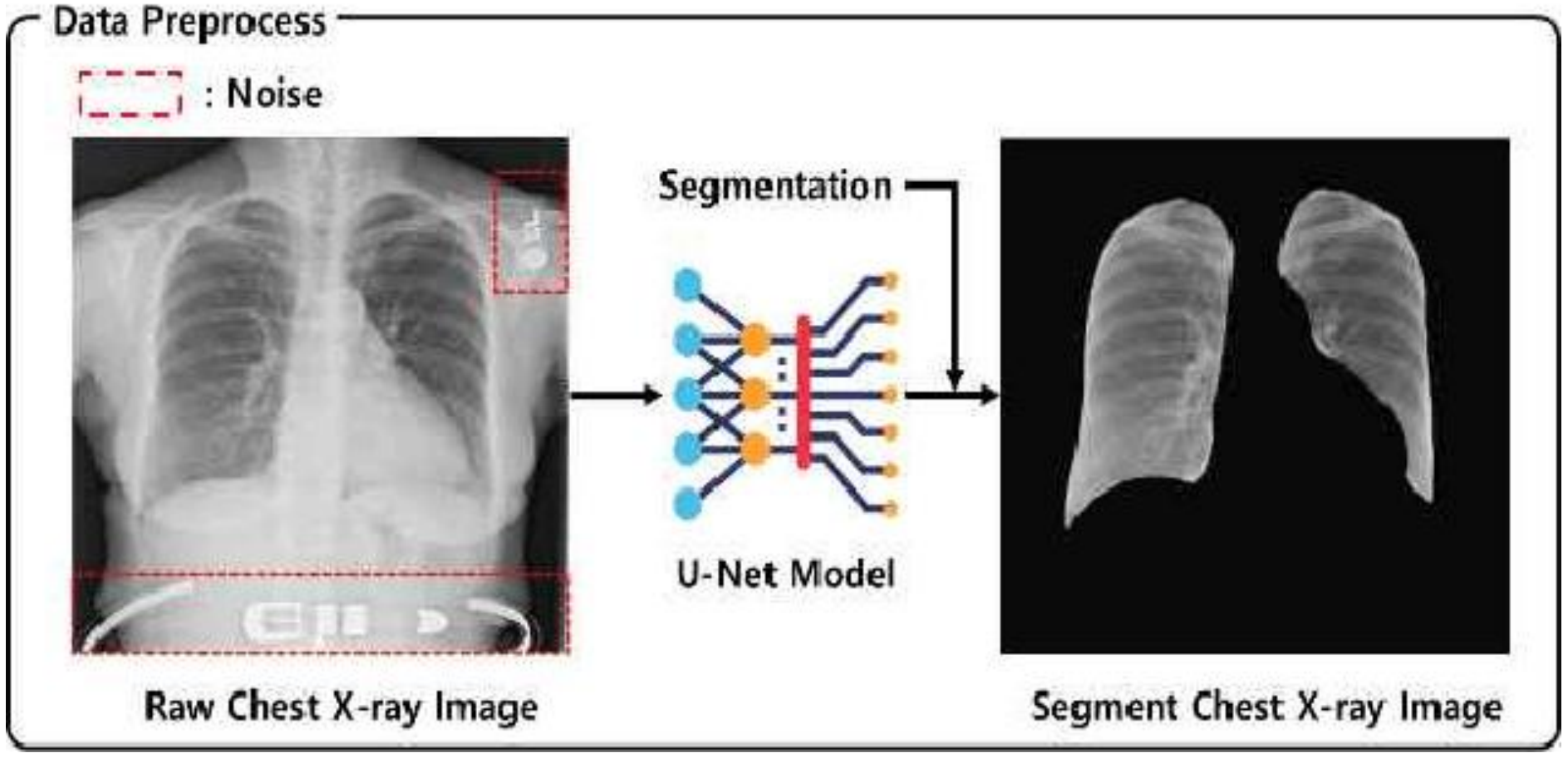
Best Practices for Medical Image Preprocessing

Challenges in Medical Image Preprocessing

Latest Trends in Medical Image Preprocessing

Tools and Software for Medical Image Preprocessing

What is Medical Image Preprocessing?



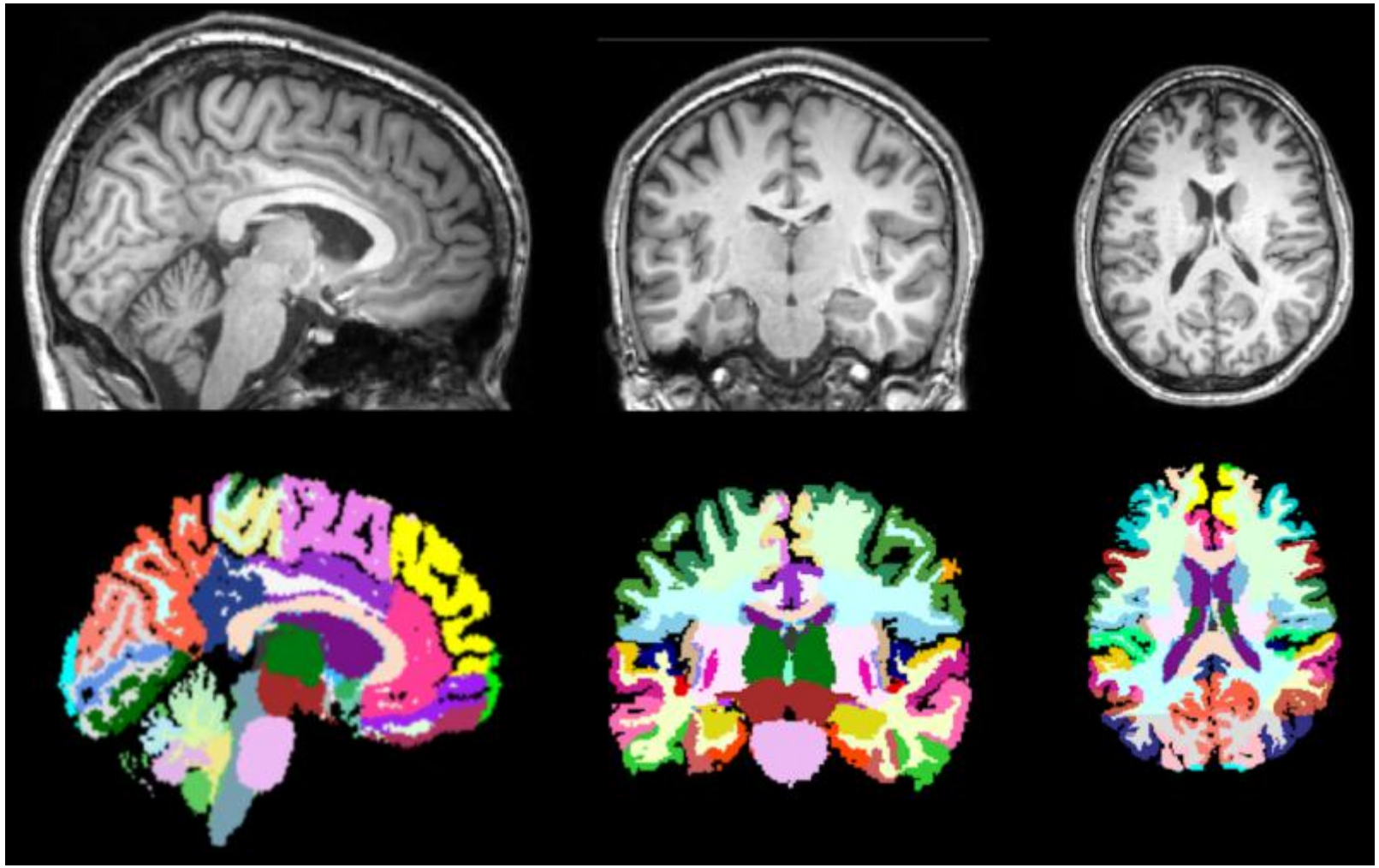
Why is Preprocessing Essential in Medical Imaging?

- Improved Image Quality
- Standardization
- Enhanced Analysis
- Increased Diagnostic Accuracy

Common Preprocessing Techniques

- Background Removal
- Denoising
- Resampling
- Registration
- Intensity Normalization

Advanced preprocessing workflows



Best Practices for Medical Image Preprocessing

- Understand Your Data
- Preserve Original Data
- Document Your Process
- Validate Your Results
- Use Standardized Protocols
- Consider the Downstream Analysis
- Be Consistent
- Handle Missing Data Appropriately
- Optimize for Performance
- Stay Updated

Challenges in Medical Image Preprocessing

- Variability in Image Quality
- Preservation of Subtle Features
- Computational Resources
- Standardization Across Institutions
- Handling of Artifacts

Latest Trends in Medical Image Preprocessing

Deep Learning-Based Preprocessing

Multi-Modal Preprocessing

Automated Preprocessing Pipelines

Edge Computing for Preprocessing

Federated Learning for Preprocessing

Tools and Software for Medical Image Preprocessing

MATLAB

SimpleITK

NiBabel

ANTs (Advanced Normalization Tools)

FSL (FMRIB Software Library)

SPM (Statistical Parametric Mapping)

TorchIO

Future of Medical Image Preprocessing

- Adaptive Preprocessing
- Real-time Preprocessing
- Quantum Computing for Preprocessing
- Integrating Clinical Data
- Explainable AI in Preprocessing

Biomedical signal Pre processing

What is Biomedical signal Pre processing

Medical signal preprocessing refers to the techniques used to enhance, filter, and prepare raw biomedical signals for analysis.

Signals:

ECG

EEG

EMG

Why Biomedical signal Pre processing in important

- Noise and Artifacts
- Inconsistent Data
- Feature Extraction

Why Biomedical signal Pre processing in important

- Heart disease diagnosis.
- Diagnosis of neuromuscular disorders.
- Control of prosthetic limbs using muscle signals.
- Gesture recognition and human-computer interaction.
- Sports science and rehabilitation for muscle monitoring.
- Epilepsy detection.
- Sleep disorder analysis and sleep cycle monitoring.
- Brain-computer interface (BCI) applications.
- Cognitive neuroscience and mental state monitoring.

ECG Pre processing

- Filtering
- Baseline Wander Removal
- Normalization
- Wavelet Transform
- Segmentation

EEG Pre processing

- Filtering
- Artifact Removal
- Epoching & Segmentation
- Baseline Correction
- Wavelet Transform

EMG Pre processing

- Filtering
- Rectification
- Smoothing
- Baseline Correction
- Segmentation

Thank You!