Data and Datasets Collection

The data and the datasets were collected from the sources listed in section 3.2 of the manuscript using the following method:

The data acquisition pipeline utilized Selenium 4.10.0 with ChromeDriver 114.0.5735.90 for automated extraction from health repositories, implementing headless mode (options.add_argument("--headless")) to minimize resource consumption and WebDriverWait with explicit waits for dynamic content loading. Custom web scrapers with XPath selectors were developed to navigate through WHO, CDC, and wastewater surveillance portals, handling pagination via driver.execute_script("window.scrollTo(0, document.body.scrollHeight)") and capturing tabular data through pandas.read_html(driver.page_source). The extracted datasets were encapsulated within a unified Python package named smarthealthtrack with modular architecture comprising dataloaders, preprocessors, and analyzers submodules, exposing a consistent API (load_patient_records(), load_pharmaceutical_sales(), etc.). This package employs SQLAlchemy ORM for dataset persistence with incremental update capabilities and implements automated data validation using pandera.SchemaModel with custom checks for temporal consistency. Leveraging this library, the analytical workflow employed dimensionality reduction via sklearn.decomposition.PCA(n_components=0.95) followed by time-series decomposition through statsmodels.tsa.seasonal_decompose() to distinguish seasonal patterns from outbreak anomalies. The integrated analysis framework executed feature importance ranking via permutation importance (sklearn.inspection.permutation_importance(n_repeats=30)), identified pharmaceutical demand spikes using scipy.signal.find_peaks(distance=7, prominence=2.0), and quantified cross-correlation between wastewater pathogen levels and clinical cases through numpy.correlate(mode='full') with time-lag analysis, ultimately enabling 91.2% precision in early outbreak detection.