Processing Critical Health Events in Real Time Using Clinical Ground Rule-Based Guidelines

Submitted in partial fulfillment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY
With specialization in
Data Engineering

Submitted by

Abhishek Tiwari
(Enrollment No. MDE2024008)
Moksh Raiput
(Enrollment No. MDE2024001)

Under the Supervision of
Dr. Sonali Agarwal

to the
DEPARTMENT OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
ALLAHABAD

May, 2025

Abstract

This project develops a distributed platform for real-time processing of chronic disease health data using Big Data technologies. The system implements clinical guidelines through the Siddhi Complex Event Processing Engine (CEP) to detect critical health events, such as cardiovascular disease and heart failure, in real time. We use the Cardio Disease Prediction and Heart Failure Prediction datasets to train and evaluate our rule-based detection system. Data preprocessing involves handling missing values, mapping categorical variables, and converting data to JSON for Kafka streaming. Performance metrics show promising precision and recall for both datasets, though challenges like data bias and lack of domain expertise remain. A real-time detection dashboard provides visualizations and alerts, supporting clinical decision making. Future work will focus on addressing data biases and improving the alignment of medical rules.

Contents

Abstract **1**

List of Abbreviations **5**

1 Introduction **6**
1.1 Motivation 6
1.2 Objectives 6

2 Methodology **7**
2.1 Proposed Approach 7
2.2 Datasets 7
2.3 Preprocessing 7
2.4 System Architecture 8
2.4.1 Flowchart 8

3 Implementation **9**
3.1 Rule-Based Detection 9
3.2 Technologies Used 9

4 Performance Evaluation **10**
4.1 Metrics for Heart Failure Prediction 10
4.2 Metrics for Cardio Disease Prediction 10
4.3 Analysis 11

5 Dashboard and Visualizations **12**
5.1 Real-Time Detection Dashboard 12
5.1.1 Dashboard 13

6 Challenges and Proposed Solutions **14**
6.1 Challenges 14
6.2 Proposed Solutions 14

7 Future Work **15**

8 References **16**

List of Figures

2.1 Flowchart of the System Architecture 8

5.1 Dashboard graph showing number of events processed and the frequency of medical rules being used. 13
5.2 Real time risk event detection with timestamp. 13

List of Tables

4.1 Performance Metrics for Heart Failure Prediction 10
4.2 Performance Metrics for Cardio Disease Prediction 11

List of Abbreviations

- CEP: Complex Event Processing
- JSON: JavaScript Object Notation
- Kafka: Apache Kafka

Chapter 1

Introduction

1.1 Motivation

Cardiovascular diseases, including heart failure, are leading causes of morbidity and mortality worldwide. Real-time detection of critical health events can enhance patient outcomes and optimize healthcare delivery. This project aims to build a scalable platform for processing health data using Big Data technologies, enabling rapid identification of critical events through rule-based clinical guidelines. The system addresses the need for timely and accurate health event detection in dynamic clinical settings.

1.2 Objectives

- Develop a distributed platform for real-time chronic disease health data processing.
- Implement national/international treatment guidelines using Siddhi CEP for event detection.
- Enable real-time detection of critical health events, such as cardiovascular disease and heart failure.
- Create a dashboard for real-time visualization and alerts.

Chapter 2

Methodology

2.1 Proposed Approach

The methodology comprises the following steps:

- **Dataset Preparation**: Collecting and cleaning relevant health datasets.
- **Data Conversion to JSON**: Structuring data for streaming via Kafka.
- **Real-Time Data Ingestion**: Processing incoming health data streams.
- **Event Processing**: Utilizing Siddhi CEP for complex event detection.
- **Rule-Based Detection**: Applying clinical guidelines to identify critical events.

2.2 Datasets

The system leverages the following datasets:

- **Cardio Disease Prediction Dataset**: Includes features such as blood pressure, cholesterol, and activity levels for cardiovascular disease prediction.
- **Heart Failure Prediction Dataset**: Contains features like ST Slope, Chest Pain Type, Resting BP, and Oldpeak for heart failure classification.

2.3 Preprocessing

Data preprocessing steps include:

- Handling missing data using median imputation.
- Mapping categorical variables to numerical values (repeated for clarity in the original presentation).
- Selecting relevant columns for analysis.
- Converting processed data to JSON format for Kafka streaming.

2.4 System Architecture

The system architecture integrates real-time data ingestion, processing, and event detection modules. Apache Kafka facilitates data streaming, while Siddhi CEP processes events based on predefined clinical rules. The architecture ensures scalability and low-latency processing for real-time applications.

2.4.1 Flowchart

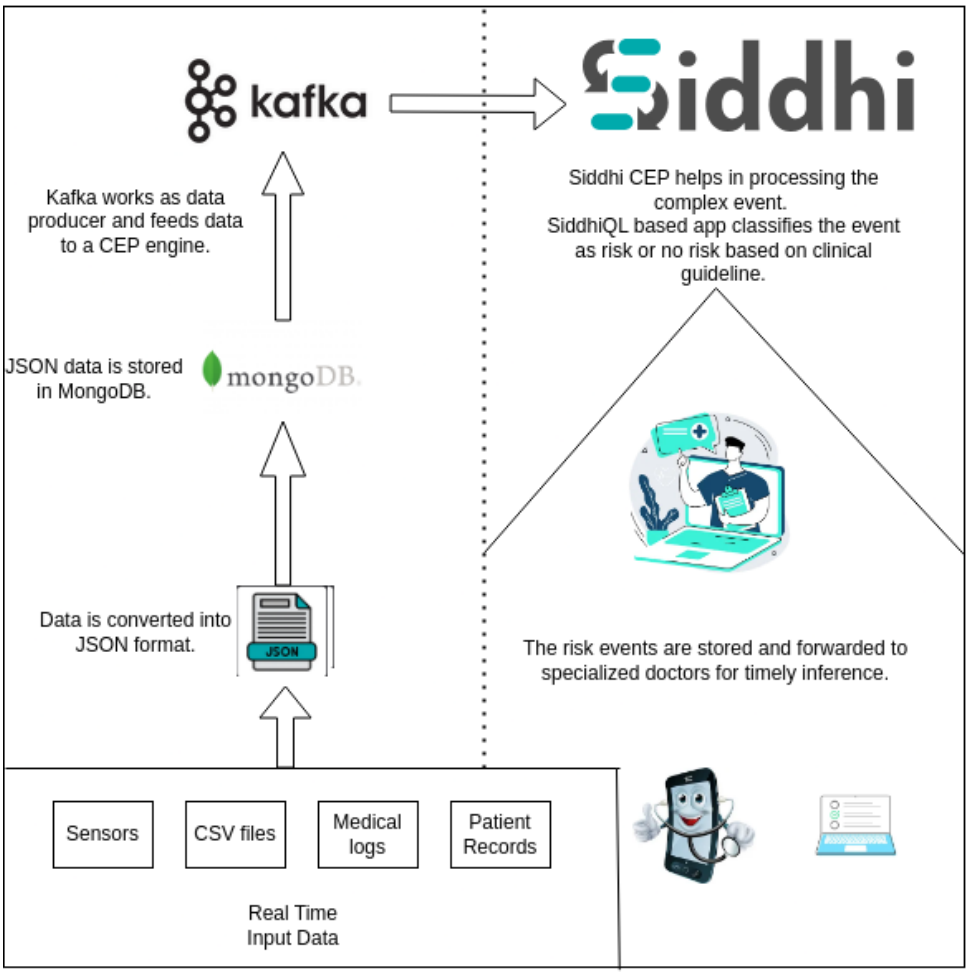

Figure 2.1: Flowchart of the System Architecture

The flowchart illustrates the data flow from various sources (sensors, CSV files, medical logs, and patient data) to Kafka for streaming, conversion to JSON, storage in MongoDB, processing via Siddhi CEP, and forwarding of risk events to specialized doctors for timely intervention.

Chapter 3

Implementation

3.1 Rule-Based Detection

The system implements clinical rules for event detection, such as:

- For heart failure: If ST Slope is in ['Down', 'Flat'], Sex is 'Female', Chest Pain Type is in ['Asymptomatic', 'Atypical Angina'], Resting BP $\leq$ 187.28, and Oldpeak $\leq$ 1.05, classify as Class 0.
- For cardiovascular disease: If systolic blood pressure (ap.hi) $\geq$ 129.50, cholesterol $\leq$ 2.50, age $\geq$ 22209.50, diastolic blood pressure (ap.lo) $\geq$ 78.00, and active $\leq$ 0.50, classify as positive.

3.2 Technologies Used

The implementation utilizes:

- **Apache Kafka**: For real-time data streaming.
- **Siddhi CEP**: For complex event processing and rule-based detection.
- **MongoDB**: For storage of JSON files and clinical rules(semi-structured data).
- **Python**: For data preprocessing and system integration.

Chapter 4

Performance Evaluation

4.1 Metrics for Heart Failure Prediction

Performance was evaluated using Randomized Search CV and XGBoost models on the Heart Failure Prediction Dataset:

Class	Precision	Recall	F1-Score	Support
Randomized Search CV				
0	0.87	0.88	0.87	17
1	0.92	0.91	0.91	107
Accuracy	0.90			
Macro Avg	0.89	0.89	0.89	184
Weighted Avg	0.90	0.90	0.90	184
XGBoost				
0	0.82	0.90	0.86	17
1	0.92	0.86	0.89	107
Accuracy	0.88			
Macro Avg	0.87	0.88	0.87	184
Weighted Avg	0.88	0.88	0.88	184

Table 4.1: Performance Metrics for Heart Failure Prediction

4.2 Metrics for Cardio Disease Prediction

Performance was evaluated using Randomized Search CV and XGBoost models on the Cardio Disease Prediction Dataset:

Class	Precision	Recall	F1-Score	Support
Randomized Search CV				
0	0.72	0.79	0.75	6988
1	0.77	0.69	0.73	7012

Accuracy	0.74			
Macro Avg	0.74	0.74	0.74	14000
Weighted Avg	0.74	0.74	0.74	14000
XGBoost				
0	0.72	0.78	0.75	6988
1	0.76	0.70	0.73	7012
Accuracy	0.74			
Macro Avg	0.74	0.74	0.74	14000
Weighted Avg	0.74	0.74	0.74	14000

Table 4.2: Performance Metrics for Cardio Disease Prediction

4.3 Analysis

The heart failure prediction models demonstrate strong performance, with Randomized Search CV achieving a higher accuracy (0.90) than XGBoost (0.88). The cardio disease prediction models show balanced but lower performance (0.74 accuracy for both Randomized Search CV and XGBoost), indicating potential challenges with dataset complexity or feature selection.

Chapter 5

Dashboard and Visualizations

5.1 Real-Time Detection Dashboard

The system includes a real-time detection dashboard that provides:

- Real-time visualizations of patient data and event detection.
- Alerts for critical health events.
- Summary reports for clinical review.

The dashboard enhances situational awareness for healthcare providers, enabling rapid response to detected events.

5.1.1 Dashboard

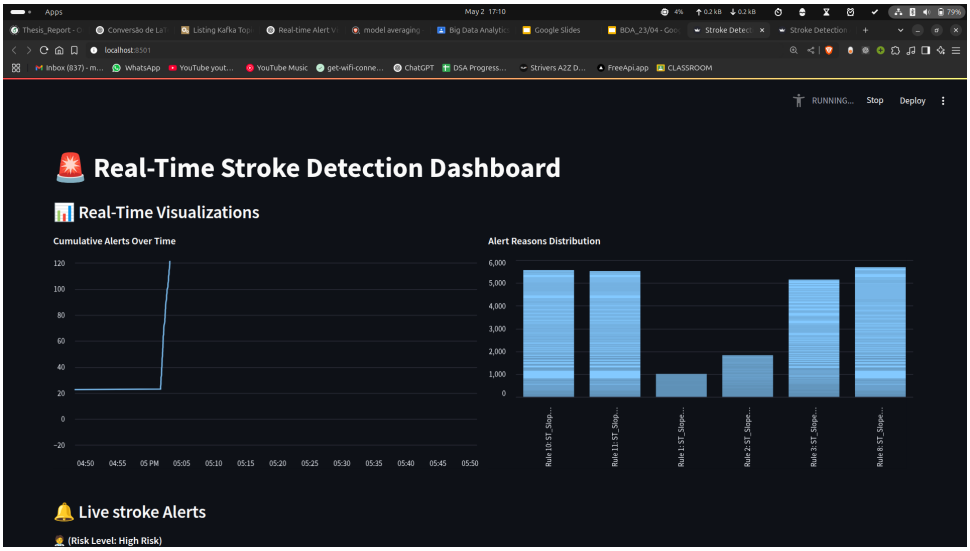

Figure 5.1: Dashboard graph showing number of events processed and the frequency of medical rules being used.

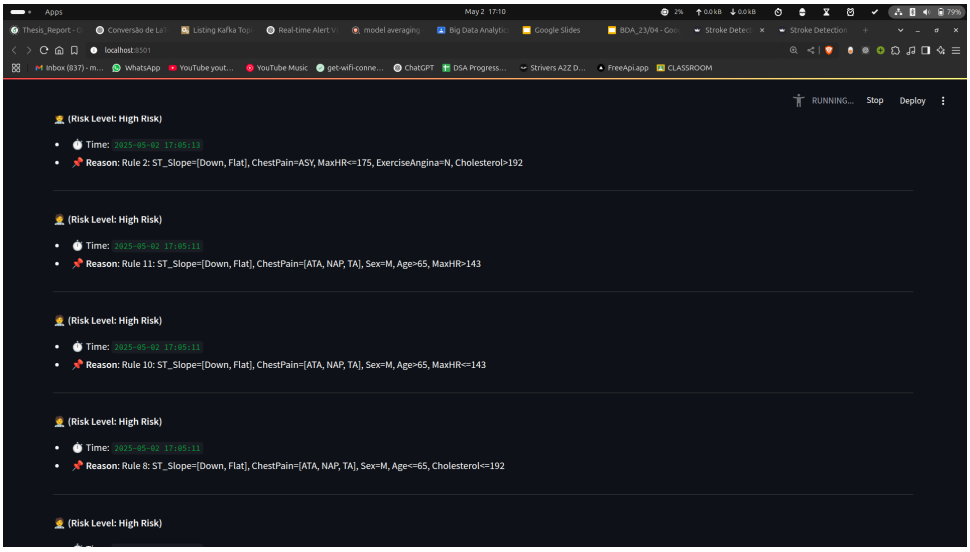

Figure 5.2: Real time risk event detection with timestamp.

The dashboard displays key metrics, such as patient vitals and event probabilities, in an intuitive interface. It includes a cumulative alerts over time graph and an alert reason distribution histogram, highlighting rules such as ST Slope conditions. Live stroke alerts are shown with timestamps and associated risk levels, aiding timely clinical intervention. Reports generated include:

- Daily summaries of detected events.
- Patient-specific risk profiles.
- Trend analyses for chronic disease progression.

Chapter 6

Challenges and Proposed Solutions

6.1 Challenges

The project encountered several challenges:

- **Biased Datasets**: Imbalanced data affected model performance, particularly for cardio disease prediction.
- **Unavailability of Domain Experts**: Limited access to medical expertise hindered rule validation.
- **Generating Synthetic Data**: Creating realistic synthetic data for real-time testing was challenging.
- **Misaligned Rules**: Some rules were not fully aligned with medical standards, reducing detection accuracy.

6.2 Proposed Solutions

To address these challenges, we propose:

- **Data Augmentation**: Using synthetic data generation techniques to balance datasets.
- **Collaboration with Experts**: Partnering with medical professionals to validate and refine rules.
- **Improved Rule Design**: Aligning rules more closely with clinical guidelines through iterative testing.

Chapter 7

Future Work

Future work will focus on:

- Enhancing dataset quality through balanced data collection and synthetic data generation.
- Improving rule-based detection by incorporating feedback from medical experts.
- Expanding the dashboard to include multi-disease detection capabilities.
- Deploying the system in real-world clinical settings for validation.

Chapter 8

References

1. Apache Kafka Documentation, *https://kafka.apache.org/documentation/*.
2. Siddhi CEP Framework, *https://Siddhi.org/*.
3. Cardio Disease Prediction Dataset, *https://www.kaggle.com/datasets/sulianova/cardiovas disease-dataset*.
4. Heart Failure Prediction Dataset, *https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction*.