SYNOPSIS

Submitted to

The IIHMR University, Jaipur

for the partial fulfillment for the award of the degree of

Master of Business Administration (MBA)

in

Healthcare Analytics

by

Sourav

Under the Supervision and Guidance of

Dr. Kumar Parimal Shrestha

2026

NAME- SOURAV KUNDU

COURSE – MBA (HEALTHCARE ANALYTICS)

ROLL NO.- 0026

Dissertation Synopsis

Title: A Comprehensive Analysis of Big Data Analytics Frameworks for Proactive Fraud Detection in the Health Insurance Sector

1. Introduction

The rapid digitalization of the healthcare industry has led to an exponential increase in health insurance claims. While this transition improves accessibility, it also opens avenues for sophisticated fraudulent activities such as "phantom billing," "upcoding," and "identity theft." Traditional manual auditing is insufficient for the massive volume of data generated today. This research focuses on how Big Data Analytics. encompassing Machine Learning (ML) and Artificial Intelligence (AI) can be leveraged to identify fraudulent patterns through secondary data analysis.

2. Problem Statement

Health insurance fraud results in significant financial leakage, leading to higher premiums for honest policyholders and reduced profitability for insurers. Current fraud detection systems are primarily rule-based and reactive, meaning they only catch fraud after the payment is made. There is a critical research gap in developing a proactive, data-driven framework that utilizes existing historical data to predict and prevent fraudulent claims in real-time.

3. Review of Literature (Secondary Approach)

As a secondary research study, this project will synthesize findings from peer-reviewed journals, industry reports (e.g., IRDAI, NHCAA), and previous case studies. The review will focus on:

- The evolution of fraud detection from simple statistical methods to advanced Neural Networks.
- Comparative studies of Supervised vs. Unsupervised learning in healthcare fraud.
- The role of Network Analysis in identifying collusion between providers and patients.

4. Research Objectives

- To categorize the most prevalent types of fraud within the health insurance ecosystem.
- To design a conceptual framework for a predictive fraud detection system based on secondary data findings.
- To identify the technical and regulatory challenges such as data privacy in implementing these systems.

5. Research Questions

- What are the key indicators (features) in secondary healthcare datasets that signal a high probability of fraud?
- Which Machine Learning algorithms demonstrate the highest accuracy and lowest "False Positive" rates in insurance claim analysis?
- How can integrated Big Data architectures improve the speed of fraud detection?

6. Research Methodology

This study will employ a **Qualitative and Analytical Research Design** based entirely on secondary sources:

- **Data Selection:** Analysis of publicly available datasets (e.g., CMS Medicare datasets, Kaggle Insurance Fraud datasets).

- **Systematic Literature Review (SLR):** A rigorous process of filtering and synthesizing academic papers from the last 7–10 years.
- **Comparative Analysis:** Benchmarking different analytical models described in existing literature to determine the most viable solution for the industry.

7. Scholar Search Strategy

- **Databases:** Primary use of Google Scholar, PubMed.
- **Boolean Operators:** Utilizing specific search strings like:
    - "Big Data" AND "Health Insurance" AND "Fraud Detection"
    - "Machine Learning" OR "Deep Learning" AND "Claim Anomaly"
- **Inclusion/Exclusion Criteria:** Focusing on Peer-Reviewed articles published between 2018–2026 to ensure the technology discussed is contemporary.

8. Proposed Conceptual Framework

1. **Data Ingestion:** Aggregation of claims, provider details, and patient history.
2. **Pre-processing:** Cleaning secondary data and handling missing values.
3. **Feature Engineering:** Identifying Risk Flags (e.g., unusually high frequency of specific surgeries).
4. **Modeling:** Applying algorithms like Random Forest or Isolation Forest for anomaly detection.
5. **Risk Scoring:** Assigning a probability score to each claim for further investigation.

9. Significance of the Study

This research is vital for healthcare administrators and policymakers. It provides a roadmap for reducing operational costs without the need for expensive primary surveys. For the academic community, it offers a consolidated view of how disparate Big Data technologies can be harmonized into a single fraud-prevention strategy.

10. Scope and Limitations

- **Scope:** The study covers technical and managerial aspects of fraud detection within the global and Indian health insurance context.
- **Limitations:** Since it is secondary research, the study is dependent on the accuracy of existing datasets. The lack of access to live private insurance data due to confidentiality (HIPAA/DISHA) may limit the real-time testing of the proposed framework.
