International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Survey Report on MediScope - Web-Based Disease Surveillance using NLP and Machine Learning

Shreeshailya Zirpe[1], Prof. A. H. Kagne[2], Aaditya Marathe[3], Prashant Pawar[4], Krishna Kawale[5]

Department of Computer Engineering, Sinhgad Academy of Engineering, Pune& College

---

Abstract –

People often struggle to access reliable and timely health information about diseases and viruses, as relevant data is scattered across multiple sources such as WHO, health portals, and government websites. This leads to delays in awareness, misinformation, and difficulty in understanding disease symptoms, affected regions, and preventive measures. Current health information systems are typically static and do not provide real-time, consolidated insights. To solve these issues, we propose MediScope: Spot Sickness Before It Spreads, an intelligent healthcare web crawling and analytics system. This system automatically extracts health data from trusted sources using web scraping techniques, processes and analyzes information using Natural Language Processing (NLP) for entity recognition and symptom extraction, and presents insights through interactive dashboards. It uses a combination of web crawling frameworks (BeautifulSoup, Scrapy), NLP libraries (spaCy, NLTK), and data visualization tools to collect, standardize, and display disease information including current outbreaks, symptoms, geographic distribution, and prevention strategies. By offering centralized access, real-time updates, and user-friendly visualization, this solution aims to improve public health awareness, enable early disease detection, and support informed decision-making, providing a powerful tool for modern healthcare information management.

Keywords: Web Crawling, Natural Language Processing (NLP), Disease Surveillance, Named Entity Recognition, Healthcare Analytics, Data Visualization.

1.INTRODUCTION

In today's fast-evolving healthcare environment, the timely detection and monitoring of disease outbreaks have become essential for preventing large-scale public health crises. However, vital health information is often scattered across various platforms such as WHO portals, wellness websites, and news articles, making it difficult to access reliable and consolidated data quickly. Traditional disease tracking systems rely heavily on manual data collection and delayed reporting, resulting in information gaps and slower response times. For instance, during recent global health emergencies, the lack of centralized information platforms contributed to delayed public awareness and response.

The rise of Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) offers powerful tools to revolutionize digital healthcare surveillance. Automated systems capable of web crawling using frameworks like BeautifulSoup and Scrapy, combined with semantic analysis through NLP libraries such as spaCy, can continuously gather, analyze, and interpret vast volumes of online data to identify emerging health threats in real time. The proposed system, **MediScope: Spot Sickness Before It Spreads**, leverages these technologies to build an intelligent, AI-powered healthcare information platform that aggregates data from trusted health sources, extracts key information through Named Entity Recognition, and visualizes disease trends geographically.

This survey paper explores how MediScope integrates web crawling, NLP-based entity extraction, and data analytics to improve early disease detection, support informed decision-making, and enhance public awareness. By automating data aggregation and providing real-time insights through interactive dashboards, the system aims to bridge the gap between global health data and actionable intelligence, contributing to a safer and more informed society. The following sections review existing literature on web crawling techniques, NLP in healthcare, and disease surveillance systems, followed by the proposed system architecture and methodology.

2. LITERATURE RIVIEW

2.1 Web Crawling and Data Extraction Techniques

Web crawling frameworks like BeautifulSoup and Scrapy enable automated extraction of structured data from HTML documents. Modern approaches use headless browsers (Selenium, Playwright) to handle dynamic content. In healthcare, intelligent crawlers with adaptive parsing maintain high accuracy despite frequent website structure changes. MediScope integrates these frameworks for continuous health data extraction from WHO, CDC, and wellness portals.

2.2 Natural Language Processing in Healthcare

Named Entity Recognition (NER) systems, particularly BERT-based models, achieve state-of-the-art performance in identifying medical entities such as diseases, symptoms, and medications from unstructured text. Healthcare-specific NLP libraries like spaCy and BioBERT, trained on medical literature, enable accurate extraction with F1 scores above 0.90. MediScope leverages these advances through custom NER pipelines to automatically identify and categorize health information from crawled content.

2.3 Machine Learning for Disease Prediction and Analytics

Machine Learning models enable predictive analytics in healthcare systems. Classification algorithms (Random Forest, SVM, Neural Networks) trained on historical disease data can predict outbreak severity and threat levels. Time series forecasting models (ARIMA, LSTM) analyze disease progression patterns to forecast future trends. Studies show that ensemble methods combining mult
iple ML models improve prediction accuracy by 15-20% compared to single models. MediScope integrates supervised learning for threat classification and unsupervised learning for pattern discovery in disease data.

2.4 Existing Disease Surveillance Systems

Platforms like WHO's GOARN, CDC's systems, and HealthMap pioneered automated disease surveillance. However, existing systems often operate as closed platforms, lack real-time visualization, or require subscriptions. ProMED-mail and BlueDot provide recent approaches but lack public accessibility and comprehensive features. Current systems rarely combine data collection, NLP processing, and visualization in a single accessible platform.

2.5 Research Gap

Although web crawling, NLP, and disease surveillance have been studied separately, few solutions integrate these technologies into a unified, publicly accessible platform. Existing systems lack either real-time data aggregation, advanced NLP processing, or interactive visualization. MediScope bridges this gap by combining automated web crawling, NLP-based entity extraction, and geographic visualization into a single cohesive, accessible platform.

3. PROPOSED SYSTEM ARCHITECTURE

The proposed system is designed as a layered architecture integrating data collection, NLP processing, machine learning, storage, and presentation components.

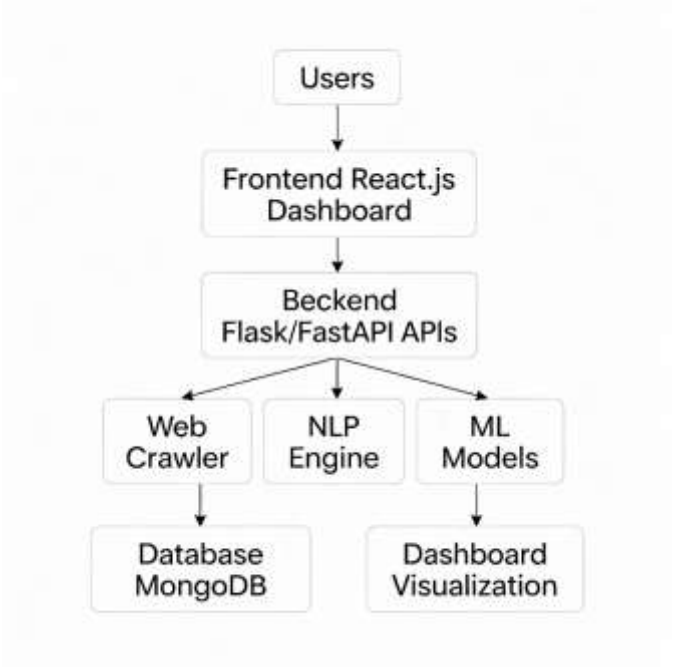

Figure 1 System Architecture Diagram

3.1 System Components

1. **Web Crawling Module:** Automatically extracts health data from trusted sources (WHO, CDC, wellness websites) using BeautifulSoup and Scrapy frameworks, handling dynamic content through headless browsers.
2. **NLP Processing Engine:** Applies Named Entity Recognition (NER) using spaCy and NLTK libraries to identify and classify diseases, symptoms, locations, and prevention measures from extracted text.
3. **Machine Learning Module:**Trains classification models using scikit-learn and TensorFlow to predict disease threat levels, perform sentiment analysis on health discussions, and forecast disease trends based on historical data patterns.
4. **Data Storage Layer:** Stores standardized, cleaned health information in MongoDB/PostgreSQL, enabling efficient retrieval and historical trend analysis.
5. **Analytics Module:** Analyzes disease patterns, tracks geographic distribution, identifies emerging trends, and calculates threat levels using ML-driven insights.
6. **Visualization Dashboard:** React.js-based interactive interface displaying disease information, symptom correlations, affected regions through maps and charts, with real-time ML-powered predictions and alerts.

3.2 System Features

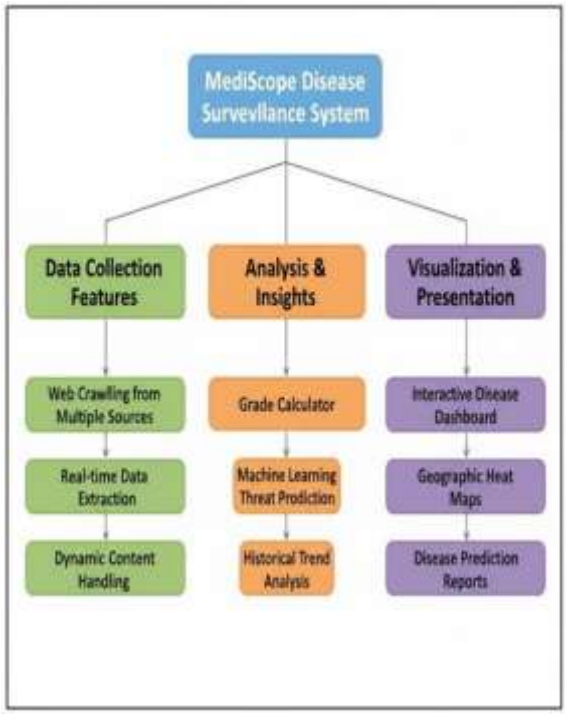

Figure 2 System Feature Diagram

- Automated data collection from multiple trusted health sources without manual intervention.
- Real-time NLP-based entity extraction and health information processing.
- Machine Learning models for threat level prediction and disease forecasting.
- Interactive geographic visualization of disease outbreaks and affected regions.
- Historical trend analysis and comparative disease tracking with ML insights.
- User-friendly web interface accessible to non-technical users.
- Scalable architecture supporting increased data volume and concurrent users.

4. METHODOLOGY

The workflow of MediScope follows a structured pipeline that ensures reliability, efficiency, and reproducibility in health information aggregation and analysis.

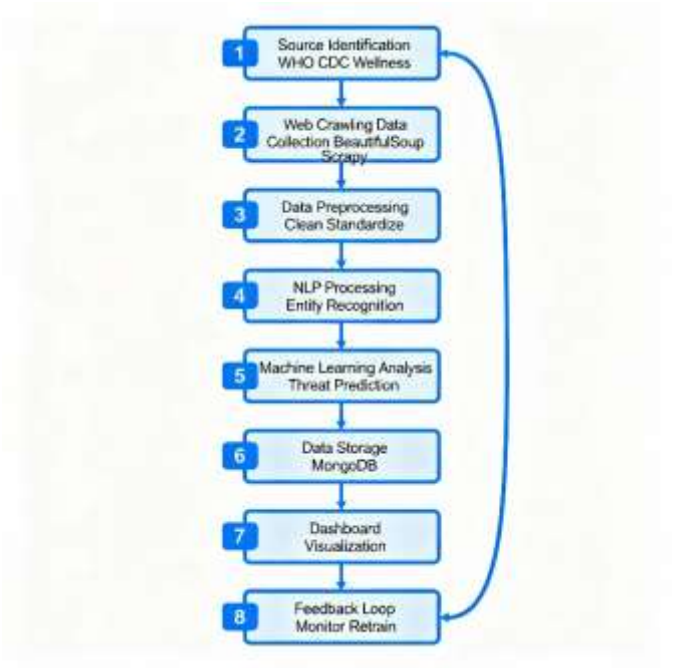

Figure 3 Data Flow Diagram

1. **Data Collection:**
   a. Source Identification: The system automatically resizes, normalizes, and augments images to enhance model generalization.
   b. Web Crawling: Automated extraction of health content from identified sources using BeautifulSoup and Scrapy frameworks. Headless browsers (Selenium) handle dynamic content. Data is stored in raw format for preprocessing.
2. **Data Processing:**
   a. Data Preprocessing: Clean extracted data by removing duplicates, standardizing formats, handling missing values, and filtering irrelevant content to ensure data quality.
   b. NLP Processing: Apply spaCy-based Named Entity Recognition to identify diseases, symptoms, locations, and prevention measures from extracted text. Classify extracted entities into structured categories.
3. **Analysis and Storage:**
   a. Machine Learning Analysis: Train classification models using scikit-learn to predict disease threat levels based on symptom frequency, geographic spread, and historical patterns. Apply time series forecasting (LSTM/ARIMA) to predict disease trends.
   b. Data Storage: Store structured, processed health information in MongoDB/PostgreSQL with metadata (source, timestamp, confidence scores) for efficient retrieval and historical tracking.
4. **Visualization and OutputPhase:**
   a. Dashboard Visualization: Present aggregated insights through React.js interactive dashboards displaying disease information, symptom correlations, geographic heatmaps, and ML-based threat predictions with real-time updates.
   b. Feedback Loop: Monitor system performance, user feedback, and emerging health trends. Retrain ML models periodically with new data to enhance prediction accuracy.

5. Technology Stack

Layer	Technology / Tool
Frontend	React.js, Chart.js, Tailwind CSS
Backend	Flask / FastAPI
Web Crawling	BeautifulSoup, Scrapy, Selenium
NLP Processing	spaCy, NLTK, BERT
Machine Learning	TensorFlow, scikit-learn, LSTM
Database	MongoDB, PostgreSQL
APIs & Data Processing	Python, Pandas, NumPy
Deployment	Docker, RESTful APIs, Gunicorn
Hardware Requirements	8 GB+ RAM, Intel i5 or higher, SSD

6. CONCLUSION

MediScope integrates web crawling, NLP processing, and machine learning analytics into a unified platform for disease surveillance. By automating data aggregation from trusted sources and providing real-time insights, it addresses gaps in existing systems. The user-friendly interface ensures accessibility for non-technical users, enabling early detection of health threats and informed decision-making for public health awareness.

Future Enhancements

1. **Predictive Analytics:** Advanced ML models (ARIMA, Prophet) to forecast disease outbreaks weeks in advance.
2. **Multilingual Support:** Process health information from non-English sources for global disease tracking.
3. **Mobile Application:** Native iOS/Android apps with push notifications for real-time disease alerts.
4. **Government API Integration:** Direct integration with WHO, CDC, and health ministry databases for data verification.
5. **Sentiment Analysis:** Monitor social media and public discussions for disease-related misinformation and public perception.
6. **Cloud Deployment:** Expand using AWS, Azure, and Google Cloud for enhanced scalability and global accessibility.
7. **Explainable AI (XAI):** Implement interpretable ML models to explain predictions and build user trust.

7. ACKNOWLEDGMENT

The authors would like to express their gratitude to **Prof. Ms.A.H.Kagane** for her continuous guidance, insightful feedback, and encouragement throughout the development of this project. The team also extends sincere thanks to the **Department of Computer Engineering, Sinhgad Academy of Engineering, Pune**, for providing the facilities and technical

support required to carry out this research successfully. Finally, the authors acknowledge their peers and families for their motivation and support during the completion of this work.

8. REFERNCES

1. N. Yanes, L. Jamel, B. Alabdullah, M. Ezz, A. M. Mostafa, and H. Shabana, "Using Machine Learning for Detection and Prediction of Chronic Diseases. IEEE 2024.

2. S. Liu et al., "Evaluating Medical Entity Recognition in Health Care," JMIR Medical Informatics, vol. 12, no. 1, art. e59782, October 2024.

3. J. Bergman and O. B. Popov, "Exploring dark web crawlers: A systematic literature review of dark web crawlers and their implementation," IEEE Access, vol. 11, pp. 35914–35941, Mar. 2023.

4. S. Raza, A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya, "Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review," IEEE Access, vol. 12, pp. 180815-180834, Nov. 2024.

5. S. Ganesh, "Web Automation in Health Care," in Proc. IEEE Int. Conf. Recent Trends Electronics Inform. Technol. (RAITEC*)*, 2019, pp. 1-5.

6. E. Khan and S. Bhide, "Web Scrapping Tools Used in Healthcare Sector," in Proc. IEEE Int. Workshop Healthcare Data Analytics, 2024.M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Proceedings of the International Conference on Machine Learning (ICML*)*, 2019.