BCSE498J Project-II

Early Warning System for Lifestyle-Related Health Risks Using Text Analytics

22BCE0490 ABHAY SINGH

22BCE2573 MANAS JOSHI

22BCE2588 YAGNIT MAHAJAN

Under the Supervision of

Dr. Yoganand S

Assistant Professor Sr. Grade 1

School of Computer Science and Engineering (SCOPE)

B.Tech.

in

Computer Science and Engineering

School of Computer Science and Engineering (SCOPE)

February 2, 2026

BCSE498J Project-II

ABSTRACT

Lifestyle related health issues usually develop gradually and are often shaped by everyday habits. People tend to reveal subtle clues about their routines through the way they write or talk about their daily lives, whether it is feeling constantly tired, spending too much time on screens, or making poor food choices. These small behaviors may not seem serious at first, but over time they can accumulate and lead to larger health problems. Issues related to sleep, weight, drinking, smoking, and excessive screen time are common, yet many people fail to notice them early because the effects are slow and not immediately visible.

This project aims to address that gap by building an early warning system that analyzes text data to identify patterns linked to unhealthy lifestyle behaviors. By using natural language processing and machine learning techniques, the system looks for repeated mentions, emotional cues, and language patterns that may suggest emerging risks. For instance, frequent references to exhaustion or staying up late can indicate poor sleep habits, while mentions of skipping meals, junk food, or lack of appetite may point to weight related concerns. The goal is not to diagnose conditions, but to highlight early signals that deserve attention.

What makes the system especially useful is its focus on clarity and transparency. Instead of producing unclear predictions, the system explains its results by showing which words or phrases influenced the outcome. This helps users understand why a certain lifestyle risk was flagged and builds trust in the system. By presenting insights in an understandable way, the system encourages users to reflect on their habits and take small, practical steps toward healthier choices. Ultimately, the system supports early awareness and gradual improvement, helping people protect their long term wellbeing before problems become serious.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

Lifestyle-related health problems such as diabetes, obesity, heart disease, and high blood pressure are increasing day by day. These problems are mainly caused by poor food habits, lack of exercise, improper sleep, and high stress. Most of these diseases do not appear suddenly. They develop slowly over a period of time and often show no clear symptoms in the early stages.

In many cases, healthcare systems focus on treating diseases after they occur. Regular health checkups are usually done only when a person feels unwell. Because of this, early signs related to unhealthy lifestyle habits are often ignored. A large amount of useful health information is available in the form of text, such as lifestyle surveys, daily health notes, and self-reported habits.

With the help of machine learning and text analysis techniques, it is possible to study this type of data and find patterns related to health risks. Early Warning Systems are used to identify possible health problems at an early stage and alert users in advance. Such systems can support better health awareness and help people take timely action to improve their lifestyle.

1.2 MOTIVATION

The primary motivation behind this project is the growing need for preventive healthcare solutions that focus on early detection rather than delayed treatment. Lifestyle-related diseases place a significant burden on individuals as well as healthcare systems, both in terms of cost and long-term health complications. Early identification of risk factors can help reduce this burden by enabling timely lifestyle modifications.

From an academic perspective, this project provides an opportunity to apply concepts of machine learning, text analytics, and data preprocessing to a real-world healthcare problem. The increasing availability of lifestyle-related data and the rapid growth of artificial intelligence in healthcare make this domain both relevant and challenging. Working on this project helps bridge the gap between theoretical knowledge and practical implementation.

Additionally, current trends in healthcare emphasize personalized and data-driven approaches. An intelligent system capable of analyzing lifestyle patterns and generating early warnings aligns well with these trends. This project is motivated by the goal of contributing to proactive health management and exploring how technology can support individuals in maintaining healthier lifestyles.

1.3 SCOPE OF THE PROJECT

The scope of this project is limited to developing an early warning system that analyzes lifestyle-related text data to predict possible health risks. The system focuses on factors such as food habits, physical activity, sleep patterns, and stress-related information provided by users in text form. Based on this information, machine learning models are used to classify users into different risk levels.

The project includes steps such as data collection, text cleaning, feature extraction, model training, and basic performance evaluation. The goal is to show how text-based lifestyle data can be used for early risk identification.

This project does not provide medical diagnosis or treatment. It does not include real-time data from sensors or medical devices. The system is meant only as a support tool for health awareness. The results may vary depending on the quality and size of the data used, and these limitations are considered within the scope of the project.

CHAPTER 2

PROJECT DESCRIPTION AND GOALS

2.1 LITERATURE REVIEW

The literature review discusses existing studies related to lifestyle-based health risk prediction, text analytics in healthcare, and early warning systems. Previous research shows that lifestyle factors such as diet, physical activity, sleep, and stress play an important role in long-term health conditions. Machine learning and text analysis techniques have been widely used to analyze such data and identify early health risks.

This review helps in understanding the methods used in earlier work and highlights their limitations, which supports the identification of research gaps for the proposed system.

2.1.1 Traditional ML Approaches

Traditional machine learning (ML) techniques have been widely used for predicting lifestyle-related diseases due to their effectiveness with structured health data. Models such as Logistic Regression, Random Forests, Support Vector Machines (SVM), and ensemble methods are commonly applied to datasets derived from health surveys and biobanks, including NHANES and the UK Biobank [1], [4], [10], [38]. These approaches are particularly suitable for tabular lifestyle features such as dietary habits, physical activity levels, sleep duration, and body mass index.

Several studies report that ensemble models consistently outperform individual classifiers by capturing complex, non-linear relationships among lifestyle variables [6], [10]. For instance, Random Forest and Gradient Boosting models have achieved AUC values ranging from 0.82 to 0.88 for diabetes and cardiovascular disease (CVD) prediction [1], [8], [11]. Additionally, traditional ML models provide interpretable feature importance measures, which support clinical validation and decision-making [4], [27].

Despite their advantages, these methods rely heavily on manual feature engineering and domain expertise. Their performance may degrade when dealing with high-dimensional or unstructured data such as clinical notes or patient-generated text, limiting their applicability in multimodal early warning systems [16], [37].

2.1.2 DL and NLP Based Approaches

Deep learning (DL) and natural language processing (NLP) techniques have enabled more advanced risk prediction by extracting latent health patterns from unstructured and high-volume data sources. These include clinical notes, social media posts, patient diaries, and wearable sensor streams [7], [20], [21]. Unlike traditional ML models, DL architectures can automatically learn hierarchical representations without extensive manual feature engineering.

Hybrid models combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have been successfully applied for mental health, stress, and depression screening using short-text and time-series inputs [13], [31], [33]. Transformer-based models, such as BERT, have further improved disease risk prediction by capturing contextual information from free-text electronic health records (EHRs), leading to AUROC improvements of approximately 5–10% compared to baseline models [21], [36].

Recent studies have also explored multimodal fusion techniques that integrate lifestyle time-series data(e.g., accelerometer signals) with textual embeddings. Such approaches have demonstrated promising results in real-time sepsis and cardiac arrest prediction systems [18], [19], [28]. Moreover, large language models (LLMs) have shown potential for zero-shot and few-shot classification of lifestyle risks from self-reported data, although their deployment requires careful fine-tuning and computational resources [22], [42], [49].

2.1.3 Explainable AI in Healthcare

Explainable Artificial Intelligence (XAI) plays a critical role in addressing the transparency and trust issues associated with black-box predictive models in healthcare. For preventive and early warning systems, clinicians require clear explanations of how lifestyle factors influence predicted risk scores [17], [25]. Techniques such as SHAP and LIME provide feature-level attributions that highlight the contribution of variables like diet quality, physical activity, and stress levels to disease outcomes [9], [24].

XAI has been effectively integrated into lifestyle disease prediction frameworks to enhance clinical interpretability and adoption. For example, explainable early warning systems (xAI-EWS) use trend-based explanations to support timely interventions for obesity, hypertension, and CVD [5], [37]. Multi-stage frameworks such as SpinachXAI-Rec further improve transparency by tracing personalized dietary recommendations back to raw lifestyle inputs [25].

By balancing predictive accuracy with interpretability, XAI bridges the gap between advanced deep learning models and regulatory or ethical requirements in healthcare. Clinical validation studies demonstrate that explainable models significantly improve clinician confidence and usability in real-world decision support systems [17], [39].

2.2 GAPS IDENTIFIED

Paper	Focus Area	Key Limitation	Proposed Solution
Adlakha et al. (2025) [1]	The study focuses on predicting diabetes risk using structured lifestyle and health factors such as age, body mass index, and dietary type through traditional machine learning models.	The approach relies entirely on structured numerical data and does not consider unstructured or user-written lifestyle information, such as stress experiences, sleep quality, or daily habits, which limits its ability to capture real-world behavioral signals.	This work incorporates unstructured text data collected from Reddit posts to capture behavioral and lifestyle-related information. Text-based indicators related to stress, diet, and sleep are extracted and used to generate an interpretable lifestyle risk score.
Cardamone et al. (2025) [7]	The paper applies natural language processing techniques to electronic health record text for predicting mental health conditions.	The study is limited to clinical EHR data and mental health outcomes, making it unsuitable for analyzing public, user-generated text or for assessing lifestyle-related disease risks.	This work extends text-based health analysis to publicly available Reddit data and applies NLP techniques to identify lifestyle-related risk signals, demonstrating a low-cost and scalable approach for lifestyle risk indication outside clinical environments.
Research Team (2022) [31]	The study focuses on detecting depression using social media text through NLP and hybrid deep learning models.	Although effective for mental health analysis, the approach does not link social media text to lifestyle-related physical health risks and does not provide a general lifestyle risk interpretation.	This work adapts social media text analysis to estimate lifestyle-related health risk by identifying language patterns associated with stress, diet, and sleep, and mapping these signals to a composite lifestyle risk score.
Nandwani and Verma (2025) [23]	The research applies sentiment analysis techniques to assess mental health conditions using social media content.	The analysis is limited to emotional or psychological interpretation and does not associate sentiment outcomes with lifestyle-related health risk or preventive health assessment.	This work utilizes sentiment scores as one component of lifestyle risk assessment, combining emotional tone with lifestyle-related textual indicators to provide a broader interpretation of behavioral health risk.

Liu et al. (2025) [21]	The study focuses on extracting structured information from free-text clinical notes using advanced NLP techniques.	The approach depends on structured clinical documentation and annotated medical records, making it less effective for handling informal, noisy, and unstructured social media text.	This work applies lightweight NLP preprocessing and keyword-based feature extraction techniques to Reddit text, demonstrating the feasibility of extracting lifestyle-related health signals from noisy, real-world user-generated content.

2.3 OBJECTIVES

● To collect and preprocess user-generated text data from Reddit related to lifestyle factors such as stress, diet, and sleep habits.

● To extract meaningful textual features, including sentiment and lifestyle-related keywords, from the collected Reddit posts.

● To compute a simple and interpretable lifestyle risk score based on the extracted text features.

● To design and implement a basic frontend user interface that displays the risk score and key contributing factors, allowing users to evaluate lifestyle-related risk insights.

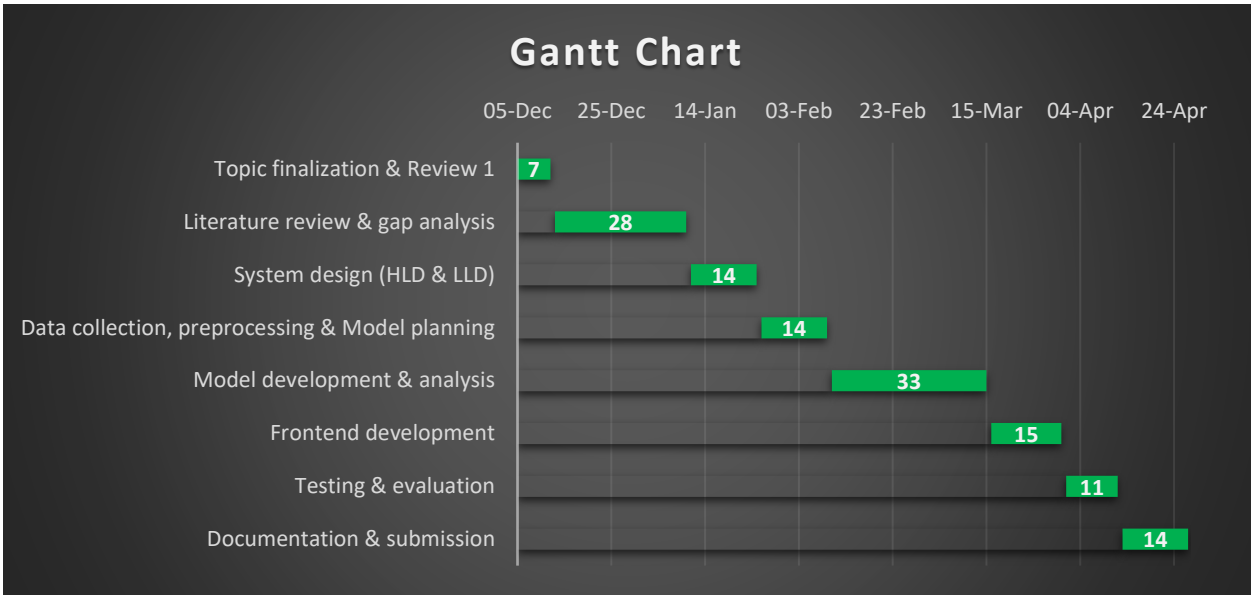
2.4 PROBLEM STATEMENT

Lifestyle-related diseases are often influenced by daily habits, stress levels, and behavioral patterns that are not always captured through structured health records. While many existing approaches rely on clinical or numerical data, user-generated text from online platforms remains underutilized for lifestyle risk assessment. Therefore, there is a need for a simple and interpretable approach that can analyze social media text to identify lifestyle-related risk indicators and provide an understandable risk score that supports early awareness rather than clinical diagnosis.

2.5 PROJECT PLAN

The project is carried out in clear and manageable phases to keep the development process organized and easy to evaluate. The work begins with collecting Reddit text related to lifestyle aspects such as stress, diet, and sleep, followed by basic preprocessing to remove noise and irrelevant content. In the next phase, simple text analysis techniques, including sentiment analysis and keyword extraction, are used to identify lifestyle-related signals from the data. These signals are then used to design a straightforward lifestyle risk scoring approach. In the final phase, a frontend user interface is developed to present the risk score along with the key factors that influence

it, making the results easy for users to understand and interpret. Each phase builds on the results of the previous one, helping ensure smooth integration and steady progress throughout the project.



CHAPTER 3
REQUIREMENTS AND SPECIFICATIONS

3.1 REQUIREMENTS

3.1.1 Functional Requirements

The main functional requirements of the Early Warning System for Lifestyle-Related Health Risks are listed below:

1. The system shall collect publicly available text data from Reddit related to lifestyle and health discussions.
2. The system shall preprocess the collected text by performing cleaning, tokenization, stopword removal, and lemmatization.
3. The system shall analyze the sentiment of the input text to determine the emotional tone (positive, neutral, or negative).
4. The system shall identify lifestyle-related keywords such as sleep issues, stress, diet habits, smoking, and alcohol consumption.
5. The system shall compute a lifestyle risk score using a rule-based scoring mechanism.
6. The system shall classify the risk level as Low, Moderate, or High based on the computed score.
7. The system shall generate an explainable output indicating the factors contributing to the risk score.
8. The system shall display the analysis results through a simple frontend user interface.

3.1.2 Non-Functional Requirements

The following non-functional requirements are considered:

1. **Usability:**
   The system should provide a simple and easy-to-understand user interface suitable for non-technical users.
2. **Performance:**
   The system should analyze and generate results within an acceptable time for small to medium-sized text inputs.

3. **Scalability:**
   The system should be capable of handling an increased number of text inputs with minimal modification.
4. **Reliability:**
   The system should produce consistent and repeatable results for the same input data.
5. **Explainability:**
   The system should clearly explain how the lifestyle risk score is calculated to ensure transparency.
6. **Security and Ethics:**
   The system should use only publicly available data and ensure ethical handling of user-generated content.

3.2 FEASIBILITY STUDY

3.2.1 Technical Feasibility

The proposed system is technically feasible as it uses well-established and widely available technologies. The implementation relies on Python-based NLP libraries such as NLTK for text preprocessing and sentiment analysis. Reddit data is collected using the PRAW (Python Reddit API Wrapper) library, which provides a reliable and documented way to access publicly available Reddit posts. The system architecture is modular, allowing each component to be developed and tested independently. Since the project does not require specialized hardware or real-time sensor data, it can be implemented using standard computing resources available to students.

3.2.2 Economic Feasibility

The project is economically feasible as it involves minimal cost. All tools and technologies used in the system, including Python, NLP libraries, and frontend frameworks, are open-source and freely available. Data is collected from publicly available sources, eliminating any data acquisition costs. The project can be developed using personal computers without the need for additional hardware or paid software licenses. Therefore, the overall development cost of the system is low and suitable for an academic project.

3.2.3 Social Feasibility

The system is socially feasible as it aims to promote awareness of lifestyle-related health risks and encourages early preventive action. By analyzing publicly shared text data, the system does not intrude on personal privacy or require direct user involvement. The explainable nature of the output helps users understand potential risk factors rather than providing medical diagnoses. As a result, the system supports positive health awareness while maintaining ethical and social responsibility.

3.3 SPECIFICATION

3.3.1 Hardware Specification

The proposed system can be developed and executed on a standard personal computer. A minimum of 8 GB RAM, an Intel i5 or equivalent processor, and sufficient storage space are adequate for handling text data and running NLP-based analysis. No specialized hardware such as GPUs or external devices is required, as the system processes only textual data.

3.3.2 Software Specification

The system is implemented using Python as the primary programming language. Text preprocessing and sentiment analysis are carried out using NLTK, while Reddit data is collected using the PRAW library. The backend is developed using Flask, and the frontend interface is built using React. The development environment includes VS Code, and the system is executed on a Windows/Linux operating system.

CHAPTER 4

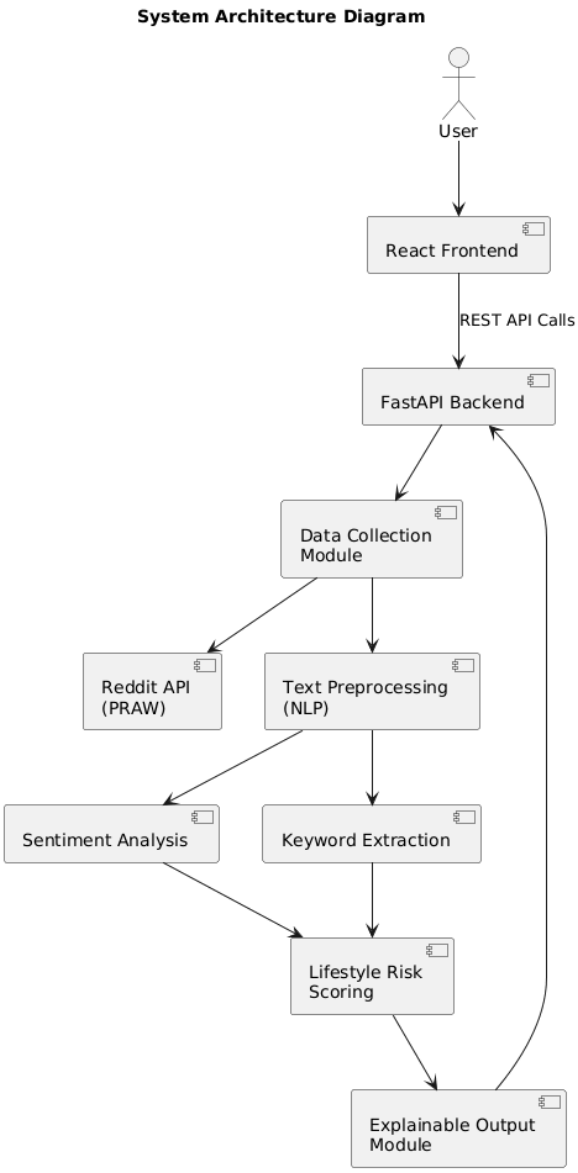
DESIGN APPROACHES AND DETAILS

4.1 SYSTEM ARCHITECTURE

The proposed system follows a modular and layered architecture that supports efficient data processing, machine learning–based analysis, and user interaction. The system is divided into well-defined components, including data acquisition, text preprocessing, analytical processing, backend services, and frontend visualization, allowing each module to function independently.

Lifestyle-related textual data is collected from Reddit using the PRAW API. The collected data is then passed to the preprocessing module, where Natural Language Processing techniques such as text cleaning, normalization, tokenization, and lemmatization are applied to remove noise and prepare the data for analysis. The processed text is subsequently analyzed to determine sentiment polarity and to identify keywords associated with lifestyle risk factors.

Based on the extracted features, a lifestyle risk scoring module computes a relative risk score that reflects potential lifestyle-related health concerns. To promote transparency and user understanding, the system also generates explainable outputs that highlight the key textual factors influencing the assigned risk score. The backend of the system is implemented using FastAPI, which provides RESTful APIs for efficient communication with a React-based frontend, enabling users to view insights and risk indicators through an interactive interface.



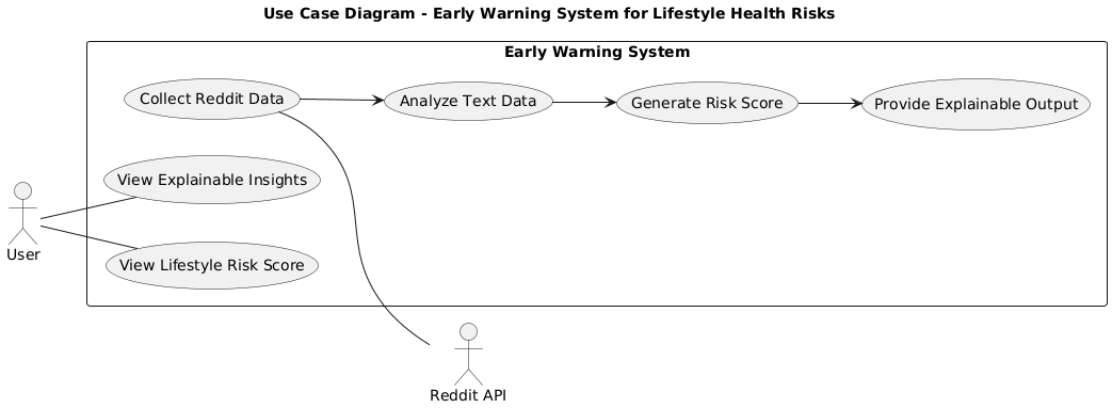
4.2 DESIGN

The system design focuses on a data-driven and explainable workflow, ensuring that each stage of processing contributes clearly to the final output. The design emphasizes simplicity and interpretability rather than complex black-box models, aligning with the objective of providing early warnings instead of medical diagnosis.

The design is divided into functional modules that interact through well-defined interfaces. This modular approach allows individual components to be modified or extended without affecting the overall system functionality.

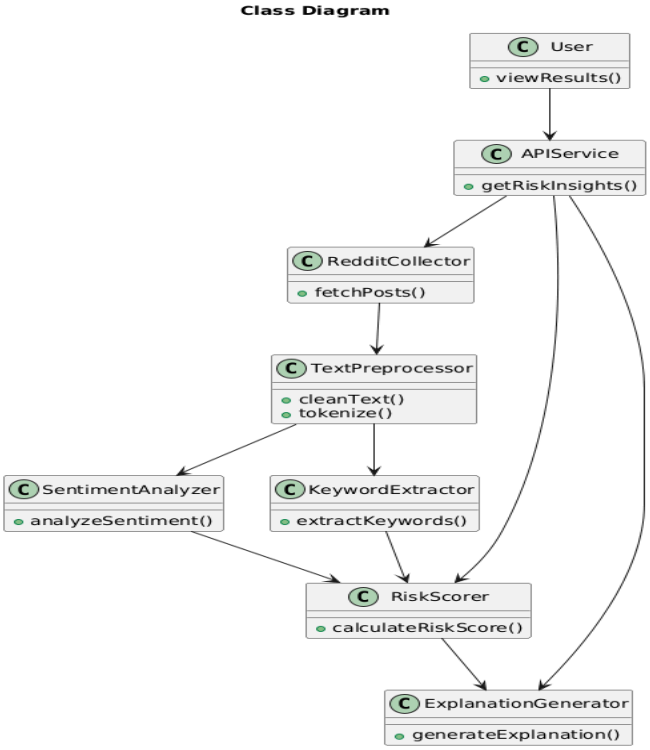
4.2.1 Use Case Diagram

The Use Case Diagram illustrates the functional requirements of the system by identifying the primary actors and the interactions they have with the system. It provides a high-level view of the system's functionality from the user's perspective, without exposing internal processing details.



4.2.2 Class Diagram

The class diagram represents the static structure of the system by depicting the main classes, their attributes, and relationships. It provides an object-oriented view of the system design and helps in understanding how different components interact at the code level.



CHAPTER 5

METHODOLOGY AND TESTING

5.1 PROPOSED METHODOLOGY

The proposed methodology adopts a systematic, data-driven approach to analyze lifestyle-related health discussions from Reddit. It focuses on understanding the data, normalizing textual content, and extracting meaningful features to support reliable and interpretable analysis.

5.1.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted on 19,080 Reddit posts collected across seven lifestyle-related health categories, including fitness, diet, substance use, sleep, mental health, diabetes, and weight management. The objective was to examine data distribution, engagement patterns, and content characteristics prior to further processing.

The dataset spans nearly three years (April 2023 to January 2026). On average, each post contains 812 characters and 145 words, indicating detailed user discussions. Engagement analysis revealed a skewed distribution, with a majority of posts receiving low engagement, while a smaller fraction achieved high interaction levels. Category-wise analysis showed that weight management posts had the highest engagement, whereas sleep-related posts showed lower engagement. Data was sourced from 21 subreddits, with the *running* subreddit being the most active.

Various visualizations such as category distribution plots, engagement metrics, temporal trends, word clouds, and word frequency charts were used to gain insights into lifestyle-related discussion patterns.

5.1.2 Text Preprocessing

Following EDA, the textual data was prepared using a structured NLP preprocessing pipeline. This included text cleaning, tokenization, stopword removal using a customized list of 244 stopwords, and lemmatization to normalize the text.

Preprocessing resulted in a reduction of 41.3% in characters and 51.2% in words, indicating effective noise removal. After discarding insufficient or empty posts, the final dataset consisted of 19,060 posts. The processed corpus contains approximately 1.35 million words with 40,800 unique terms, and an average lexical diversity score of 0.855, reflecting rich vocabulary usage.

Additional engineered features such as text length statistics, sentence count, punctuation indicators, uppercase ratio, and lexical diversity were extracted to enhance subsequent analysis.

5.1.3 Planned Model Development

After data preprocessing, the next step will focus on developing a lifestyle risk assessment model using the processed Reddit text. The goal is to identify patterns in user discussions that may indicate potential lifestyle-related health risks.

Text features will be extracted using TF-IDF for content representation and sentiment scores to capture emotional tone. Additional features such as lifestyle-related keyword frequency and basic text statistics (for example, post length and lexical diversity) will also be included.

A risk score will be computed by combining these features and used to classify posts into Low, Moderate, or High risk categories. Different machine learning models will be evaluated to balance performance and interpretability, and simple explanations will be provided to show which features influenced the risk predictions.

5.2 PERFORMANCE EVALUATION

The proposed system will be evaluated to understand how effectively text analytics and machine learning techniques can identify lifestyle-related health risk indicators from Reddit discussions. The evaluation approach focuses on comparing baseline machine learning models with more advanced models in order to observe improvements in performance and reliability.

Model performance will be assessed using commonly accepted evaluation metrics such as accuracy, precision, recall, and F1-score. Confusion matrices will also be used to analyze class-wise predictions and identify areas where the system performs well or requires improvement. This evaluation process helps ensure that the system produces consistent and meaningful results while supporting reliable early warning generation.

REFERENCES

JOURNALS

1. Adlakha, M., Singh, S., & Sharma, M. (2025). A comparative study of diabetes prediction based on lifestyle and health factors. *IEEE Access*. Afreen, F., Shaikh, A., & Patel, R. (2025). Calorie intake prediction using machine learning for smart diet planning. *International Journal of Creative Research Thoughts*.

2. Afreen, F., Shaikh, A., & Patel, R. (2025). A prediction of sleep cycle pattern by using machine learning based on work-related stressors. *International Journal of Engineering Research & Technology*.

3. Alafari, F., Khan, M. A., & Rahman, S. (2025). Advances in natural language processing for healthcare: A comprehensive review. *Computer Science Review*.

4. Alhumaidi, N. H., Al-Zahrani, F. A., & Mohammed, H. A. (2025). Machine learning–based analysis of lifestyle risk factors for cardiovascular disease. *JMIR Medical Informatics*.

5. Bozyel, S., Erdogan, T., & Yilmaz, K. (2025). Artificial intelligence-based clinical decision support systems in cardiovascular prevention. *Anatolian Journal of Cardiology*.

6. Budihal, S. V., Shetty, P., & Rao, M. (2025). A systematic review of machine learning models for heart disease prediction. *International Journal of Computer Applications*.

7. Cardamone, N. C., Wilson, J. P., & Smith, K. L. (2025). Classifying unstructured text in electronic health records for mental health prediction models. *JMIR Mental Health*.

8. Chowdhury, E., Rahman, M., & Islam, S. (2025). Risk prediction of cardiovascular disease for diabetic patients. *IEEE Access*.

9. Görmez, Y., Işık, Z., & Temiz, M. (2025). Prediction of obesity levels based on physical activity and diet with XAI. *Applied Sciences*.

10. Huang, W., Li, X., & Zhang, Y. (2022). Application of ensemble machine learning algorithms on lifestyle factors for cardiovascular risk prediction. *Scientific Reports*.

11. Hwang, S. H., Park, J., & Kim, H. (2024). Machine learning–based prediction for incident hypertension. *JMIR Medical Informatics*.

12. IEEE Sensors Team. (2023). A novel embedded deep learning wearable sensor for fall detection. *IEEE Internet of Things Journal*.

13. IEEE Sensors Team. (2025). Deep learning approach for detecting work-related stress using multimodal signals. *IEEE Sensors Journal*.

14. Javed, R., Khan, A., & Ali, M. (2025). Enhancing chronic disease prediction in IoMT-enabled healthcare. *IEEE Access*.

15. Katori, M., Shi, S., & Spira, A. P. (2022). The 100,000-arm acceleration dataset in the UK Biobank for sleep phenotyping. *Proceedings of the National Academy of Sciences (PNAS)*.

16. Krones, F., Weber, L., & Schmidt, M. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion*.

17. Lauritsen, S. M., et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*.

18. Lausanne University Hospital Research Team. (2026). An artificial intelligence-powered learning health system to enhance sepsis care. *npj Digital Medicine*.

19. Lee, H., Park, S., & Choi, J. (2023). Real-time machine learning model to predict in-hospital cardiac arrest using continuous physiological signals. *npj Digital Medicine*.

20. Liu, J., Chen, X., & Wang, Y. (2025). Comparing text-based clinical risk prediction in critical care. *IEEE Transactions on Biomedical Engineering*.

21. Liu, L., Thompson, P., & Manning, C. D. (2025). Using natural language processing to extract information from free-text clinical notes. *Journal of the American Medical Informatics Association (JAMIA)*.

22. Liu, Q., Zhang, H., & Li, W. (2024). A review of applying large language models in healthcare. *IEEE Access*.

23. Nandwani, P., & Verma, R. (2025). Research on sentimental analysis on mental health using social media. *International Journal of Scientific Research*.

24. Nature Scientific Reports Research Team. (2025). ObeRisk: A machine learning framework for predicting susceptibility to obesity. *Scientific Reports*.

25. Nature Scientific Reports Research Team. (2025). SpinachXAI-Rec: A multi-stage explainable AI framework for diet recommendation. *Scientific Reports*.

26. Nauman, M., Abbas, A., & Khan, S. (2025). The role of big data analytics in revolutionizing diabetes prediction. *IEEE Access*.

27. Nipa, N., Rahman, T., & Hossain, S. (2024). Clinically adaptable machine learning model to identify diabetes. *Informatics in Medicine Unlocked*.

28. Niu, K., Guo, X., & Tian, Y. (2023). Deep multi-modal intermediate fusion of clinical record and time-series data in early prediction of sepsis. *Journal of Biomedical Informatics*.

29. PeerJ Computer Science Research Team. (2025). Enhancing healthcare data privacy and interoperability with federated learning. *PeerJ Computer Science*.

30. Qin, Y., Wu, J., & Xiao, W. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *Frontiers in Public Health*.

31. Research Team. (2022). Depression detection from social media text analysis using NLP and hybrid deep learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

32. Research Team. (2022). Integration and validation of an NLP machine learning suicide risk prediction model in clinical workflows. *Frontiers in Digital Health*.

33. Research Team. (2024). Screening for depression using natural language processing of clinical notes. *Journal of Medical Internet Research*.

34. Research Team. (2025). Explainable AI-driven depression detection from social media text. *Information Processing & Management*.

35. Research Team. (2025). Psychological and behavioral insights from social media users: NLP-based study. *JMIR Medical Informatics*.

36. Sung, S. F., et al. (2021). Natural language processing improves prediction of stroke in unstructured text from electronic health records. *Journal of the American Medical Informatics Association*.

37. Systematic Review Team. (2021). Machine learning–based early warning systems for clinical deterioration: A systematic scoping review. *JMIR Medical*

Informatics.

38. Vangeepuram, N., Liu, B., et al. (2021). Predicting youth diabetes risk using NHANES data and machine learning. *Scientific Reports*.

39. Wang, X., Chen, L., & Zhang, M. (2025). Effect of artificial intelligence driven therapeutic lifestyle change: A review. *Frontiers in Public Health*.

40. Yuan, H., Tai, W. P., & Jonasson, C. (2024). Self-supervised learning of accelerometer data provides new insights for sleep and its relationship with mortality. *npj Digital Medicine*.

41. Zeydan, E., Kilic, S., & Arslan, S. S. (2024). Managing distributed machine learning lifecycle in healthcare: From data engineering to deployment. *IEEE Communications Surveys & Tutorials*.

42. Zhou, S., Wang, Y., & Liu, H. (2025). Large language models for disease diagnosis: A scoping review. *npj Digital Medicine*.

43. Zou, Y., Li, J., & Wang, Z. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-based topic model. *Scientific Reports*

CONFERENCE

44. Adlakha, M., Singh, S., & Sharma, M. (2023). Deep learning approach for accurate prediction of diabetes. *ACM International Conference Proceedings*.

45. Hu, G., Chen, Y., & Wang, L. (2025). Exploring approaches to computational representation and analysis of diet. *IEEE International Conference on Healthcare Informatics (ICHI)*.

46. Research Team. (2025). Optimizing stroke risk prediction: A machine learning framework. *arXiv (Submitted to IEEE Conference)*.

47. Research Team. (2026). A graph-language based benchmark for nutritional health. *arXiv (Submitted to IEEE Transactions on Affective Computing/Conference)*.

48. Shekhar, R., Kumar, V., & Singh, A. (2024). Human activity recognition using deep learning techniques for healthcare applications. *IEEE Conference on Computing*.

49. Shoham, O. B., Avital, G., & Shomron, N. (2024). CPLLM: Clinical prediction with large language models. *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.