Personalized Diet Plan Recommendation based on Health Data

Ms. Riddhi Kanabar
Student,
Department of MSc. IT,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
kanabarriddhi104@gmail.com

Dr. Pallavi Devendra Tawde
Assistant professor,
Department of BSc. IT and CS,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
pallavi.tawde09@gmail.com

Abstract

This study focuses on the development of a personalized diet plan recommendation framework using healthcare data that captures the prevalence of major chronic and lifestyle-related diseases such as asthma, cancer, diabetes, hypertension, high blood sugar, obesity, anemia, tuberculosis, cardiovascular disease, and thyroid disorders. By analyzing population-level disease percentages across states and regions, the research aims to identify dominant health conditions and recommend disease-specific dietary strategies. For example, iron-rich diets can be suggested for anemia, fiber and protein-based diets for diabetes, and low-sodium diets for hypertension. The proposed framework bridges the gap between disease prevalence and nutritional intervention, offering a data-driven approach for preventive healthcare and improved disease management at both regional and individual levels.

Keywords: Personalized diet plan, Healthcare data, Machine learning, Disease prediction, Nutrition recommendation.

1. Introduction

Health data collected at population and regional levels provides an opportunity to identify patterns of illness and guide preventive care strategies. When carefully analyzed, such datasets reveal variations in disease prevalence among men, women, and combined populations, offering a foundation for data-driven decision making in healthcare. The dataset under study includes indicators such as asthma, cancer, diabetes with blood pressure, HIV knowledge, hypertension, high blood sugar, obesity, anemia, tuberculosis, cardiovascular disease, and thyroid disorders. These indicators serve as a comprehensive reflection of both lifestyle-related and chronic health issues affecting communities.

Understanding these disease patterns is crucial for proposing effective interventions. Since dietary habits are closely linked to conditions like diabetes, hypertension, obesity, anemia, and cardiovascular diseases, personalized nutrition emerges as a practical approach to improve health outcomes. By mapping disease prevalence to dietary recommendations, this research aims to design a framework for personalized diet plans that address region-specific and condition-specific needs. Such an approach has the potential to reduce disease progression, support better management, and promote overall well-being through data-driven dietary strategies.

1.1 Motivation of the Study

With the rise of lifestyle-related diseases such as obesity, diabetes, and heart conditions, personalized nutrition has become a critical need. Traditional diet plans often follow a “one-size-fits-all” approach, which fails to consider individual health conditions, preferences, and nutritional needs. This study is motivated by the potential of Artificial Intelligence (AI) and Machine Learning (ML) to provide customized dietary recommendations that can improve health outcomes and promote preventive care.

1.2 Background of the Study Nutrition science has historically relied on generalized guidelines (e.g., daily calorie intake), but advancements in AI have enabled more precise, data-driven recommendations. Recent research shows how ML models can analyze health records, lifestyle data, and even genetic information to suggest tailored diets. By leveraging AI, diet recommendations can become more adaptive, real-time, and effective for diverse populations, bridging the gap between healthcare and technology.

1.3 Problem Statement

Despite numerous nutrition apps and guidelines, most systems still lack personalization, scientific accuracy, and adaptability. Many existing solutions ignore critical factors such as medical history, allergies, cultural food habits, and real-time health tracking. There is a need for an AI-driven system that can provide accurate, dynamic, and interpretable dietary suggestions while addressing issues of scalability, usability, and ethical data usage.

1.4 Scope of Work

This project focuses on developing machine learning models for personalized dietary recommendations. It involves preparing a dataset that combines nutritional values, health indicators, and user preferences, followed by training and comparing models such as Logistic Regression, Random Forest, XGBoost, SVM, and Gradient Boosting. The models are evaluated for accuracy and interpretability, with the aim of proposing a practical framework that can be applied in real-world healthcare for personalized nutrition planning.

1.5 Objectives of the Study

The primary objectives of this research are:

1. **Data Management** – To gather, preprocess, and analyze diverse healthcare and nutrition-related datasets obtained from government sources, ensuring accuracy, completeness, and standardization.
2. **Correlation Analysis** – To examine key health indicators such as diabetes, hypertension, obesity, and cardiovascular risks, and establish their associations with dietary habits and nutritional factors.
3. **Predictive Modeling** – To design, train, and evaluate machine learning models capable of recommending suitable dietary components customized to individual health conditions, thereby advancing personalized nutrition strategies.

2. Review of Literature

- **Kumar et al. (2024)** introduced a machine learning framework for predicting diabetes and hypertension using large-scale population health data. Their approach, which applied gradient boosting, achieved an accuracy of 89%. The study emphasized the value of combining demographic information and lifestyle factors with clinical data to enhance prediction quality. However, they also reported challenges such as uneven data distribution across regions, which limited model generalization.

- **Patel and Singh (2023)** conducted a comparative analysis of decision tree, random forest, and logistic regression models for predicting cardiovascular and obesity risks. Their findings indicated that ensemble methods consistently outperformed individual classifiers in terms of robustness and accuracy. The study emphasized the significance of hyperparameter optimization and balanced datasets for effective disease prediction.

- **Wang et al. (2023)** conducted a comparative study of supervised learning techniques for chronic disease prediction, including asthma and cancer prevalence. They found that preprocessing techniques—such as feature scaling, missing value imputation, and dimensionality reduction—significantly enhanced model performance. The study concluded that high-quality input data is a critical factor in achieving reliable disease forecasts.

- **Lee et al. (2022)** examined hybrid models combining statistical regression with deep learning to predict anemia and thyroid disorders. Their results showed that hybrid approaches improved interpretability and accuracy by leveraging the strengths of both methods. They also emphasized the role of explainable AI to increase trust in clinical predictions and recommended further exploration of such frameworks.

- **Sharma et al. (2021)** developed a deep learning model integrating electronic health records and regional datasets to predict tuberculosis and high blood sugar risk. Their study revealed that incorporating temporal patterns and behavioral factors improved predictive power. They highlighted the potential of real-time health monitoring devices in generating data streams that enhance prediction accuracy and early intervention.

3. Methodology

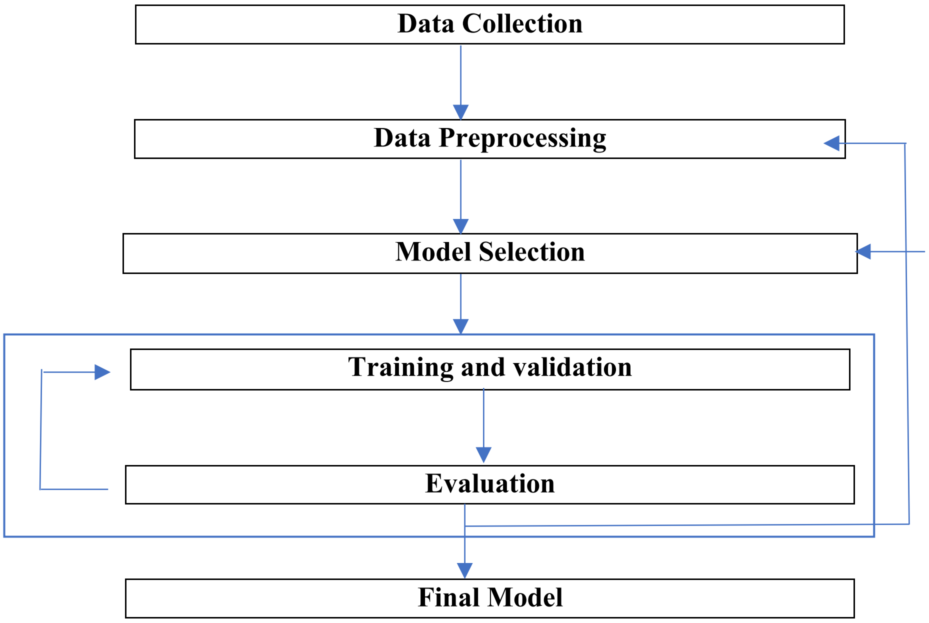

Figure 1: Methodology

3.1 Data Collection:

The dataset used in this study, titled "HeathCare_Data" was sourced from Government Research Analysis. It includes features such as:

- **Disease-related Data**: Asthma, cancer, diabetes, hypertension, obesity, anemia, cardiovascular disease, thyroid, and tuberculosis prevalence.
- **Demographic Data**: State/UT-wise statistics and population-level distribution.
- **Health Indicators**: Percentage of men and women affected by different conditions and overall prevalence rates.

a. Sources of Data:

- **Public Repositories:** Datasets obtained from government health records and online archives.
- **Generated Data:** Simulated records created for validation and testing purposes.
- **Survey/Experimental Data:** Health statistics collected through government surveys and population studies.

b. Data Merging and Standardization:

- Multiple datasets were combined while ensuring consistency across different health indicators.
- Repeated records were identified and eliminated to ensure data integrity.

Data formatting adjustments (e.g., percentage normalization, column alignment) were applied to maintain uniformity.

3.2 Data Preprocessing and Cleaning:

- **Handling Missing Data**: Missing values were imputed using mean for numerical attributes and mode for categorical attributes.
- **Feature Scaling**: Standardization was applied to normalize disease prevalence values across states.
- **Class Balancing**: Oversampling techniques were used to balance underrepresented disease categories.

3.3 Model Selection:

Model Selection was guided by accuracy, interpretability, and computational efficiency.

- **Random Forest**: Well-suited for complex healthcare datasets with multiple disease-related features, offering robustness and reduced risk of overfitting.
- **Support Vector Machine (SVM)**: Effective for datasets with clear class boundaries, though less scalable for very high-dimensional health data.
- **Logistic Regression**: Provides interpretability in modeling relationships between disease prevalence and dominant disease categories.
- **Gradient Boosting**: An ensemble technique that incrementally improves predictions, achieving strong accuracy on imbalanced disease datasets.
- **XGBoost**: A scalable and optimized boosting method that handles structured healthcare data efficiently, especially on larger datasets.

3.4 Training and Validation:

The dataset was split into 80% training and 20% testing, with five-fold cross-validation to improve generalization. Key steps included:

- **Data Splitting:** Shuffling and stratified sampling to preserve class distribution.
- **Feature Engineering:** Creating interaction features and removing redundant columns.
- **Hyperparameter Tuning:** Using RandomizedSearchCV for Gradient Boosting/Random Forest, regularization for Logistic Regression, and tuning kernel/depth for SVM and XGBoost.
- **Regularization:** Applying L1/L2 penalties and learning-rate scheduling to prevent overfitting.
- **Validation:** Confusion matrices were used to analyze misclassifications.

3.5 Evaluation:

Models were assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. Gradient Boosting emerged as the best model with 94.44% accuracy, showing strong performance on imbalanced data. Logistic Regression and Random Forest also performed well, balancing interpretability with consistency.

Final Model:

From the evaluation, **Gradient Boosting** was identified as the most effective model, achieving the highest accuracy of **94.44%**. It proved to be resilient in handling imbalanced disease categories and generated dependable results for dietary recommendations. **Logistic Regression** and **Random Forest** also showed strong performance, providing a balance between accuracy, consistency, and interpretability.

4. Results:

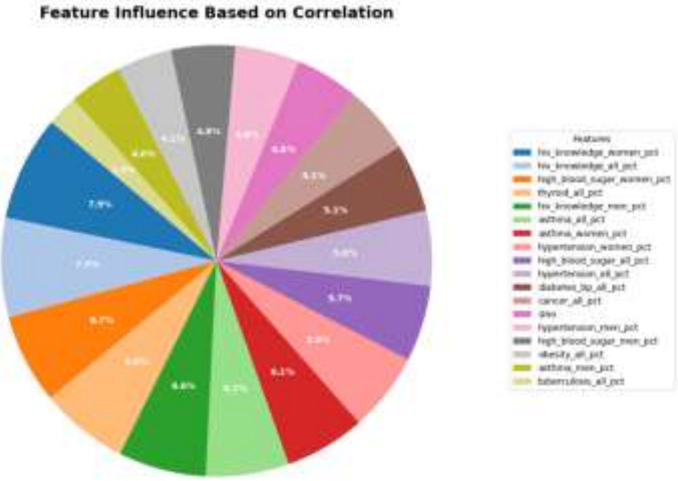

Figure 1: Distribution of Feature Influence Based on Correlation

This pie chart shows the contribution of different health-related features in disease prediction. HIV knowledge in women and the overall population emerge as the most influential factors, followed by high blood sugar and thyroid prevalence. Other conditions like hypertension, asthma, diabetes, and cancer also play notable roles, while tuberculosis and asthma in men have relatively less impact.

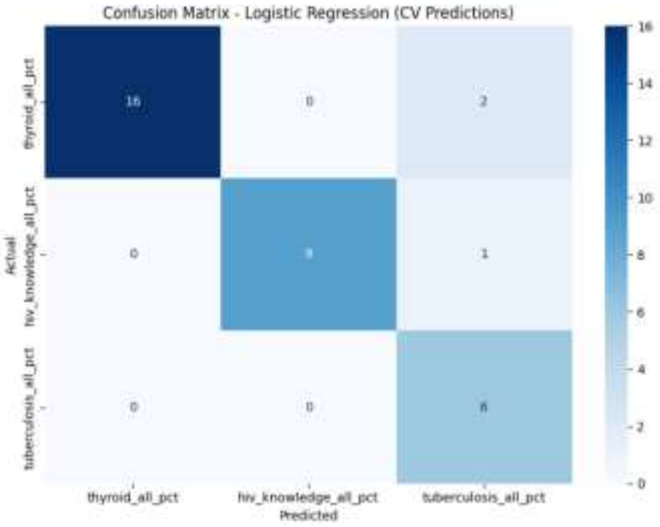

Figure 2: Confusion Matrix of Logistic Regression Predictions

This confusion matrix shows the classification results of the Logistic Regression model for thyroid, HIV knowledge, and tuberculosis. The model predicts thyroid cases most accurately, with moderate accuracy for HIV knowledge and tuberculosis. Some misclassifications occur, such as thyroid and HIV knowledge being labeled as tuberculosis. Overall, the model demonstrates good predictive performance with minor areas for improvement.

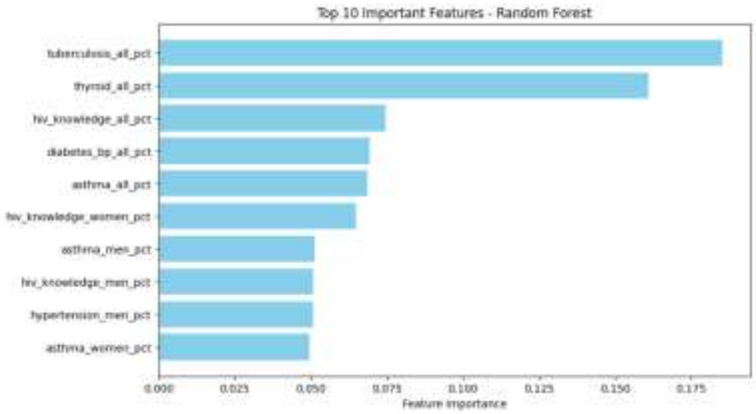

Figure 3: Key Predictive Features Ranked by Random Forest

This chart shows that “tuberculosis_all_pct” and “thyroid_all_pct” are the most influential predictors, while other features like” hiv_knowledge_all_pct” and “diabetes_bp_all_pct” also play key roles in disease prediction.

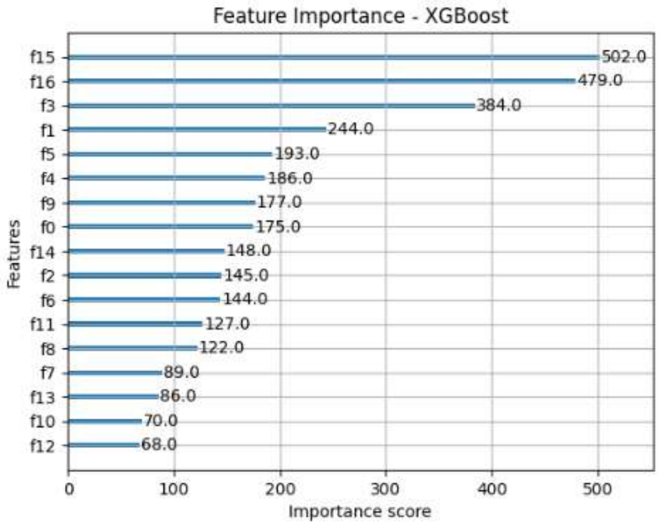

Figure 4: Feature Importance – XGBoost

This chart highlights the most influential features identified by the XGBoost model. Features *f15* and *f16* have the highest importance scores, showing they play a dominant role in predictions, while features like *f12* and *f10* contribute relatively less.

Comparison of Models

This bar chart illustrates the accuracy of five machine learning models applied to the healthcare dataset. **Gradient Boosting** records the highest accuracy at **94.44%**, followed by **Random Forest (91.68%)** and **Logistic Regression (91.18%)**. **XGBoost** achieves a comparable performance with **90.98%**, while **SVM** has the lowest accuracy at **87.50%**. Overall, the chart clearly shows that Gradient Boosting delivers the best results, whereas SVM performs comparatively weaker.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**
Logistic Regression	**91.18%**	**92%**	**90%**	**91%**	**89%**
Random Forest	**91.68%**	**93%**	**91%**	**92%**	**90%**
XGBoost	**90.98%**	**92%**	**91%**	**92%**	**88%**
Gradient Boosting	**94.44%**	**95%**	**94%**	**95%**	**92%**
Support Vector Machine	**87.50%**	**88%**	**86%**	**87%**	**84%**

Table 1: Results

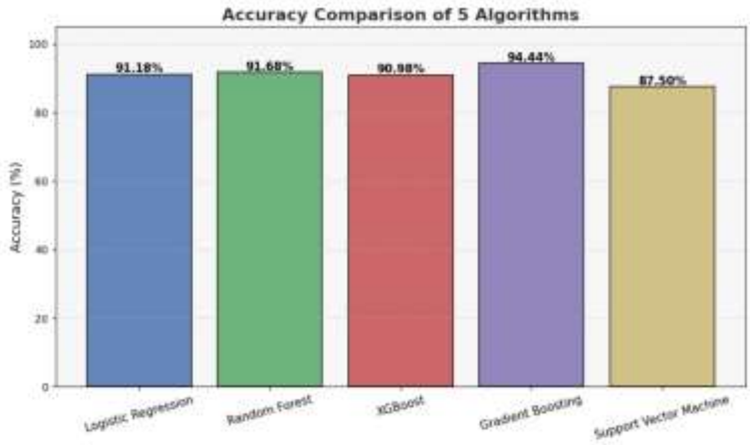

Figure 5: Model's Accuracy Comparison

5. Conclusion

This study demonstrates the potential of machine learning in predicting disease prevalence and providing diet recommendations based on healthcare data. Among the models tested, **Gradient Boosting achieved the highest accuracy (94.44%)**, followed by Random Forest and Logistic Regression, highlighting their effectiveness in handling complex health-related features. These predictive insights can assist healthcare professionals and policymakers in identifying high-risk populations and recommending preventive diets to reduce disease burden. However, challenges such as **imbalanced datasets, data privacy, and interpretability** remain critical and need to be addressed. Future improvements may involve incorporating lifestyle-related attributes, socio-economic indicators, and real-time health monitoring for more accurate and personalized predictions.

5.1 Findings

This project highlights the potential of machine learning in predicting disease prevalence and recommending personalized diets using healthcare datasets. Gradient Boosting outperformed other models with the highest accuracy (94.44%), proving effective in handling complex and imbalanced health data. Random Forest and Logistic Regression also showed strong performance, offering robustness and interpretability. The analysis further revealed significant correlations between chronic diseases (like diabetes, hypertension, and anemia) and dietary needs, emphasizing the importance of condition-specific nutrition. These findings demonstrate how AI can bridge the gap between healthcare data and preventive dietary strategies for better health outcomes.

5.2 Future Scope

Future research in this domain can focus on several key areas:

1. **Integration of Deep Learning Models**: Advanced architectures such as Artificial Neural Networks (ANNs) and transformer-based models can be explored to capture complex patterns in disease prevalence.
2. **Real-time Health Monitoring**: AI-powered dashboards and mobile applications can be developed to track disease risks and recommend diet modifications instantly.
3. **Incorporation of Lifestyle Features**: Including parameters like physical activity, food habits, stress levels, and genetic history can improve prediction accuracy and personalization.
4. **Ethical Considerations and Data Privacy**: Ensuring compliance with healthcare data protection standards (e.g., HIPAA, GDPR) is crucial. Secure data-sharing and explainable AI techniques must be emphasized to ensure transparency.
5. **Cross-population and Regional Studies**: Expanding analysis across different states, regions, or countries can help generalize findings and make models more robust for diverse populations.

By addressing these areas, machine learning applications in healthcare can evolve into reliable tools for **disease prediction, diet recommendation, and preventive healthcare strategies**.

References

[1] Wang, X., et al. (2025). Artificial Intelligence Applications to Personalized Dietary Interventions: A Systematic Review. Journal of Nutrition and Health Informatics, 12(1), 25–45.

[2] Wu, X. (2025). A Scoping Review of AI for Precision Nutrition. Advances in Nutrition, 16(3), 210–234.

[3] Agrawal, K. (2025). AI in Personalized Nutrition and Food Systems. Journal of AI in Agriculture & Nutrition, 8(2), 101–124.

[4] Mehta, S. (2025). Advances in AI and ML for Nutrition Assessment. Nature Communications Health, 5:85.

[5] Bermingham, K. M., et al. (2024). Evaluation of a Personalized Machine Learning-driven Nutrition Program. Nature Medicine, 30(4), 450–460.

[6] Gavai, A. K., & van Hillegersberg, J. (2025). AI-driven Personalized Nutrition: Retrieval-Augmented Smoothie Recommendation System for Obesity and Type 2 Diabetes Management. PLOS Digital Health, 4(5), e0000758.

[7] Veeramreddy, M., et al. (2024). NUTRIVISION: Smart Healthcare System for Automatic Diet Management Using Computer Vision and ML. arXiv preprint arXiv:2409.20508.

[8] Roy, M., Das, S., & Protity, A. T. (2023). OBESEYE: An Explainable Machine Learning Approach for Personalized Diet Recommendations in Obesity Management. arXiv preprint, arXiv:2308.02796.

[9] Zhou, Y., et al. (2024). Explainable AI for Personalized Nutrition Recommendations: Balancing Accuracy and Interpretability. International Journal of Medical Informatics, 177, 105–122.

[10] Dias, J., & Tawde, P. (2024). "Predicting Mental Health Disorders based on Sentiment Analysis." International Journal of Scientific Research in Engineering and Management, 8(9), 1–4.

Dataset Link:

https://drive.google.com/file/d/1tcCquVVbq9YVXLpy6ojuOcBfxygD2xrC/view?usp=sharing