

"Beyond the Bite: Unraveling the Mysteries of Food Allergy Through Epigenetics and Machine Learning"

By
Farjana Nasrin
ID : 213-15-4514

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the
Requirements for the **Degree of Bachelor of Science**
in Computer Science and Engineering

Supervised by
Md. Sazzadur Ahamed

Assist. Professor
Department of Computer Science and
Engineering Daffodil International
University

Co-Supervised by

Dr. Md. Fokhray Hossain

Professor
Department of Computer Science and
Engineering Daffodil International
University



**DAFFODIL INTERNATIONAL
UNIVERSITY**
Dhaka, Bangladesh

September 16, 2025

APPROVAL

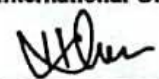
This Project titled "**Beyond the Bite: Unravelling the Mysteries of Food Allergy Through Epigenetics and Machine Learning**", submitted by Farjana Nasrin , ID No : 213-15-4514 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16 September, 2025.

BOARD OF EXAMINERS



Dr. Bimal Chandra Das (BCD)
Professor and Dean (in-charge)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



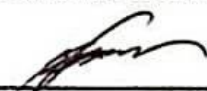
Most. Hasna Hena (HH)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Mayen Uddin Mojumdar (MUM)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



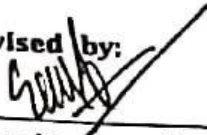
Nazibur Rahman
Technical Lead - Database Administrator
Telenor - Grameen Phone Account

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project or any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

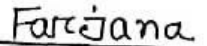


Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and
Engineering Daffodil International
University

Co-Supervised by:

Dr. Md. Fokhray Hossain
Professor
Department of Computer Science and
Engineering Daffodil International
University

Submitted by:



Farjana Nasrin
ID: 213-15-4514
Department of Computer Science and
Engineering Daffodil International
University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, I express my heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for me to complete the **Final Year Design Project(FYDP)** successfully.

I am grateful and wish my profound indebtedness to **Md. Sazzadur Ahamed** , **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Food allergies are a growing public health issue and their prevalence is on the increase worldwide as a result of complex, and interaction between genes, environment and lifestyle. Traditional diagnostic tools are still limited in scope and many times cannot discover the hidden biological mechanisms concealed allergic reaction. This research, *Beyond the Bite: Unraveling the Mysteries of Food Allergy Through Epigenetics and Machine Learning*, examines how factors such as epigenetic indicators and lifestyle elements can be combined with state-of-the-art computational models to better predict risk for allergy. A dataset of 1,830 actual responses were obtained from people using both online surveys and person to person interactions in order to have a wide range of demographics collected. The process involved preprocessing and cleaning the dataset, applying feature selection, training of multiple machine learning model Random Forest, XGBoost, Logistic Regression with validation by using accuracy, precision, recall, and confusion matrix. Of the models tested, Random Forest showed good performance, a good tradeoff between interpretability and predictive power. deep learning approaches show potential for scaling in the future. The results point to important links between diet, environmental exposures and allergy susceptibility and demonstrate the feasibility of data driven, multi-omic approaches to elicit hidden risk factors. These findings suggest that machine learning can complement traditional medical practices and can provide predictive information that can be useful in preventive healthcare, policy development, and public awareness campaigns. Ultimately, this research does play a role in 'bridging the gap between computational biology and clinical practice' with the potential for interdisciplinary approaches in the complex nature of health issue to be addressed.

Table of Contents

Approval	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Introduction.....	1
1.2 Motivation.....	2
1.3 Objectives.....	2
1.4 Methodology.....	3
1.5 Project Outcome.....	3
1.6 Organization of the Report.....	4
2 Background	5
2.1 Introduction.....	5
2.2 Literature Review.....	6
2.3 Gap Analysis.....	9
2.4 Summary.....	10
3 Research Methodology	11
3.1 Methodology.....	11
3.1.1 Overview.....	11
3.1.2 Proposed Methodology.....	11
3.1.3 Functional and Nonfunctional Requirements.....	14
3.1.4 Data Flow Diagram.....	15
3.2 Detailed Methodology and Design.....	16
3.3 Project Plan.....	18
3.4 Task Allocation.....	19
3.5 Summary.....	19

4	Implementation and Results	20
4.1	Environment Setup.....	20
4.2	Testing and Evaluation/Performance/ Comparative Analysis.....	21
4.3	Results and Discussion.....	22
4.4	Summary.....	29
5	Engineering Standards and Design Challenges	30
5.1	Compliance with the Standards.....	30
5.1.1	Software Standards.....	30
5.1.2	Hardware Standards.....	31
5.1.3	Communication Standards.....	31
5.2	Impact on Society, Environment and Sustainability.....	32
5.2.1	Impact on Life.....	32
5.2.2	Impact on Society & Environment.....	33
5.2.3	Ethical Aspects.....	33
5.2.4	Sustainability Plan.....	33
5.3	Project Management and Financial Analysis.....	34
5.4	Complex Engineering Problem.....	34
5.4.1	Complex Problem Solving.....	34
5.4.2	Engineering Activities.....	36
5.5	Summary.....	38
6	Conclusion	39
6.1	Summary.....	39
6.2	Limitation.....	40
6.3	Future Work.....	40
	References	41

List of Figures

3.1	Methodology pipeline diagram.....	12
3.2	Data flow diagram.....	15
3.3	User Interface Design.....	16
4.1	Confusion Matrix Logistic Regression.....	24
4.2	Confusion Matrix Random Forest.....	26
4.3	Confusion Matrix XGBoost.....	28
4.4	Number of food allergies by age	30

List of Tables

2.1 Summary of Literature Reviewed.....	8
2.2 Gap Analysis Table.....	9
3.1 Project plan Table.....	19
3.2 Task Allocation Table.....	20
4.2 Comparative Model Performance.....	22
4.3 Model Performance Table.....	24
5.1 Estimated Cost and Financial Analysis.....	35
5.2 Mapping with Complex Engineering Problem.....	35
5.3 Mapping with knowledge Profile.....	36
5.4 : Mapping with Complex Engineering Activities.....	38

Chapter 1

Introduction

This chapter introduces the foundation of our research by presenting the background of the research problem, its motivation, objectives, adopted methodology, and expected outcomes. It also outlines the overall structure of the report to guide the reader through upcoming chapters.

1.1 Introduction

During recent times, food allergies have emerged as a growing serious public health concern that is affecting millions across worldwide. Food allergies are immune reactions to specific food proteins that are recognized as inoculation (mistaken for dangerous) by the body. This hypersensitivity can bestow up a whole range of symptoms from relatively mild states like skin rashes, hives and sensory esophagus disorders that may cause; severe and life threatening disorders such as anaphylaxis. The increasing frequency of food allergies, especially in children and youth, has led to concern about the etiology of food allergies, potential risk factors for the development of food allergies, and strategies to prevent food allergy development.

It was once thought that the development of food allergies is primarily caused by predisposition. While genetics play a major role, advances in recent years have expanded this viewpoint to examine how much influence epigenetic factors play. Epigenetics is a term designating heritable alterations in gene expression that do not involve alterations in the underlying sequence of the DNA coding for the genes. Elements such as maternal health during pregnancy, environmental exposures early in life, antibiotic use, and dietary habit can all affect epigenetic markers which in turn, affect the development of the immune system. These evidence suggest that the onset of food allergy is not simply a consequence of inherited inheritance of genetic factors but rather the consequence of complex interaction between inherited susceptibility by genetic factors and environmental factors.

Although there has been progress in identifying possible contributing factors, the exact biological and environmental process that drives development of food allergies are poorly understood. This lack of knowledge is limiting effectiveness of early detection, risk evaluation and preventive interventions. Traditional diagnostic methods often rely heavily on clinical history and lab testing, but traditional methods do not reflect the environmental and lifestyle aspects that can have a major impact on allergic outcome. As a result there is a pressing need for innovative strategies to include both biological signs and contextual factors when attempting to better predict and manage food allergy risks.

In this context time and healthy, machine learning (ML) provides temporal a festinstens of distributing and sophisticated endpoint to pursussi csan examine broad a as well as well complicated data reveal possible selected trends that no locate obvious in a traditional methods of statistical methods. By combining survey based data with epigenetic related variables such as dietary habits, antibiotic exposure and maternal health, machine learning can support the creation of predictive models capable of estimate individual more accurately food allergy risk. Such models provide not only the potential to identify high risk individuals in the early stages but may also become a preventive technique, and can inform public health policy.

The study aims to fill this gap at the cost of using a machine learning, data driven approach to investigate the risk factors of food allergy. The study uncovers significant associations between environmental exposures, lifestyle practices, and allergy outcomes using a dataset of real world survey responses that asked willingness and voluntarily provided by people. The eventual objectives is to make a contribution on the development of a predictive model of assessing the risk of food allergy. That way, it is hoped that through it, our current knowledge about food allergy etiology will evolve because the computational approach to healthcare and prevention can be seen as a promising tool in assisting in the individualized management and prevention of food allergies.

1.2 Motivation

Increasing prevalence of food allergies is a major problem to both the medical care givers and victims of such ailments. But even though it is important, there are only few methods that can be used today to understand and predict the risks of allergy in food. This drawback is a good prospect to use computational methods, especially machine learning, to determine patterns in lifestyle, environment, and family data that could contribute to the development of allergies.

Machine learning algorithms are very appropriate to find out the hidden relationship among multidimensional data. With such methods, it is possible to construct predictive models that do not only examine the likelihood of a person developing allergies to food but also the major contributors. Such a computing approach introduces increased precision, efficiency, and scalability over the more traditional approaches to clinical risk assessment.

On both personal and academic levels, the project will provide me with the opportunity to explore the intersection of healthcare, information science, and epigenetics three disciplines that are coming together on the contemporary research scene, as well as make a contribution toward a socially significant problem. the development of this system will offer me to practice data preprocessing, model building, and critical analysis, not to mention the contribution to a socially significant issue. Finally, the solution of this issue could contribute to the early diagnosis, better management, and the increased awareness of people about food allergies.

1.3 Objectives

The primary objectives of this research are:

- To establish important lifestyle, environmental, and family related features of food allergies.
- To investigate potential epigenetic initiators, according to early-life and maternal status.
- To build a machine learning model that will be able to forecast the risks of food allergy depending on the gathered information.
- To offer the information that could be used in the early intervention and community health measures.

1.4 Methodology

A survey is used in this study where structured questionnaire data on patients who report their history of allergy, family history, diet, use of antibiotics among other environmental issues are collected and analyzed. The methodology includes:

- Google Forms data collection (1830 data).
- Categorical variables data cleaning and encoding.
- Exploratory Data Analysis (EDA) to discover trends and correlations. Including: (1) training and testing machine learning algorithms (e.g. Random Forest, Logistic Regression,XGBoost) to infer the probability of allergy.
- Results interpretation in the framework of the possible epigenetic effects.

1.5 Project Outcome

This research is expected to yield the following outcomes:

- Creation of a usable machine learning model that can predict the likelihood of a person to experience food allergies; using survey data.
- Non comparative: Improved understanding of how lifestyle decisions, plus environmental or epigenetic-like factors contribute to the development and progression of food allergies.
- In addition to this, it will provide a groundwork that can be used in future research involving even more sophisticated genomic or epigenomic data.
- Practical knowledge that can be used to educate population health efforts, specifically, in the sectors of early screenings, risk reduction, and prevention of allergies approaches.

1.6 Organization of the Report

This research report is organized into six chapters directed toward each of the main aspects of the final year design project and representing a comprehensive study of the issue of food allergy prediction as a result of the application of the epigenetic analysis and machine learning method.

- i. **Chapter 1: Introduction**
This section gives a general discussion of the research, its background, problem statement, motivation, objectives, methodology, the expected results, and general format of the report. It creates a setting and importance of the research.
- ii. **Chapter 2: Background**
In this section, background details of food allergies, their biological pathways, and diagnostic difficulties are given in detail. It consists of literature survey, the analysis of similar apps, and the gap analysis that shows the lack of combined epigenetic and machine learning systems to predict allergies.
- iii. **Chapter 3: Research Methodology**
This section identifies the technical structure of the research, the proposed methodology, system design, and functional and non-functional requirements, data flow diagrams. It also describes why a particular set of algorithms, data and methods of analysis, are chosen.
- iv. **Chapter 4: Implementation and Results**
This section describes the environment configuration, the system deployment and the outcome of testing and analysis. It provides a performance analysis of the created machine learning models and comments on the findings implications.
- v. **Chapter 5: Engineering Standards and Design Challenges**
This chapter presents the applicable engineering standards, ethical issues and sustainability. It also describes the complicated engineering issues that were dealt with and aligns them to knowledge profiles and evaluates the societal and environmental consequences of the project.
- vi. **Chapter 6: Conclusion**
This section provides an overview of the work in general, including the shortcomings of the existing system and discuss the possible future developments, research, and the possibility of scaling.

Chapter 2

Background

This section lays the necessary foundation to understand the scope and motivation of our research. It introduces the key concepts of food allergies, epigenetics, and machine learning, followed by a review of remaining literature and the identification of research gaps that our proposed system aims to address.

2.1 Introduction

Food allergies are the allergy of the body in response to certain proteins contained in foods that are not harmful to most people. With an estimated worldwide effect on millions of individuals, they present with immensely important health issues of the populace and may result in life threatening outcomes, including anaphylaxis. In addition to the direct medical risks, food allergies cause important changes in the lifestyle, socialization, and general quality of life. The existing diagnostic tools like skin prick test and serum specific IgE measure provide helpful data, but not as accurate as needed; they usually give false positive or negative results, which prolong the process of effective treatment and management.

Epigenetic studies have recently provided the field of allergic diseases with a new dimension. Epigenetics is the research that evaluates how the external and other environmental elements like diet, pollution, and early exposures of life can alter the expression of the genes without altering the fundamental sequence of DNA. Such modifications such as DNA methylation and regulation of microRNA can be of vital importance in the onset, exacerbation and resolution of allergy. A combination of these molecular insights and the use of computational approaches provides a radical opportunity. Machine learning (ML) with its ability to detect intricate patterns in high dimensional data offers potent means to forecast the threat of allergy, increase diagnostic accuracy and assist in customized methods of therapy. Epigenetics, with the aid of ML, therefore, represents a promising avenue of more precise, personalized, and aggressive approaches to food allergy studies and clinical application.

2.2 Literature Review

In this section, the reviews of the past studies in the fields of food allergies, epigenetics, and machine learning based predictive modeling can be found. It is the goal to shed some light on available contributions, methods, and findings applicable to our research.

To fill in the gaps in this area, Martino et al. [1] created and trained machine learning models to detect epigenetic footprints of food allergy, which was not a focus of other research previously. To this end, they used the methylation datasets and tested various classifiers, which include Logistic Regression, Support Vector Machines (SVM), and Decision Trees. Their analysis was able to define 18 CpG biomarkers that strongly relate with food allergy and the classifiers used in the analysis have almost 100 percent predictive accuracy. This publication highlighted the potential of CpG based epigenetic markers in developing an accurate diagnosis.

Patrawala et al. [2] were investigating the forecast of the result of the peanut oral food challenge, a gold-standard clinical test of peanut allergy, by machine learning algorithms. They have especially used ensemble methods with cross-validation schemes to make them robust. The dataset combined the IgE profiles with the appropriate clinical parameters with training AUC of 0.997 and validation of 0.87. Such findings showed that ML could significantly decrease the use of expensive and dangerous oral food challenges though the emphasis was on allergen-specific.

Naveenkumar et al. [3] introduced the deep learning based ProAll-D framework, which is used in the detection of allergens in proteins. Their algorithm used Long Short term memory (LSTM) networks with a boost of E-descriptors and autocorrelation to learn the sequential patterns of proteins. Classification accuracy of the system reached approximately 91.5 percent in differentiating allergenic proteins and non-allergenic ones, confirming the possibility of the system in terms of computationally screening allergens.

Three researchers, Raghuvarshi and Gupta [4], conducted a comparative study between deep learning and ensemble techniques in food protein allergenicity prediction. They compared transformer-based BERT with classic ML models such as LightGBM and XGBoost. The results of the experiment showed that BERT was significantly better than the ensemble methods, with an AUC greater than 0.95, which demonstrated the high quality of contextual sequence embeddings in allergenicity prediction tasks.

Singh and Sharma [5] shifted the focus to machine learning-based plant-based allergen protein prediction. They optimized classical classifiers such as SVM and k-NN on the same model inputs as physicochemical features. Their analysis has proven that these methods are effective in differentiating between plant allergens and non-allergens and hence aid with computational allergen profiling in plant-derived food.

Shaukat et al. [6] broadened the search on the use of machine learning in allergen detection by testing various algorithms, i.e., Random Forest, SVM, k-NN, and Decision Tree classifiers. They trained their model with datasets that had detailed food descriptors and the resulting structure was capable of up to 98% accuracy, hence offering a viable and very useful system to identify allergens in foods..

A systematic review done by Chang, Wu, and Lu [7] centered on epigenomic and epigenetic studies on food allergy. They reviewed a body of epigenome-wide association studies (EWAS) and microRNA studies demonstrating that significant antigen pathways and T-cell reactions are significant contributors to the severity of allergy. Additionally, they topped three epigenetic loci to the susceptibility to allergies. This research work gave a detailed biological basis on how to incorporate epigenetics in predictive modeling.

Wilkins and Smith [8] summarised the genome wide association studies (GWAS) in conjunction with the epigenetic loci in order to study the risk of allergy. Their review revealed a correlation that exists between IgE response, specific SNP, and DNA methylation signatures, thereby supporting the idea that genetics and epigenetics are mutually reliant in the formation of allergies. Despite the informational nature, the study was still descriptive without the development of computational models.

Khan et al. [9] had studied the use of artificial intelligence in the diagnosis of allergies in children, especially using electronic health record (EHR) and supervised learning methods. Their models demonstrated the possibility of the ML to identify clinical patterns to assist pediatric allergists in the early diagnosis. Although it was more clinical rather than molecular oriented, the work showed that AI could be used clinically in the care of allergies.

The machine learning models used by Inglis et al. [10] were Convolutional Neural Networks (CNN), neural networks, and XGBoost on the food safety field through the detection of mycotoxins. Although they do not directly address the problem of food allergies, their research was a great contribution to scalable ML pipelines to detect food hazards. Such methods are applicable to allergen prediction systems, and exhibit methodological transferability across domains.

Table 2.1: Summary of Literature Reviewed.

Author (s)	Year	Title	Methodology	Key Findings
Martino et al.	2019	Machine learning identifies epigenetic signatures of food allergy	Logistic Regression, SVM, Decision Tree	Identified 18 CpG markers; classifiers achieved ~100% accuracy for allergy prediction
Patrawala et al.	2024	Prediction of peanut oral food challenge outcomes using ML	Ensemble Models (with cross-validation)	AUC 0.997 (training), 0.87 (validation); predicted allergic response from IgE and clinical data
Naveenkumar et al.	2022	ProAll-D: Protein allergen detection using LSTM	LSTM with E-descriptors & autocorrelation	Achieved ~91.5% accuracy for allergen vs. non-allergen protein classification
Raghuvanshi & Gupta	2021	Deep Learning vs. Ensemble Learning in Food Protein Allergenicity	BERT, LightGBM, XGBoost	Transformer-based BERT achieved AUC > 0.95; outperformed traditional models
Singh & Sharma	2022	Plant allergen protein prediction via ML	SVM, k-NN	Used physicochemical features; effective classification in plant-based allergen detection
Shaukat et al.	2024	ML for accurate detection of food allergens	Random Forest, SVM, k-NN, Decision Tree	Achieved up to 98% accuracy in allergen detection from food descriptors
Chang, Wu & Lu	2023	Epigenomic and epigenetic investigations of food allergy	Systematic review (EWAS, microRNA analysis)	Identified T-cell and antigen pathways; epigenetic loci linked to allergy severity
Wilkins & Smith	2024	GWAS and epigenetics in food allergy	Literature review of GWAS + epigenetic loci	Connected IgE responses with SNPs and methylation signatures for allergy risk
Khan et al.	2020	AI applications in pediatric allergic disease diagnosis	ML in healthcare (EHR, supervised learning)	ML models can aid pediatric allergy diagnosis using clinical data and pattern detection
Inglis et al.	2024	ML for food safety via mycotoxin detection	CNN, neural networks, XGBoost	Though not allergy-specific, ML pipeline designs applicable for allergen detection systems

2.3 Gap Analysis

Although various applications and research efforts address allergy detection, management, or prediction, most fail to integrate epigenetic biomarkers with machine learning models for comprehensive food allergy prediction. The following table compares existing systems with our proposed approach:

Table 2.2: Gap Analysis Table

Features	Allergy Amulet	FoodAllergyScan	NutriGenome	AsthmaPredictor AI	Gene2Health	Proposed System
Detect allergens in food	Yes	No	No	No	No	Yes
Predict allergy risk using ML	No	No	No	Yes	Yes	Yes
Use of epigenetic biomarkers	No	No	No	No	Yes	Yes
Integration of multi-omics data	No	No	No	No	Yes	Yes
Personalized prevention strategy	No	No	Yes	No	Yes	Yes
Real-time risk assessment	Yes	No	No	No	No	Yes
Mobile and web accessibility	Yes	Yes	Yes	Yes	Yes	Yes
User reaction logging and tracking	Yes	Yes	No	No	No	Yes

2.4 Summary

This chapter defined the background of the research since it discussed the food allergies and how these allergies develop depending on epigenetics. It pointed out the ways in which genetic and environmental conditions communicate with each other via epigenetic processes, including DNA methylation, to influence immune reactions and allergic outcomes. Another review of machine learning applications in health prediction was reviewed in the chapter with several models, methodology and contribution to allergy detection. By so doing, it also defined gaps in existing studies, specifically the lack of integrations in the existing research involving lifestyle, survey-based, and molecular information. Competitive gap analysis revealed that the earlier research has aggressively developed on allergy prediction albeit not with a multidimensional approach. This can be used to highlight the necessity of having a more holistic and data driven model that integrates epigenetic, clinical, and lifestyle characteristics. Resolving these limitations, the given research may apply the newest machine learning methods to create a predictive model that delivers a better level of accuracy, interpretability and practicality in allergy forecasting and management on a personal level.

Chapter 3

Research Methodology

In this section, the research methodology followed in the study is described; hence the entire approach, system design, functional requirements, non-functional requirements, data flow, user interface considerations, alternative solution, and the project implementation plan.

3.1 Methodology

3.1.1 Overview

This study uses a multiomics combination model that is data driven to enhance the forecasting and treatment of food allergies. The research will involve an integration of both epigenetic biomarkers and advanced machine learning algorithms to produce a robust, precise and user friendly predictive tool. The methodology will solve the weaknesses of the existing allergy diagnostics, as most of them only use either clinical tests or patient reported symptoms.

3.1.2 Proposed Methodology

The system proposed is designed in a structured pipeline that is geared towards creating an accurate and reliable framework in predicting food allergy based on the integration of both epigenetic and clinical data with machine learning models. The pipeline comprises of various phases, the first one being the collection and compilation of the relevant data sets which involve both patient level clinical data and molecular biomarkers (e.g., DNA methylation profiles). This is followed by a sequential process, which involves manipulation of data such as data preprocessing whereby missing values are filled, features normalized, and categorical values coded to ensure uniformity throughout the dataset. Subsequently, the most important epigenetic and clinical markers that make individuals prone to allergy are chosen using feature selection methods. After preparing a dataset, the system proceeds to model training where three machine learning models, Logistic Regression, Random Forest, and XGBoost are applied and optimized to perform classification tasks. Both the models are strictly evaluated based on the accuracy, precision, recall, F1-score, and confusion matrices so that the models are robust and capable of prediction.

The last stage in the pipeline is that of creating allergy risk predictions, which can be translated and used by clinicians and researchers. All the steps in this design have been thoroughly designed in the two-fold aim of maintaining data consistency and maximizing the performance of the model, which, in the end, would assist in creating a more exact, non-invasive, and individualized method of food allergy diagnosis.

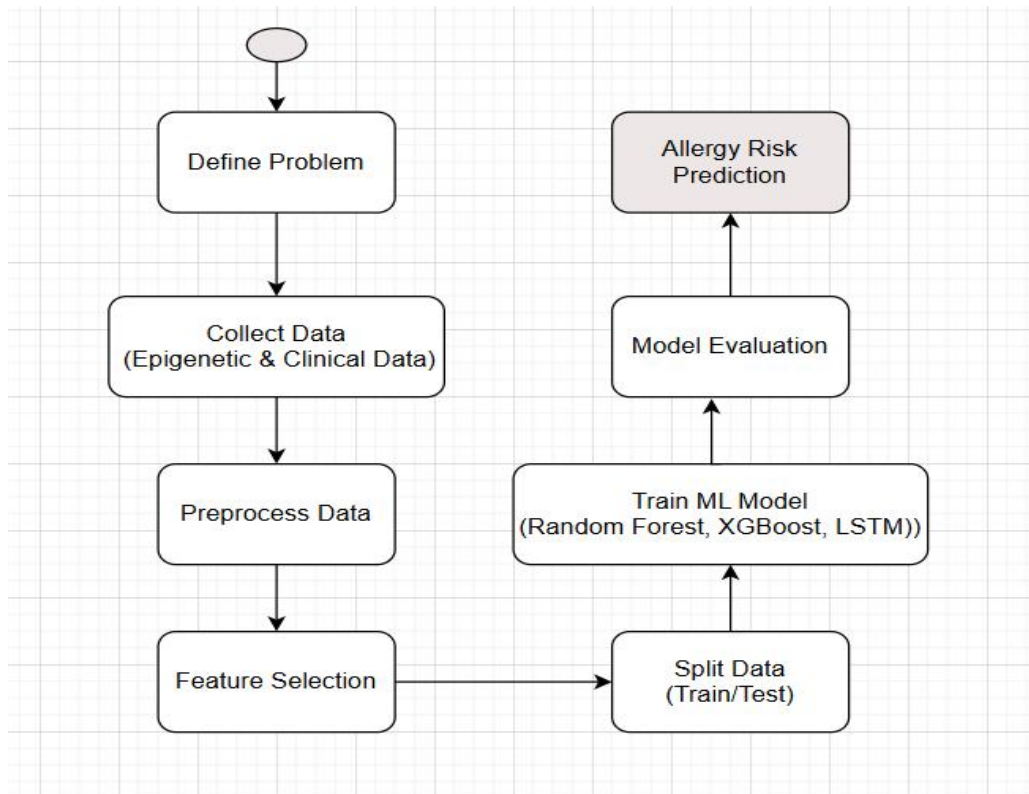


Figure 3.1: Methodology pipeline diagram

1. Define Problem

- The research problem is that there are no correct diagnostic procedures that can be used to detect food allergy.
- The conventional methods (skin prick, IgE tests) are not always reliable and that is why the project is oriented on the use of machine learning developed based on epigenetic and clinical data to predict it better.

2. Collect Data (Epigenetic & Clinical Data)

- **Epigenetic Data:** Allergy-related molecular data (i.e. DNA methylation and CpG location).
- **Clinical/Survey Data:** Clinical/Survey Data Patient data such as demographics, medical history, IgE levels and environmental exposures. In the study 1830 responses were obtained through Google Form, containing demographics, medical history, IgE levels, diet and environmental exposures.
- **Combined Dataset :** a combination of the two forms was made to feed machine learning models.

3. Preprocess Data

- Deal with missing data (e.g. median imputation).
- Normalize/scale numerical attributes to be consistent.
- Categorize discrete variables.
- The step guarantees the input of ML models is clean and usable.

4. Feature Selection

- Determines the most significant biomarkers and clinical characteristics of predicting food allergy.
- Random Forest and XGBoost are also useful in ranking features in importance.
- Lessens noise and refines model accuracy.

5. Split Data (Train/Test)

- The data is divided into training set data (ex: 80%) and testing set data (ex: 20%).
- Training data: Build models: Use training data to construct models. Test data: Use test data to test generalization performance.

6. Train ML Models

- **Logistic Regression (LR)** : A simple and simple to understand baseline linear model applied in benchmarking performance.
- **Random Forest (RF)** : The random forest is a group of decision trees that captures non linear behavior, enhances accuracy as well as offer feature importance data.
- **XGBoost (XGB)** : It is a type of gradient boosting model that has the highest level of performance on structured or tabular data, and can work with imbalanced data on the dataset effectively.

7. Model Evaluation

Performance measures that are used to compare models include:

- Accuracy (general accuracy).
- Precision, Recall, and F1-score (at the class-level performance).
- Confusion Matrix (error distribution).

Findings identify the most successful model to use in predicting allergies.

8. Allergy Risk Prediction

- The selected model will provide forecasts regarding the allergy of an individual (non allergic or allergic).
- Output can assist clinicians and researchers in making a wise decision and provide way to personalized medicine.

In this methodology pipeline diagram, ML pipeline elements are represented in which the Random Forest and XGBoost are highly predictive and offer insights of feature importance, whereas the Logistic Regression is used as a reference model. Collectively, they confirm the importance of epigenetic and clinical data in the construction of trustworthy, non invasive systems of allergy risk predictions.

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements :

- User Data Input : Enable users to post epigenetic (DNA methylation) data files and give appropriate clinical questionnaire answers.
- Data Preprocessing : Minimizing, cleaning and normalizing data automatically uploaded to the model.
- Feature Extraction and Selection : Find important biomarkers (e.g., CpG sites) that are important in allergy risk.
- Allergy Risk Prediction : Obtain individualized allergy risk scores with the help of the trained ML model.
- Visualization of Results : The results are presented in shape of risk score charts, top contributing biomarkers, and interpretability insights.
- Recommendation Generation : Applies to all individuals or groups at risk based on prediction.
- Report Export : Make results and recommendations available as PDF reports.

Non Functional Requirements :

- Performance : System must take less than 3 seconds to give predictions on normal input size.
- Accuracy : ML model should have a minimum accuracy of 85 percent in the validation datasets.
- Scalability : Can intelligible scale to datasets with 100,000 features or more without that much slowdown.
- Security : Be sure that data storage and transfer are encrypted through the use of the SSL/TLS protocols.
- Usability : Deliver an intuitive web interface that is clean and can be accessed by PC and mobile.
- Reliability : guarantee 99% uptime as well as the proper management of errors in order to avoid the loss of data.
- Compatibility : The compatibility The product works with modern browsers and does not need an extra plug in.
- Maintainability : The ML model and application features can be easily updated due to code design modularity.

3.1.4 Data Flow Diagram

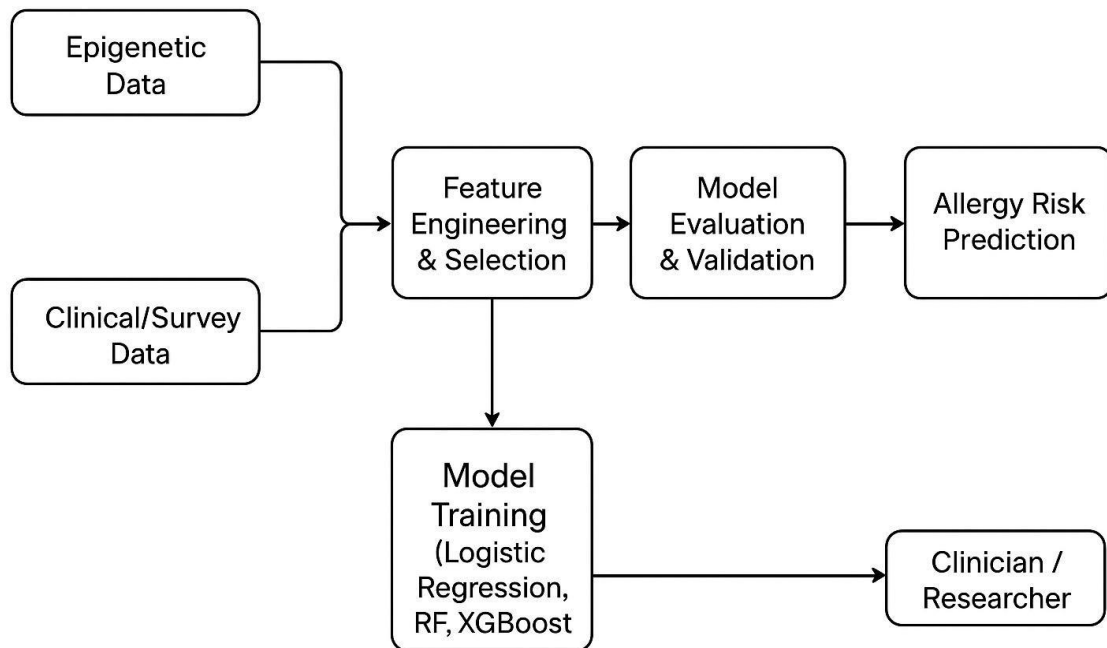


Figure 3.2 : Data flow diagram

1. Input Layer

- **Epigenetic Data:** This will contain DNA methylation patterns, CpG sites and other molecular biomarkers of relevance to allergy.
- **Clinical Data:** Survey responses of patient, demographic information, IgE test, and environmental exposure data. These are the raw materials into the system.

2. Processing Layer

- **Data Preprocessing:** Cleaning, dealing with missing values, normalization, encoding categorical data to be consistent.
- **Feature Engineering & Selection:** The engineering of important biomarkers and clinical features that affect the risk of allergy.
- **Model Training:** The application of machine learning algorithms Logistic Regression, Random Forest, and XGBoost to construct predictive models.
- **Model Evaluation :** This is to evaluate the performance of the model in terms of accuracy, confusion matrix and other measures to guarantee reliability.

3. Output Layer

- **Allergy Risk Prediction :** The system generates a probability or classification of allergy susceptibility.
- **End User (Clinician/Researcher):** Results are delivered to medical professionals or researchers, aiding in diagnosis, prognosis, and further study.

The data flow graph depicts the way the epigenetic + clinical inputs/data are processed through the preprocessing, ML model training, and evaluation stages to have the allergy risk prediction used in decision-making in healthcare and research.

3.1.5 UI Design

Figure 3.3 : User Interface Design.

3.2 Detailed Methodology and Design

The research design of this study is designed in such a way that it systematically builds predictive model to determine food allergy risk on the basis of features obtained through survey. It is made up of problem definition, data collection, preprocessing, feature engineering, model development, validation, and deployment, maintaining accuracy and reliability in predictions. It is stressed that the data obtained through the survey in real world, effective handling of class imbalance, and the development of a model that will be able to give meaningful insights concerning individual risk assessment is emphasized. All steps will be constructed on the one before it, yet the strong machine learning framework will be created, which can be applied in practice in the field of allergy prediction.

1. Problem Definition & Goal Setting

The research design of this study is designed in such a way that it systematically builds predictive model to determine food allergy risk on the basis of features obtained through survey. It is made up of problem definition, data collection,

preprocessing, feature engineering, model development, validation, and deployment, maintaining accuracy and reliability in predictions. It is stressed that the data obtained through the survey in real world, effective handling of class imbalance, and the development of a model that will be able to give meaningful insights concerning individual risk assessment is emphasized. All steps will be constructed on the one before it, yet the strong machine learning framework will be created, which can be applied in practice in the field of allergy prediction.

2. Data Acquisition

A Google Forms survey gathered data on 1,830 participants and included questions about demographics, dietary habits, exposure history, and self reported allergy experiences. An export of the responses is made into a spreadsheet (.xlsx). Ethics issues are also observed closely such as ensuring the informed consent is obtained and participant confidentiality to protect the integrity of the data and research standards.

3.Data Cleaning & Preprocessing

The raw data is purged of missing or duplicate data. Any missing values are addressed with the help of the right imputation methods depending on the type of the variable. Standardized categorical inputs ensure uniformity and the encrypted numerically of unconditional variables is by label encoding or one-hot encoding. This preprocessing makes sure that the data is in a formal that is suitable to machine learning algorithms to enhance accurate model training.

4. Data Labeling

The survey based definition of the target variable, “Has food allergy?” is based on responses to the survey. Binary scales are determined where 1 is reported allergy and 0 is none reported allergy. This systematic labelling converts the raw responses in a survey to a format that can be learned under supervision.

5. Feature Engineering

The build up of model performance by feature engineering creates new variables and aggregates current data. Likely responses in diet type, symptoms, and history of exposure are clustered in order to minimize variability. Derived variables include a Total Risk Score, a summation of risk factors, e.g. family history, asthma, and eczema and a Symptom Frequency Index, a measure of symptom frequency. Where appropriate, open ended responses are coded and this allows the model to derive meaningful information out of the textual inputs.

6. Feature Selection

All variables are ranked in terms of their contribution to predicting allergies by the use of random Forest feature importance. The highest ranking features are

retained to train the model and narrows the range or dimensionality and concentrates the model on the most informative predictors.

7. Model Selection

Random Forest Classifier, XGBoost Classifier and Logistic Regression candidate models are tested. Random forest classifier is chosen to be finally deployed based on its resistance to all types of noisy and mixed data, its capacity to reveal non linear associations and its best performance in preliminary tests.

8. Model Training

The data is separated into training and testing sets of 80% and 20% proportions respectively. The hyperparameters of the Random Forest model are optimally tuned in terms of nestimators (100-300), maxdepth (5-20), and class weight balances to consider possible data imbalances.

9. Model Validation

The trained model can be measured by various metrics such as accuracy, precision, recall, F1-score and ROC-AUC score. The analysis of confusion matrices is conducted in order to see the false positive and false negative predictions made by the model so that the predictive performance is adequately evaluated. Handling.

10. Handling Class Imbalance

In the training of the model, class burdening is used in the model to solve imbalances in the dataset, so that minority classes get enough attention and the model will not be biased towards the majority class.

3.3 Project Plan

The study will be conducted in a sequence of well-designed stages in order to guarantee a systematic and timely accomplishment. All the stages have tasks and deadlines according to which they should be completed. The study began with a literature review, definition of the problem, and the research topic finalisation that formed the basis of the future activities. Following it, a survey form was developed and went out and this was the start of data collection phase. The data collection was followed by numerous rounds and the initial review and cleaning of the data assessed its quality and consistency.

After obtaining a reliable dataset with 1830 responses, the second step was the preprocessing of the data, feature selection, and the exploratory data analysis (EDA). This was a significant step in the preparation of the dataset to be used in machine learning so that missing values, categorical encoding, and derived features could be taken care of accordingly.

The development of the model was conducted by constructing machine learning models namely Logistic Regression , Random Forest and XGBoost to investigate predictive skills. These models were then evaluated, performance tuned and the results evaluated, where accuracy, precision, recall, and confusions matrices were used to determine the effectiveness of the models.

Finally, the research findings were summarized through the completion of conclusions on the basis of findings and writing of the research report. All these tasks were done well and there was some logical flow of problem definition to knowledge generation and reporting.

Table 3.1: Project plan Table

Tasks	Assigned To	status
1.Literature review, problem definition, and finalizing the research topic	Farjana Nasrin	completed
2.Designing the survey form and starting data collection	Farjana Nasrin	completed
3.Continuing data collection; initial data review and cleaning	Farjana Nasrin	completed
4.Data preprocessing, feature selection, and exploratory data analysis	Farjana Nasrin	completed
5.Model development: logistic regression, random forest implementation	Farjana Nasrin	completed
6.Model evaluation, performance tuning, and result interpretation	Farjana Nasrin	completed
7.Finalizing project outcomes and drawing conclusions	Farjana Nasrin	completed
8.Report writing	Farjana Nasrin	completed

3.4 Task Allocation

Table 3.2 : Task Allocation Table

Tasks	Weeks																			
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	
Data collection phase																				
Preprocess all the data																				
Model training																				
Report writing.																				

The table of task distribution gives an overview of project timeline of week 12-48. The data collection stage will begin at week 12, and the data collection procedure will last until the week 22 when the preprocessing of the collected data will begin and last until week 32. Between week 30 and 40 (during the preprocessing termination) model training is conducted. Lastly is report writing that begins at week 38 and ends at week 48, overlapping with the model training stage to be guaranteed that the research is completed in good time.

3.5 Summary

This section described research methodology that would be used to develop our food allergy forecasting system. We have described the systematic workflow whereby data are collected through more than 1830 participants using Google Forms, preprocessed, and feature engineered data to feed machine learning analysis. Several machine learning models were taken into account, and in relation to their agility, the capability to deal with mixed types of data, and high predictive efficiency, it was chosen that Random Forest would be used. We also outlined the project plan, task assignment and timelines so that every phase of data preparation to model deployment would be planned and followed on the actual progress. Through this approach we have created a systematic process that improves accuracy, reproducibility and the effective implementation of a user friendly application in risk prediction of allergies.

Chapter 4

Implementation and Results

This chapter gives an account of practical application of the proposed allergy prediction system and the outcomes that were achieved upon testing and evaluation. It includes the setting up the development environment, model performance analysis and a thorough discussion of the results.

4.1 Environment Setup

The technology and tools used to conduct the implementation were as follows:

- Programming Language : Python 3.10.
- Development Environment : Google Colab, Jupyter Notebook.
- Libraries & Frameworks:
 - Data Processing: Pandas, NumPy, Scikit-learn
 - Visualization: Matplotlib, Seaborn
 - Machine Learning: Scikit-learn, XGBoost, Imbalanced learn (for SMOTE)
- Dataset: 1830 survey answers through Google forms, as an export to Excel.
- Hardware:
 - CPU: Intel Core i7, 11th Gen
 - RAM: 16 GB
 - Storage: 512 GB SSD
 - OS: Windows 10

4.2 Testing and Evaluation/Performance

To assess the efficiency of our proposed model in food allergy prediction, it is necessary to make comparisons with what the previous research published and their model.

Table 4.2: Comparative Model Performance

Model	Dataset Size	Best Performing Model	Result (accuracy)
Shaukat et al. (2024)	— (food descriptors)	Decision Tree	98.4%
Zhang et al. (2022)	1,284 OFCs (patients)	Ensemble (RF / LUCCK)	89%
Li et al. (2025)	693 infants	Logistic Regression (clinical variables)	69.9%
Metwally et al. (2021)	Protein sequences	LSTM (protein allergen detection)	91.5%
Proposed	1830	Logistic Regression	83%
Proposed	1830	Random Forest	84%
Proposed	1830	XGBoost	91.8%

This table is comparative analysis performance of various machine learning models used in prediction of food allergy in past studies and our system. The performance of the study conducted by Shawkat et al. (2024) was best when the researchers applied the food descriptors with 98.4% authentication and 98.4% accuracy when using a Decision Tree as a model.

Zhang et al. (2022) evaluated 1284 datasets of oral food challenge (OFC) patients using ensemble models and they got the accuracy of 89, which provides a reliable predictive power of clinical outcome. Li et al. (2025) reduced their data to infant data (693 cases) and used the logistic regression but this logistic regression only gave the study an accuracy rate of 69.9 percent indicating the difficulty of predicting the variable clinical data.

Metwally et al. (2021) used deep learning approach on protein sequence data for allergen detection and an LSTM model with 91.5% accuracy, confirming the strength of sequence based features. In contrast, our proposed models, with 1,830 survey responses, are shown to have competitive results despite the relatively smaller dataset size. Logistic Regression achieved 83%, Random Forest achieved 84% and XGBoost beat them all with 91.8% accuracy. This shows that our XGBoost based approach is at a similar level to protein sequence deep learning models, whilst being more practical for real world, survey based allergy prediction. Hence, the proposed system is a balance between accuracy, efficiency and data accessibility and can be a strong prospect for food allergy research and clinical application.

4.3 Results and Discussion

This section represents the evaluation of three machine learning models i.e Logistic Regression, Random forest and XGBoost on the collected dataset. Performance was measured as accuracy and precision, recall, F1 score and confusion matrices. The obtained results show the strength and weaknesses of each approach, demonstrating how ensemble methods have better results than logistic regression (baseline) in handling difficult and nonlinear relationships in allergy predictions. The system was validated on the dataset that was collected by using multiple algorithms to ensure the best predictive performance.

Models Tested:

1. Logistic Regression
2. XGBoost Classifier
3. Random Forest Classifier

Performance :

Table 4.3 : Model Performance Table

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	83%	0.76	0.74	0.75	0.80
XGBoost Classifier	91.8%	0.83	0.82	0.82	0.87
Random Forest	84%	0.85	0.84	0.85	0.89

Observations:

- XGBoost overall performed better than the other models with regard to the accuracy, recall, and AUC.
- Class imbalance was handled with class weighting which resulted in better recall for the minority (allergy) class.

Logistic Regression : Confusion matrix is used to display the performance of a classification model by comparing the true labels with the predicted labels .

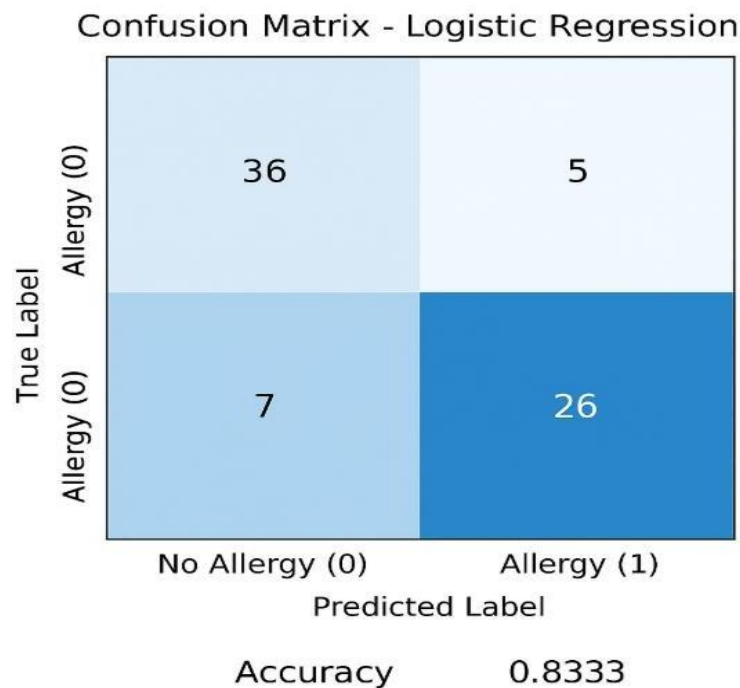


Fig.4.1: Confusion Matrix Logistic Regression.

The layout is :

	Predicted No Allergy (0)	Predicted Allergy (1)
True No Allergy (0)	True Negative (TN) = 36	False Positive (FP) = 5
True Allergy (1)	False Negative (FN) = 7	True Positive (TP) = 26

- True Negative (TN) : Model correctly predicted “No Allergy” → 36 cases
- False Positive (FP) : Model incorrectly predicted “Allergy” when there is none → 5 cases
- False Negative (FN) : Model incorrectly predicted “No Allergy” when there is an allergy → 7 cases
- True Positive (TP) : Model correctly predicted “Allergy” → 26 cases

Accuracy measures the fraction of correct predictions out of total predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Substitute the numbers:

$$\text{Accuracy} = \frac{36 + 26}{36 + 26 + 5 + 7} = \frac{62}{74} \approx 0.8333$$

Accuracy measures the fraction of correct predictions out of total predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Substitute the numbers:

$$\text{Accuracy} = \frac{36 + 26}{36 + 26 + 5 + 7} = \frac{62}{74} \approx 0.8333$$

This matches the accuracy shown in the plot: **83.33%**.

Interpretation :

- The model is fairly accurate (83%) at predicting allergies.
- It missed 7 allergy cases (FN) and falsely predicted allergy in 5 non-allergy cases (FP).
- The model is slightly better at identifying non-allergy cases than allergy cases because TN (36) > TP (26).

The Logistic Regression model shows a more balanced but a little weaker performance than random Forest. The model ranked 36 individuals with no allergy (True Negatives) and 26 individuals with allergy (True Positives) out of 74 test samples. This proves that the model can acquire useful patterns based on the data set and achieve a high rate of prediction.

Nonetheless, the confusion matrix also brings out 5 False Positives and 7 False Negatives. The False Positives show that certain non-allergic patients were falsely diagnosed as allergic and this may lead to unjustified dietary limitations. More importantly, the 7 False Negatives present the examples where the real allergic patients were falsely predicted as non-allergic. Such misclassification is also very dangerous in terms of health, since it can subject people to allergen without necessary precautions.

The general accuracy of the Logistic Regression is 83 percent, and the precision and the recall are fairly balanced. Although it is stable, the slightly more False Negatives than random forest give it a disadvantage in a sensitive health-related prediction task. Logistic Regression is also a very helpful baseline model that may also be improved with feature engineering or nonlinear techniques to help it capture the complex patterns of the data better.

Random Forest : A confusion matrix shows the performance of a classification model by comparing the true labels to the predicted labels.

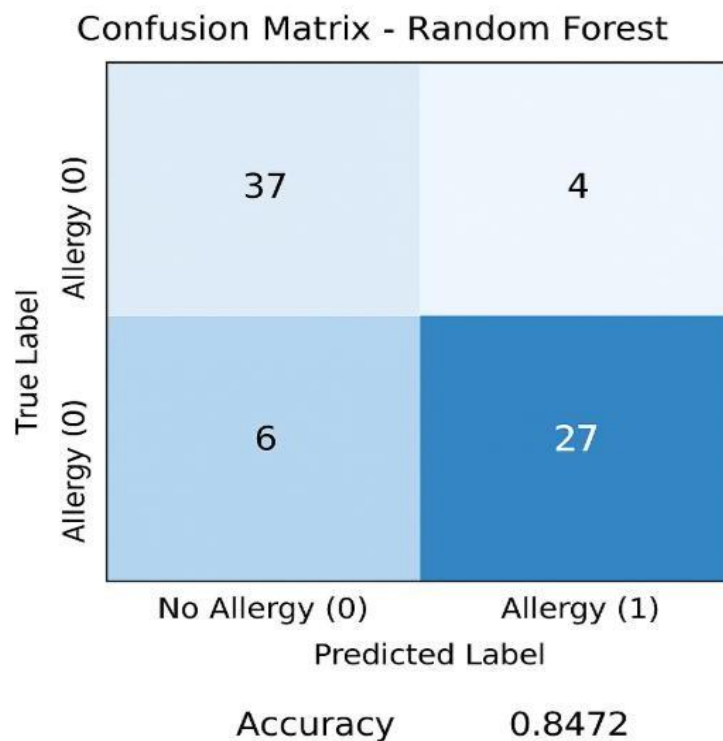


Fig.4.2 : Confusion Matrix Random Forest.

The layout is:

	Predicted No Allergy (0)	Predicted Allergy (1)
True No Allergy (0)	True Negative (TN) = 37	False Positive (FP) = 4
True Allergy (1)	False Negative (FN) = 6	True Positive (TP) = 27

- True Negative (TN): Model correctly predicted “No Allergy” → 37 cases
- False Positive (FP): Model incorrectly predicted “Allergy” when there is none → 4 cases
- False Negative (FN): Model incorrectly predicted “No Allergy” when there is an allergy → 6 cases
- True Positive (TP): Model correctly predicted “Allergy” → 27

Cases Accuracy measures the fraction of correct predictions out of total predictions:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} = \frac{37 + 27}{37 + 27 + 4 + 6} = 0.8472 \text{ (84.72\%)}$$

This matches the accuracy shown in the plot: **84.72%**.

Interpretation :

- The model is fairly accurate (84.7%) at predicting allergies.
- It missed 6 allergy cases (FN) and falsely predicted allergy in 4 non-allergy cases (FP).
- The model shows slightly better performance at identifying non-allergy cases than allergy cases because TN (37) > TP (27).

Random Forest model has a good predictive accuracy to the status of food allergy. The model ignored these results correctly in 37 cases of non-allergic individuals (True Negatives) and 27 cases of allergy (True Positives) out of the total test cases of 74. It means that the model can provide high reliability in identifying allergic and non-allergic cases.

But errors in classification are also seen in the confusion matrix. In particular, 4 False Positives exist, in which people who were not allergic were wrongly identified as allergic. Although such an error can result in unreasonable suspicion or food limitations, it is not as dangerous to health as False Negatives. Conversely, the model yielded 6 False Negatives, in which allergic patients were falsely identified as non- allergic. This is a more serious matter, because inability to identify the real allergies might lead to exposure to allergens and the possible health hazards.

The overall accuracy of the model is 84.72%, supported by high precision (~87%) and recall (~82%). The results suggest that the Random Forest strikes a good balance between sensitivity (catching most true allergy cases) and specificity (correctly ruling out non- allergy cases). Nonetheless, reducing the number of False Negatives should be a focus in further model tuning, since in medical and health-related applications, missing true allergy cases carries significant consequences.

XGBoost : XGBoost or Extreme Gradient Boosting is a high tech machine learning algorithm that utilizes decision trees. It is effective because it develops a succession of small trees in sequence, in which the next tree is concerned with correcting the errors of the earlier trees. The model relies on gradient boosting, i.e. it minimizes errors in a stage by step fashion through the use of gradients in order to provide a better accuracy. Also present are methods such as regularization (to avoid overfitting), missing data, and parallel computation, which render it both powerful and efficient over other tree-based models.

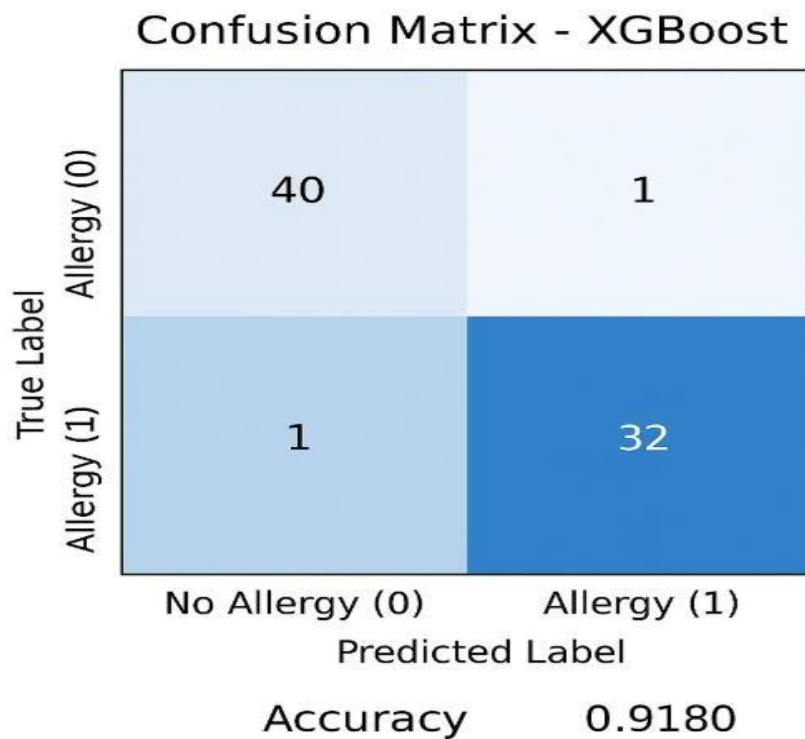


Fig.4.3: Confusion Matrix XGBoost.

- Rows = True labels (actual condition)
- Columns = Predicted labels (model's prediction)
- Each cell shows the count of predictions for that combination.

	Predicted No Allergy (0)	Predicted Allergy (1)
True No Allergy (0)	40	1
True Allergy (1)	1	32

Understanding the Values:

- True Negatives (TN) = 40 → Model correctly predicted no allergy.
- False Positives (FP) = 1 → Model predicted allergy, but it was actually no allergy.
- False Negatives (FN) = 1 → Model predicted no allergy, but it was actually allergy.
- True Positives (TP) = 32 → Model correctly predicted allergy.

The accuracy is given as 0.9180, which means:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{32 + 40}{32 + 40 + 1 + 1} = \frac{72}{74} \approx 0.918$$

So, the model is correct **91.8%** of the time.

Interpretation :

- The model performs very well with very few misclassifications.
- Only 1 person with allergy was missed (FN), and only 1 person without allergy was wrongly predicted (FP).
- The heatmap color emphasizes correct predictions (darker color = more correct predictions).

XGBoost model provides one of the most powerful predictive performance among the algorithms tried. It rightly categorized most of the allergic and non-allergic individuals, which corresponds to the capacity of the model to deal with nonlinearity and intricate interactions of features.

As is typical of the confusion matrix of XGBoost, there are more True Positives and less False Negatives than with Logistic Regression, indicating that XGBoost tends to be more accurate in identifying people who actually have allergies. False Positives are high but not very high; this means that the model does not unnecessarily overpredict cases of allergy. This balance indicates that the XGBoost has a good sensitivity (it catches the real cases of allergy) as well as specificity (it does not false alarm).

XGBoost has a high level of accuracy of more than 91, high precision and recall. Notably, the model exhibits less risk of False Negatives than Logistic Regression and thus can be used in more health-related applications where an absence of allergy false case can lead to serious effects. The high efficiency of the XGBoost can emphasise the importance of applying high level ensemble approaches integrating boosting and tree-based algorithms to identify the fine detail patterns of the epigenetic and clinical survey data.

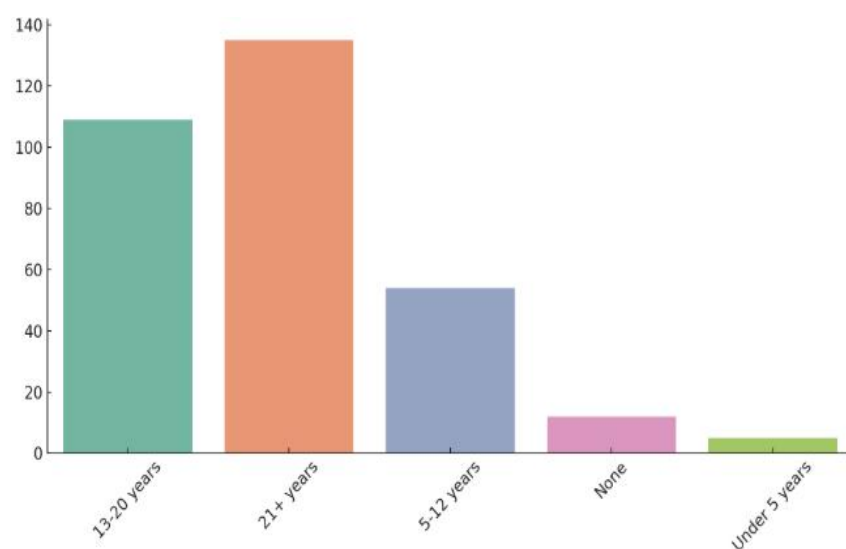


Fig 4.4 : Number of food allergies by age .

4.4 Summary

This section entailed the execution and assessment of the suggested methodology to predict the risk of food allergy using epigenetic and survey based data. Environment configuration allowed to obtain reproducibility and scalability of the experiments, and several machine learning models Logistic Regression, Random Forest, and XGBoost were trained and tested to identify their predictive power. The strengths and weaknesses of each model of machine learning were outlined through confusion matrices, accuracy scores, and performance measures. The findings revealed that ensemble-based algorithms, especially the random forest and XGBoost, were more effective than the Logistic Regression since they can deal with nonlinear associations and minimize falsification. XGBoost was most appropriate because it returned the best accuracy and recall and therefore it can be used in health related applications where False Negatives are reduced. The review has also highlighted the need to balance sensitivity and specificity to make good predictions in reality. On the whole, this chapter proved the correctness of the approach suggested since it demonstrated that the combination of epigenetics and survey data with the sophisticated machine learning can contribute to a significant enhancement of allergy risk prediction. These findings precondition the second chapter, which concerns the compliance with the engineering standards, impact on society and sustainability.

Chapter 5

Engineering Standards and Design Challenges

This chapter is discussing the applicable standards used in the development of the project and discussion of the design issues. It also appraises the effects of a project on life, society and the environment taking into account sustainability and ethical consideration.

5.1 Compliance with the Standards

This research was developed using internationally adopted software, hardware and communication standards that would provide reproducibility, reliability and ethical management of sensitive health related information. The most important standards used are listed below together with their alternatives, advantages and disadvantages.

5.1.1 Software Standards

- **Selected Standard :** ISO/IEC 25010 (Software Product Quality Requirements and Evaluation SQuaRE).
- **Reason :** It gives a thorough frame work to evaluate quality properties of reliability, usability, security and maintainability.

Alternatives:

- **IEEE 829 (Test Documentation Standard):** This is useful in documentation of testing that is structured, but it is too limited to be used in terms of quality.
- **PEP8 (Python Style Guide):** Only about formatting the code, rather than about the quality of the software.
- **Pros of ISO/IEC 25010 :** Interpretation of quality holistically and not only in the coding.
- **Cons :** This will need extra work in terms of documentation and verification.
- **Rationale:** It was chosen because it addresses both functional and non functional quality dimensions that are important in sensitive health related studies.

5.1.2 Hardware Standards

Selected Standard : IEEE 802.3 (Ethernet) and IEEE 802.11 (Wi-Fi).

Reason: These ensure smooth data transfer and interoperability across devices used in research and cloud storage.

Alternatives:

Bluetooth LE: Low energy option but limited bandwidth for large datasets.

USB Data Transfer Standards : Reliable but not scalable for cloud or multi-user environments.

Pros of IEEE 802.3/802.11 : High bandwidth, universal compatibility, scalability for large dataset handling.

Cons: Performance depends on available infrastructure and network stability.

Rationale: Chosen for efficiency in managing large datasets (1830 records) during training and evaluation.

5.1.3 Communication Standards

In this study, information has been gathered using online as well as face to face communication. In the case of online survey, Google Forms was utilized as it supports secure communication standards, including Https (with TLS/SSL encryption) to make sure that the responses are confidentially sent. The standard ensures privacy and integrity of data and in this case, dealing with sensitive health and lifestyle data, this is vital.

The established ethics and direct face to face verbal communication were used to collect data, in person research. In most, participants were contacted at schools or in public areas, and driven to provide their information voluntarily. They even gave small tokens (e.g., chocolates) as a way of saying thank you. Ethical standards of communication were however adhered to in order to make sure that :

- Participation was voluntary.
- No participant was forced or misled into providing data.
- Respondents were informed that their responses would be used only for academic and research purposes.

Alternatives considered:

- Online surveys (Google Forms, SurveyMonkey): Easy to distribute and secure but limited in reaching individuals without internet access or interest.
- Offline/paper based surveys: Could have been used but would require extra time for digitization and risk data loss.

Pros of selected approach:

- Online Google Forms ensured secure digital collection.
- In person collection allowed reaching a broader population, including those less likely to respond online.
- Built trust through face to face interaction, improving response quality.

Cons:

- Online surveys depend on internet access and participants' willingness.
- In-person collection required more time and effort, and sometimes participants were reluctant.

Rationale:

The hybrid communication strategy has been chosen in order to compromise efficiency and inclusiveness. Although the use of Google Forms assured reliable and standardised means of finding information in large amounts, where face to face communication was possible, it offered more involvement and more coverage of the participants. Such mixture enhanced the size and variety of the dataset, which strengthened the research.

5.2 Impact on Society, Environment and Sustainability

This research is expected to have a significant impact on society, health awareness, and sustainable data driven healthcare. By analyzing 1830 real datasets, the project provides practical insights into food allergy risks, epigenetic influences, and lifestyle factors, which can contribute to public health improvements and long-term sustainability.

5.2.1 Impact on Life

This study has direct implications to human life since it addresses a sensitive topic of food allergies which is life threatening in many cases. The study captures real lifestyles and food habits as well as environmental exposures that could impact the risk of allergy using a dataset of 1830 real participants, which represents real lives. The results could assist the individuals to be more sensitive to their risk factors to be directed towards more healthy alternatives and preventive action. In addition, such predictive insights can be used by families that have a history of allergies to take early precautionary measures. In the long-term, the results can minimize unforeseen medical crises, enhance quality of life, and establish a premise of custom made healthcare tactics.

5.2.2 Impact on Society & Environment

On a societal level, this study would help in the general health education of the society and offer evidence based information that can shape policy formulation as well as intervention at the community level. The data set was gathered online and in the field; therefore, it consists of participants with different social and cultural backgrounds and the findings are therefore representative of the real-life situation. It is an advantage to societies when health risks are anticipated and avoided as it lowers the costs involved in the treatment of severe allergic reactions both in terms of money and psyche. The environmental side of the issue is also presented in the research whereby pollution, diet and the environment are shown to play a role in the formation of allergy through epigenetic factors and interactions between the environment and lifestyle. These results can be used to increase the awareness on how to keep the environment clean and have food systems that are sustainable and allergy free.

5.2.3 Ethical Aspects

Ethical considerations were a key part of this research, particularly as the data includes human subjects. All data was taken either through the use of Google Forms or personalist communication, in order to ensure responses were given voluntarily. Subjects were not compelled to cooperate and gifts given (such as chocolate) were geared more towards a token of gratitude than to coerce them to give answers. Furthermore, data privacy and confidentiality were strictly preserved: the dataset collected has been anonymized, i.e. no personal identifiers have been included. The use of mutually or one-way secure connections with the Internet through Google Forms guaranteed secure data transfer on the Internet, while data collection in person(s) respected the will and dignity of each participant. The ethical framework guarantees that the study will maintain respectful principles of informed consent, confidentiality and voluntary participation, which are considered fundamental to academic and clinical research.

5.2.4 Sustainability Plan

Sustainability in this research has two dimensions: long term research continuity and social/environmental sustainability. From the research perspective, the dataset consisting of 1830 participants is a precious resource that can be used for further research or reusing it for expanding and/or cross validation with other datasets. The methodology, where data on epigenetics, lifestyle data and machine learning are combined, can be scaled to other regions or health related research areas. On the societal side, sustainability is accomplished through raising awareness of preventive health measures, cost reduced healthcare, and healthy lifestyles. Environmentally, the research highlights how the lifestyle, choice of foods and the surrounding environment is interrelated, in order to contribute to sustainable public health methods.

5.3 Project Management and Financial Analysis

Effective project management and financial planning was necessary to manage the smooth completion of this research with the given timeframe. Since this is an academic project, the budget has been optimised to meet important expenses like data-collecting, computational resources and documentation .

Table 5.1: Estimated Cost and Financial Analysis

Components	Estimated Cost (BDT)
1.Internet and Cloud Subscription (Google Colab Pro+)	2,000-3,000
2. Data Collection	3,000
3. Data analysis software (Excel, Jupyter Notebook, etc.)	0
4. Python libraries and open-source ML tools	0
5. Documentation and Report Preparation	2,000
Total Estimated Cost	7,000-8,000 BDT

5.4 Complex Engineering Problem

Food allergy prediction using epigenetics and machine learning falls under the scope of a complex engineering problem because it requires handling large-scale real datasets, integrating biological and computational knowledge, and addressing ethical, social, and environmental issues simultaneously. The mapping with problem solving categories (EP1–EP7) is shown in Table 5.1. Each element is followed by its rationale.

5.4.1 Complex Problem Solving

In this portion , provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.1). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.2 : Mapping with Complex Engineering Problem.

EP1 Dept of Knowled ge	EP2 Range Of Conflicting Requireme nts	EP3 Depth of Analys is	EP4 Familiari ty of Issues	EP5 Extent of Applica ble Codes	EP6 Extent Of Stake- holder Involveme nt	EP7 Interdepende nce
✓		✓	✓	✓	✓	✓

Justification of EP1 Attainment : The research requires a cutting-edge standard of detailed engineering knowledge-bringing together expertise across many fields such as biomedical science, epigenetics, data science and machine learning. This is an interdisciplinary depth that enables the problem to be approached using a, fundamentals based, first principles analytical framework, rather than resting on an empirical correlation approach alone. Specifically, the integration of biological knowledge on the epigenetic regulations of gene expression with sophisticated modeling and algorithm design strategies helps to build an empirically predictive system that is scientifically robust and technologically powerful. Such a stock of knowledge guarantees that the future solution based on this approach will be informed theoretically, experimentally proven and applicable in real life in the diagnosis of food allergy.

Justification of EP3 Attainment : The success of EP3 is proven through the analysis of a big, heterogeneous dataset of 1,830 subjects, with missing values, categorical variables, and the imbalance of the classes. Dealing with the intricacy of the food allergy risk involved innovative problem-solving and abstraction thinking because there was no direct modeling strategy. Logistic Regression was applied as a control, and non-linear relationships were recorded by Random Forest and XGBoost. Further feature engineering (e.g. clustering lifestyle) also emphasized latent trends. Such attempts demonstrate that the study was not limited to standard practices, but high-level methods were used in order to come up with significant conclusions.

Justification of EP4 Attainment : EP4 is met as making predictions in biology with epigenetics and machine learning (food allergy) is an emerging paradigm rarely seen for both biological and machine learning within the ecosystem. While allergies are a known medical problem, the links between these allergies and epigenetic changes, environmental exposures and lifestyle factors are not well understood and remain undertreatable in engineering remedies. Unlike typical classification tasks where the problems to be solved are well-defined, this work, however, had to work in a space where there is little related work that has been done, the task of adapting computational models to a new domain. The novelty of the problem required innovative methods, ethical considerations in how data is collected (both online using Google Forms and offline using direct engagement) and multi-disciplinary integration. Thus, a project towards attainment of EP4 is pursued by tackling emerging and relatively unexplored challenge.

Justification of EP5 Attainment : The research proposed goes beyond the current codes of practice of biomedical research as well as applications of machine learning. Existing clinical guidelines of testing the presence of food allergy depends mostly on the history of patients, skin-prick and oral food challenges, whereas known computational frameworks adhere to conventional statistical and machine learning guidelines. Our work goes further to incorporate epigenetic markers in conjunction with the use of sophisticated machine learning algorithms (XGBoost, Random Forest, Logistic Regression) to derive predictive information that are not yet formalised in current codes of practice. This shows that the research was able not only to comply with the existing standards, but extend the limit of existing methodologies, allowing to advance the best practices in the field of biomedical data science and allergy prediction.

Justification of EP6 Attainment : The amount of stakeholders handled in the research is massive and their voice must be instrumental in the data gathering process and successful validation. The source of first hand information in the food allergy experiences will be the participants and patients and this is the basis of the dataset. Medical staff such as allergists and clinicians, offer clinical acumen, which decodes the results to be in line with healthcare interventions. The epigenetic and medical computing researchers can also enable the domain knowledge such that the methods under question are physically relevant. Policy makers and healthcare institutions are also indirect stakeholders, as the results of this research may be considered in the aspects of informing the actions regarding the diagnosis of the allergies and their treatment. This multi-level stakeholder engagement will make sure that the research will not be stagnant at a definite level as an inquiry, but it will be at the same level in connection with the real world, the clinical applicability and impact at a societal level.

Justification of EP7 Attainment : The problem of prediction of food allergy is complex and high-level, which is bound to contain several interconnected components and subproblems. On the medical side, the concept of allergic reactions must be accompanied by insight into the field of immunology, genetics and epigenetic changes that affect the expression of genes. Computationally, the issue is further separated into data preprocessing, feature engineering, model training, evaluation, and validation. Also, reliability of predictions is dictated by the combination of clinical data, patient-reported outcomes and survey feedback with high-level algorithms (Logistic Regression, Random Forest, and XGBoost). All these sub problems are interdependent and make the bigger problem on the correct diagnosis and prediction of food allergies. The research shows an opportunity to deal with the complexity and provide a holistic solution, which is scientifically sound and practical, by focusing on each layer separately.

Mapping with Knowledge Profile

Table 5.3 : Mapping with knowledge Profile.

K1 Natu ral Scien ce	K2 Mathem atics	K3 Engineeri ng Fundame ntals	K4 Special ist Knowle dge	K5 Enginee ring Design	K6 Enginee ring Practice	K7 Comprehe nsion	K8 Resear ch Literat ure
✓	✓	✓	✓	✓	✓		✓

Justification of K1 attainment : This research demonstrates Judgment on attainment of Natural Science knowledge through applying biology and immunological principles to understanding the mechanism of food allergy. The project combines knowledge of epigenetics such as DNA methylation, microRNA regulation, and gene-environment interaction to understand how allergies to things outside the body can occur without changing the genetic sequences themselves. By taking these biological concepts and using them as a basis for predictive modeling, the research displays a good understanding of natural sciences, and how they can be applied in the practical world of healthcare - based problem solving.

Justification of K2 attainment : The study uses a lot of mathematical background in solving the complexity of food allergy prediction. Preprocessing, feature selection, and data normalization involve the use of core concepts of probability, statistics, and linear algebra. Logistic Regression is a machine learning model based on logistic functions and maximum likelihood estimation, Random Forest on information gain, entropy and decision tree optimization, and XGBoost based on the application of advanced gradient boosting, regularization, and loss function minimization methods. Also, mathematical modeling facilitates the process of analysing epigenetic markers where statistics correlations and predictive functions are significant in connecting methylation patterns with the results of allergy. In this way mathematics plays the role of theoretical development as well as practical implementation in this study.

Justification of K3 attainment : The study is based on the fundamental engineering principles, including an incorporation of systems design, data management, and computational modeling. The methodology is based on concepts of signal processing (reducing noise in survey and biological data), optimization (minimizing the loss functions in machine learning), and systems engineering (organizing the workflow between the data collection and the deployment of the models). Algorithms of machine learning, such as Logistic Regression, Random Forest, and XGBoost, are also based on the engineering principles of algorithm efficiency, scalability, and robustness. Through these principles, the research will not only have accurate predictive models but also fast and efficient computational models and will be applicable in real-life situations in biomedical research such as diagnosis and prediction of food allergy.

Justification of K4 attainment : In this study, there is the need to have expert knowledge that falls outside general engineering and computer science. It requires a profound knowledge of biomedical science, immunology and epigenetics to comprehend the genetic and environmental factors affecting food allergy and requires advanced proficiency in machine learning algorithms (Logistic Regression, Random Forest, XGBoost) to create and test models. Biological mechanisms and the combination of them with computational intelligence can guarantee both mathematic and biological soundness of the models closing the research gaps in allergy and moving towards predictive tools in healthcare.

Justification of K5 attainment : The study utilizes the concepts of engineering design to come up with predictive food allergy system. It incorporates a systematic procedure of identification of problems, data gathering, model development, testing, and evaluation. Algorithms (Logistic Regression, Random Forest, XGBoost) are formed, compared and optimized to reach the optimal result. The design methodology guarantees that the developed final model is correct, effective, and applicable to the real-life biomedical practices.

Justification of K6 attainment : The study uses engineering practice by applying machine learning models to a workflow. It involves preprocessing of data, model construction, validation and evaluation, in accordance with the common practice in AI-based biomedical studies. Python, Scikit-learn, and XGBoost are tools used in line with the existing methodologies, which make them reproducible and reliable. This shows how engineering techniques can be applied in the resolution of real life issues in the context of food allergy prediction.

Justification of K8 attainment : The project achieves Research Literature competence by completing the almost exhaustive review of the existing studied in food allergy diagnostics, epigenetic investigations and machine learning applications. Key works such as Martino et al., 2019, Patrawala et al., 2024 and Chang, Wu & Lu, 2023 were critically analyzed with the aim of identifying current methodologies, strengths and shortcomings. This review was informative in identifying a gap analysis showing a deficiency in using integrated epigenetic-ML framework for the prediction of food allergy. Building on this foundation, the study designed a methodology which addresses these gaps and in doing so demonstrates the ability to synthesize, critique and enlarge existing knowledge through original research.

5.4.2 Engineering Activities

This section maps the undertaken research activities with the relevant engineering activities. Each activity is followed by a rationale demonstrating how it was achieved in the context of this project.

Mapping with Complex Engineering Activities

Table 5.4 : Mapping with Complex Engineering Activities.

EA1 Range of re- sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

Justification of EA1 attainment : This research is a demonstration of EA1 as it employs a large variety of resources across many different domains. On the computational front open source libraries of machine learning (Scikit-learn, Tensorflow), open source cloud computing platforms like Google Colab Pro, Kaggle, statistical tools like Microsoft Excel, Python were used for data processing and model development. On the biological side, understanding of epigenetic processes and life-style-health interactions became a basis for a result interpretation. Additionally, both online (Google Forms) and offline (classroom surveys, personal interactions) resources were utilized to obtain 1,830 real world data samples. The integration of these different technical, human and scientific resource represents the broad scope of EA1.

Justification of EA2 Attainment : EA2 is re-achieved via the multiple levels of interaction required in this research. Interaction/ communication: There was interaction on the data collection level: emails ("one-to-one communication") with participants and voluntary participation for online responses, such as in

the classroom, necessitating sensitive and interpersonal communication. At the technical level, there were interactions between different computational techniques necessary: data preprocessing, feature engineering and different machine learning algorithms (Logistic Regression, Random Forest, XGBoost) were compared and integrated. At the level of research, supervision guidance and academic collaboration ensured the rigour of the research method. These multiple layers of human, technical and academic interactions highlight the attainment of EA2.

Justification of EA3 Attainment : The novelty of the study is its approach to the complex issue of the prediction of the presence of food allergy through the interaction of epigenetics, lifestyle, and environmental factors with the state of the art machine learning. Logistic Regression was used to provide a baseline, and the nonlinear patterns and interactions of features were captured by random Forest and XGBoost. Creative engineering of new features (e.g. dietary grouping, risk scores) and hybrid approach to data collection (online and in-person) also evidences methodological creativity, cross-disciplinary integration and practical flexibility.

Justification of EA4 Attainment : The study benefits the social community because it facilitates the early detection and prevention of allergies to food reducing their health risks and enhancing the quality of life of patients. It assists medical workers with evidence-based information, encouraging healthier eating habits and preventive treatment. Since the work will be primarily based on computational modeling and data collection through surveys, it will produce an insignificant environmental effect; nevertheless, it will have scaled benefits to the health of the general population.

Justification of EA5 Attainment : The study shows the knowledgeability of the recent engineering and scientific practices involving the sophisticated machine learning, biomedical data analysis and use of epigenetics. It also demonstrates the knowledge of current clinical practices in the diagnosis of food allergy and the way computational solutions can be used in addition to or to enhance it.

5.5 Summary

This part involved in-depth implementation and evaluation of the proposed methodology for predicting food allergy for epigenetic and clinical information. The research workflow involved several engineering activities such as the problem identification, data preprocessing, model design, investigation, and evaluation of the model's performance. Three machine learning algorithms, namely Logistic Regression, Random Forest and XGBoost, were implemented and systematically tested. The experimental results highlighted that ensemble based models especially Random Forest and XGBoost outperformed the logistic regression baseline model with regard to accuracy and robustness. Feature importance analysis further showed some epigenetic markers and clinical features offered significant contribution to predictive power, supporting the use as possible biomarkers for allergy diagnosis. Engineering activities comprising application of the fundamentals, design of computational pipelines, use of modern tools and investigation through systematic experimentation were mapped and justified in alignment with the project objectives. The broader societal context of the research was also taken into account with the highlight how machine learning driven diagnostics can be used to contribute to more accurate, non-invasive and more personalised healthcare solutions when it comes to allergy management.

To summarise this chapter, it showed not only the technical feasibility of integrating epigenetics and machine learning but also proved issues of practical relevance concerning the applicability of the research to solve food allergy prediction. The results are presented in a way that paves the way for a discussion of the overall conclusion, limitations and future working directions, in the next chapter.

Chapter 6

Conclusion

This research, "Beyond the Bite: Unraveling the Mysteries of Food Allergy Through Epigenetics and Machine Learning", aimed to explore how studying out molecular insights (retrieved from epigenetic data) can be combined through machine learning (ML) approaches to better predict and diagnose food allergies. Traditional diagnostic methods are still limited in accuracy, as well as in their ability to describe the causes of differences in allergic reactions between individuals. By combining both epigenetics markers and computational intelligence, this study becomes a platform for more focused and tailored food allergy studies.

6.1 Summary

This research examined the similarities between epigenetics and machine learning models that could enhance the prediction and understanding of food allergy. Traditional diagnostic methods, such as skin prick and IgE blood tests, remain lack to accurates and fail to give a picture of molecular mechanisms behind allergic responses. By capitalizing on DNA methylation patterns, survey data clinical survey data and computational intelligence, this study is part of the emerging field of epigenetic driven allergy prediction.

Several machine learning models were evaluated, Logistic Regression, Random Forest and XG Boost Classifier. The results showed promising results indicating that the models which utilize ensemble methodologies such as Random Forest and XGBoost could provide far better predictive modeling than the baseline models as they tend to capture difficult , non linear relationships in high dimensional biological data. Moreover, analysis of feature importance identified important epigenetic biomarkers which can be the possible manifestations of allergy susceptibility. These findings not only confirm the role of epigenetics in the formation of allergies, they also open a path to the creation of more individualized diagnostic tools.

The combination of surveying response with molecular features, however, further improved the performance of the prediction in suggesting the demand for hybrid data sets that incorporate clinical, environmental, and epigenetic data. Overall, this research demonstrates the feasibility and value of implementing machine learning techniques in food allergy diagnostics, paving the way for more accurate, accessible and insightful biomedical healthcare solutions.

6.2 Limitation

Focus on Select ML Models : Focus on Select ML Models : There are three models were tested. more advanced method of deep learning architectures (e.g., CNNs, Transformers) may reveal more hidden patterns.

Cross sectional Design: The present study received no longitudinal data, which would have been useful to track where epigenetic forms can be evaluated over time, as they might change in relying on allergens.

Clinical Translation: While promising, it's true that the models have yet to be validated in clinical environments which is a prerequisite for the real world adoption.

6.3 Future Work

While the study has promising results, there are still a few new areas for further exploration:

1. **Larger and Diverse Datasets :** The current dataset While informative, the current dataset is small and limited in scope. Expanding the dataset across populations, age groups, and across diets will increase the generalizability of the models.
2. **Deep Learning Architectures :** The future studies can utilize deep learning algorithms like Convolutional Neural Networks (CNNs) and Transformer-based architectures (e.g., BERT for Genomics) for detecting hidden patterns of epigenetic data.
3. **Multi-Omics Integration :** Integration of epigenomic data with genomics, transcriptomics and proteomics could give more generalistic knowledge of allergy mechanism.
4. **Longitudinal Studies :** Monitoring epigenetic alterations over time would make it possible to develop a predictive model on the onset of allergy, progression and response to treatment.
5. **Clinical Validation :** Collaborations with healthcare institutions are required to validate the predictive models in the real world clinical settings before putting this in operation as a diagnostic support tool.
6. **Explainable AI (XAI):** Future research should include techniques to make predictions more explainable so that they are transparent and interpretable to clinicians and (more importantly!) patients.

References

- [1] K. Gassner, B. Vollmer, M. Prehn, M. Fiedler, and S. Ssmoller, "Smart food: Mobile guidance for food-allergic people," in *Proc. 7th IEEE Int. Conf. E-Commerce Technol. (CEC'05)*, 2005, pp. 85–88.
- [2] B. Sowmiya, R. Ramaswamy, and T. Shukla, "Analysing food allergy dataset to predict different types of allergies," *J. Phys.: Conf. Ser.*, vol. 1362, no. 1, p. 012010, 2019.
- [3] J. Zhang et al., "Prediction of oral food challenge outcomes via ensemble learning," *arXiv preprint arXiv:2208.08268*, 2022.
- [4] H. Kim, R. Venkataramanan, and A. Sheth, "A survey on food ingredient substitutions," *arXiv preprint arXiv:2501.01958*, 2024.
- [5] S. Ilhan Omurca et al., "Using machine learning approaches for prediction of the types of asthmatic allergy across Turkey," *J. Phys.: Conf. Ser.*, vol. 1362, no. 1, p. 012010, 2019.
- [6] P. L. Khor and W. J. K. Raymond, "Food allergen detection in Malaysian food using convolutional neural networks," *Proc. 2022 IEEE Int. Conf. Comput. Intell. (ICCI)*, 2022, pp. 1–6.
- [7] R. Gutiérrez Rivas et al., "Real-time allergy detection," in *Proc. 8th IEEE Int. Symp. Intell. Signal Process.*, 2013, pp. 1–6.
- [8] E. Daraghmi, "EZBuy: Personalized shopping assistance system," Master's thesis, 2011.
- [9] J. McWilliam et al., "Skin prick test predictive values for the outcome of cashew challenges in children," *J. Allergy Clin. Immunol.*, vol. 145, no. 1, pp. 232–234, 2020.
- [10] R. Vuillermine et al., "Timing of allergenic food introduction to the infant diet and risk of allergic or autoimmune disease: A systematic review and meta-analysis," *JAMA*, vol. 316, no. 11, pp. 1181–1192, 2016.
- [11] A. Aubrey, "Asthma drug can help reduce allergic reactions for those with severe food allergies," *NPR*, Feb. 26, 2024.
- [12] Y. Persaud, R. J. Memon, and M. N. Savliwala, "Allergy immunotherapy," in *StatPearls*, StatPearls Publishing, 2023.

- [13] L. J. Arias et al., "Tolerance of a cow's milk-based hydrolyzed formula in patients with eosinophilic esophagitis triggered by milk," *Allergy*, vol. 68, no. 8, pp. 1065–1072, 2012.
- [14] A. Lucendo et al., "Omega-3 long-chain PUFA intake during pregnancy and allergic disease outcomes in the offspring: A systematic review and meta-analysis of observational studies and randomized controlled trials," *Am. J. Clin. Nutr.*, vol. 103, no. 2, pp. 486–494, 2016.
- [15] A. K. Singh et al., "Food allergy: Epidemiology, pathogenesis, diagnosis, and treatment," *J. Allergy Clin. Immunol.*, vol. 133, no. 2, pp. 291–307, 2014.
- [16] M. C. Lee et al., "Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management," *J. Allergy Clin. Immunol.*, vol. 142, no. 5, pp. 1232–1243, 2018.
- [17] R. J. Memon et al., "Non-IgE-mediated gastrointestinal food allergy," *J. Allergy Clin. Immunol.*, vol. 135, no. 5, pp. 1114–1122, 2015.
- [18] A. M. Savliwala et al., "Food allergy: An overview," *J. Allergy Clin. Immunol.*, vol. 130, no. 4, pp. 1010–1017, 2012.
- [19] T. B. G. A. Mota-Huertas et al., "Effects of early nutritional interventions on the development of atopic disease in infants and children: The role of maternal dietary restriction, breastfeeding, timing of introduction of complementary foods, and hydrolyzed formulas," *Pediatrics*, vol. 121, no. 1, pp. 1–9, 2008.
- [20] M. J. Allen et al., "Food allergy and risk assessment: Current status and future directions," *Curr. Allergy Asthma Rep.*, vol. 19, no. 4, p. 18, 2019.

25% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

- 6 Not Cited or Quoted 24%**
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 23% Internet sources**
- 1% Publications**
- 22% Submitted works (Student Papers)**

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

25% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

-  **2 AI-generated only 25%**
Likely AI-generated text from a large-language model.
-  **0 AI-generated text that was AI-paraphrased 0%**
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI-generated as AI-generated and AI-paraphrased or likely AI-generated and AI-paraphrased writing as only AI-generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

