Design and Development of Computationally Intelligent Framework for Protein Allergen Classification

A Thesis

submitted in partial fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

in

Department of Computer Science and Engineering

by

Arun Kumar
(Reg no: 951603006)

Under the Supervision of

Dr. Prashant Singh Rana
(Associate Professor)

Thapar Institute of Engineering and Technology, Patiala
147005, Punjab, India
June 2024

Candidate Declaration

I hereby certify that the work, which is being presented in the thesis, entitled **Design and Development of Computationally Intelligent Framework for Protein Allergen Classification** , in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed University), Patiala, India is an authentic record of my own work carried out under the supervision of **Dr. Prashant Singh Rana**. I have also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

The matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any institution.

Date: 19 June 2024

Arun Kumar
Regd. No: 951603006

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Dr. Prashant Singh Rana
Associate Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

To the green canvas of my farms and the air that breathes life into the streets of my village, where my consciousness is deeply rooted—enabling me to spread happiness and peace.

Abstract

The prevalence of allergic reactions is a serious public health problem that may cause diseases like high fever, rhinitis, asthma, dermatitis etc., and affects a sizeable fraction of the world's population. Protein allergens are a key contributor to the development of allergic reactions; hence, locating and describing these allergens is essential to the research and development of efficient diagnostic and therapeutic methods. Currently, the methods that can be used to cure allergies are not completely understood, and the only strategy that is known for preventing allergies is to abstain from substances that contain allergens. The traditional techniques of allergen prediction take a lot of time and frequently depend on methods of experimentation that require a lot of manual resources. Also, the inaccurate detection of allergens may result in excessive dietary restrictions, which can then lead to nutritional issues. Hence, it is clear that an efficient mechanism that can amalgamate multiple parameters for accurate allergen detection is necessary for allergy control. The computational methods combined with bioinformatics have the ability to analyze multiple characteristics of sequences and structures of protein allergens for detecting potentially allergenic areas with better detection accuracy. The work presented in this thesis describes a thorough analysis of various state-of-the-art literature mechanisms and related approaches, resulting in the creation of various proposed mechanisms for the detection, prevention and control of allergies. The mechanisms proposed in the thesis used various computationally intelligent approaches like machine learning, deep learning, ensemble learning and stochastic modelling. A deep learning based ensemble approach for protein allergen classification is proposed in the thesis and has an accuracy of 89.16%, which is higher than other related work in the field. A plant and animal food allergen classification using extra tree based ensemble model is also presented, and the proposed model achieved an accuracy of 94.23% with 10 fold cross-validation. A numerical simulations of a six-compartment model was also conducted in order to predict the likelihood of developing allergies. The numerical presentation of the system's solution was achieved through the utilisation of the stochastic computational artificial neural network (ANN) and the Levenberg-Marquardt backpropagation (LMBP). The validation of the model was achieved by comparing the results with a reference model. The overall combination of various proposed mechanisms and techniques in the thesis presents a computationally intelligent framework for the efficient detection of protein allergens.

Keywords: Allergy, Allergen, Machine Learning, Ensemble Learning, Stochastic Model

Acknowledgement

I am grateful to the Almighty for granting countless blessings, opportunities, knowledge, and strength, without which I could not have accomplished this work.

I am indebted to my supervisors, Dr. Prashant Singh Rana (Associate Professor, TIET), for his continuous guidance and support. I have benefitted greatly from his wealth of knowledge and meticulous editing. The time and energy they devoted to this work have immensely improved the quality of this research. His faith in me and encouragement have helped me stay focused over the years. His honest approach towards research and life is a source of inspiration, which I hope to carry throughout my career.

I offer my sincere gratitude to our director Prof. Padmakumar Nair, Dr. N. Tejo Prakash (Dean, Research Sponsored Projects) and Dr. Shalini Batra (Professor and Head, CSED) for providing the necessary academic and administrative assistance in the completion of this work. I am thankful to my Doctoral committee members- Dr. A. K. Verma (Professor, CSED), and Dr. M.S. Reddy (Professor, Biotechnology),for ensuring the progress of my research work. I am also thankful to Ph.D. Coordinator Dr. Sushma Jain (Associate Professor, CSED) and all the faculty and staff members of CSED for always helping me. My special thanks to Dr. Sanmeet Kaur (Associate Professor, Eastern Washington University), Dr. Parteek Bhatia(Professor, CSED TIET) and Dr. Akhilesh Tiwari(Professor, MITS Gwalior) for holding my hands during my difficult time.

It would not have been possible without my mother's unconditional love and belief in me. My deepest thanks to my dear father, mother, brother, and sisters for encouraging me to start this journey and providing me with every possible support. The love showered by family and cousins has also been a great motivation in this journey. I will always be grateful to them for their patience and compassion in every sphere of my life.

I wish to extend my thanks to all the lab mates and peers for their constant support and for making this journey memorable. I treasure Komal Singh Gill, Arwinder Dhillon, Mexson Fernandes, Rajesh Kondabala, Mir Shahnawaz Ahmad, Shahid Ahmad Bhat, Vagish Dwibedi, Piyush Sharma, Davinder Maan, Yadveer Kaur Batth and Spandita Dwivedi for their support and encouragement. I gratefully acknowledge their cooperation and blessings, which motivated me throughout this degree.

Arun Kumar
June, 2024

Table of Contents

Certificate . . . . . **i**

Abstract . . . . . **iii**

Acknowledgement . . . . . **iv**

Table of Contests . . . . . **v**

List of Figures . . . . . **viii**

List of Tables . . . . . **x**

List of Abbreviations . . . . . **xi**

Chapter 1 Introduction **1**

1.1 Allergy 1

1.1.1 The Background of Allergies 2

1.1.2 Types of Allergies 6

1.2 Allergy Prevalence 8

1.2.1 Factors Contributing in Increasing Prevalence of Allergic Diseases 8

1.3 Global Agencies involved in Allergy Prevention, Detection and Control . 9

1.3.1 World Health Organization (WHO) 9

1.3.2 International Union of Immunological Societies (IUIS) 11

1.3.3 Food and Agriculture Organization (FAO) 12

1.3.4 World Allergy Organization (WAO) 13

1.3.5 American Academy of Allergy, Asthma, and Immunology (AAAAI) 13

1.3.6 European Academy of Allergy and Clinical Immunology (EAACI) 13

1.3.7 Allergy and Asthma Network (AAN) 14

1.3.8 Asthma and Allergy Foundation of America (AAFA) 14

1.4 Allergen 14

1.4.1 Allergen Types 15

1.4.2 Allergen Nomenclature 18

1.5 Computational Intelligent Approaches for Allergen Detection 19

1.5.1 Machine Learning based Approach 19

1.5.2 Deep Learning based Approach 20

1.5.3 Ensemble Learning based Approach 20
1.5.4 Stochastic Mathematical Model based Approach 20
1.6 Thesis Contribution 21
1.7 Thesis Organization 21

Chapter 2 Literature Survey 23
2.1 Computational Methods for Allergy Detection 23
2.1.1 Machine Learning methods for Allergy Detection, Prevention, and Control 23
2.1.2 Deep Learning methods for Allergy Detection, Prevention, and Control 31
2.1.3 Tools used for allergy detection, prevention, and control 34
2.2 Problem Statement 38
2.3 Research Gaps Identified 39
2.4 Research Objectives 39

Chapter 3 Research Methodology 40
3.1 Requirement Specifications 40
3.1.1 Software Requirements 41
3.1.2 Hardware Requirements 42
3.2 Computational Intelligent Approaches 42
3.3 Summary 45

Chapter 4 A Deep Learning based Ensemble Approach for Protein Allergen Classification 46
4.1 Overview 46
4.2 Proposed Mechanism for Allergen Detection 47
4.3 Results and Discussion 54
4.3.1 Performance Parameters 54
4.3.2 Result Analysis 55
4.3.3 Performance Comparison 62
4.4 Summary 63

Chapter 5 Plant and Animal Food Allergen Classification using Extra Tree based Ensemble Model 64
5.1 Overview 64
5.2 Methodology 65
5.2.1 Data set and its Features 65
5.2.2 Features Extraction 65

5.2.3 Data Cleansing . . . 65
5.2.4 Model Building . . . 68
5.2.5 Ensemble Model Building . . . 68
5.2.6 Methodology . . . 70
5.3 Proposed Framework for Food Allergen Classification . . . 72
5.4 Model Evaluation Parameters . . . 73
5.5 Result Analysis and Discussion . . . 75
5.6 Summary . . . 79

Chapter 6 Stochastic Mathematical Model for Allergy Prevalence Prediction . . . 80
6.1 Overview . . . 80
6.2 Designed mathematical model and stochastic framework . . . 82
6.3 Numerical performances of the model . . . 89
6.3.1 Performance Analysis . . . 91
6.3.2 Comparative Analysis . . . 98
6.4 Summary . . . 101

Chapter 7 Conclusions and Future Work . . . 102
7.1 Conclusion . . . 102
7.1.1 Feature Extraction and Selection . . . 102
7.1.2 Deep Learning based Ensemble Approach for Protein Allergen Classification . . . 103
7.1.3 Plant and Animal Food Allergen Classification . . . 103
7.1.4 Allergy Prevalence Prediction using Stochastic Mathematical Model 103
7.2 Future Work . . . 103

List of Publications . . . 105

References . . . 106

List of Figures

Figure No.	Title	Page No.
1.1	Allergic Reaction Mechanism	2
1.2	Key historical aspects of Allergy	3
1.3	Types of Allergies	6
1.4	Agencies related to Allergy Prevention, Detection and Control	10
1.5	Allergen Types	15
3.1	Flow of Research Methodology to attain the proposed objectives	40
4.1	System Model	51
4.2	Architecture of Neural Networks used in the study	53
4.3	Training Loss for Deep Learning Models	55
4.4	Allergen Classification Accuracy of various models	57
4.5	False Alarm Rate of various models	57
4.6	F1-Score of various models	58
4.7	Specificity of various models	58
4.8	MCC of various models	59
4.9	ROC curves of various models (Part-I)	60
4.10	ROC curves of various models (Part-II)	60
4.11	PR curves of various models (Part-I)	61
4.12	PR curves of various models (Part-II)	61
4.13	Confusion matrix representing model validation results	62
4.14	Comparison of the proposed model with related work	62
5.1	System Model	73
5.2	Features Importance in Extra Tree Classification Model	76
5.3	Confusion Matrix for Extra Tree Classifier Model after Data Cleansing Operations	77
5.4	Extra Tree classifier Model Boundary	78
5.5	Roc for Extra Tree Classifier	78
5.6	Class Error Report for Extra Tree Classifier	79
6.1	Model framework representations	83
6.2	Designed procedure for the nonlinear system	85

6.3 A structure of the single neuron 85
6.4 Neural structure for solving the mathematical system 88
6.5 Best validation and gradient for the nonlinear model 92
6.6 Outputs and EHs for solving the nonlinear model 93
6.7 Regression of the model (1) 94
6.8 Regression of the model (2) 95
6.9 Regression of the model (3) 95
6.10 Results comparison for solving the allergen sensitization occurrence model 96
6.11 Performances of AE for the allergen sensitization occurrence model . . . 98

List of Tables

Table No.	Title	Page No.
2.1	Analysis of machine learning models for allergy diagnosis	28
2.2	Analysis of Deep learning models for allergy diagnosis	33
2.3	Tools used for allergy detection, prevention, and control	37
4.1	Description of Dataset Features	48
4.2	Confusion Matrix	54
4.3	Classification Results of Various Learning Models on the Test Dataset	56
5.1	Physicochemical Properties Extracted from Allergen Sequences	66
5.2	Stack Generation Member classifier learning	69
5.3	Performance Comparison of various Model on Accuracy, F1 Score and MCC Parameters	75
5.4	Description of confusion matrix	77
6.1	Parameter details of the prevalence prediction of allergies model	84
6.2	Proposed scheme for the nonlinear model	94
6.3	Comparison of Proposed Mechanism with Related Work	99

List of Abbreviations

3D	Three Dimension
AAAAI	American Academy Of Allergy, Asthma, And Immunology
AAC	Amino Acid Composition
AAFA	Asthma And Allergy Foundation Of America
AAN	Allergy And Asthma Network
AB	Adaboost
AE	Absolute Error
AI	Artificial Intelligence
ANN	Artificial Neural Network
APP	Allergy Prevalence Prediction
ASPs	Allergen-Specific Peptide Sequences
AUC	Area Under The Curve
BCE	B-Cell Epitopes
BLAST	Basic Local Alignment Search Tool
CatBoost	Categorical Boosting
CKSAAP	Composition Of K-Spaced Amino Acid Pairs
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CRAN	Comprehensive R Archive Network
CRD	Component-Resolved Diagnostics
DBN	Deep Belief Network
DL	Deep Learning
DPC	Dipeptide Composition
DT	Decision Tree
EAACI	European Academy Of Allergy And Clinical Immunology
EH	Error Histogram
ELISA	Enzyme-Linked Immunosorbent Assay
FAO	Food And Agriculture Organization
FDA	Food And Drug Administration
FN	False Negative
FP	False Positive
GB	Gradient Boosting
IgE	Immunoglobulin E
IUIS	International Union Of Immunological Societies

kNN	K- Nearest Neighbour
LASSO	Least Absolute Shrinkage And Selection Operator
LDA	Linear Discriminant Analysis
LightGBM	Light Gradient Boosting Machine
LMBP	Levenberg-Marquardt Backpropagation
LR	Logistic Regression
LS	Linear Svm
LSTM	Long Short-Term Memory
LUCCK	Learning Using Concave And Convex Kernels
MATLAB	Matrix Laboratory
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multilayer Perceptron
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Square Error
mw	Molecular Weight
NB	Naive Bayes
nftool	Neural Fitting Tool
NLP	Natural Language Processing
NumPy	Numerical Python
OFCs	Oral Food Challenges
OIT	Oral Immunotherapy
PC	Principal Component
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
pLM4Alg	Protein Language Model-Based Predictors For Allergenic Proteins And Peptides
PLS	Partial Least Squares
PPIs	Protein-Protein Interactions
PR curves	Precision-Recall Curve
protFP	Protein Fingerprint
PseAAC	Pseudo Amino Acid Composition
QDA	Quadratic Discriminant Analysis
QSAR	Quantitative Structure-Activity Relationship
RAST	Radioallergosorbent Test
RBM	Restricted Boltzmann Machines
RF	Random Forest
RMSE	Root Mean Squared Error

SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SEIR	Susceptible, Exposed, Infected, And Recovered,
SMOTE	Synthetic Minority Oversampling Technique
SPTs	Skin Prick Tests
STs	State Transitions
SVM	Support Vector Machine
TLC	Thin-Layer Chromatography
TN	True Negative
TP	True Positive
WAO	World Allergy Organization
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting
β – **LG**	β – Lactoglobulin

Chapter 1

Introduction

Allergy is an exaggerated response of the immune system to a foreign substance, referred to as an allergen. Upon the entry of an allergen into the body, the immune system generates antibodies, mainly immunoglobulin E (IgE), that attach to the allergen and initiate the discharge of histamine and other substances [1]. These substances induce inflammation and other symptoms such as pruritus, sneezing, and dyspnea. Allergy can present itself in several manifestations, including dermatitis, urticaria, bronchial asthma, and anaphylactic shock [2].

This chapter gives an overview of allergy and causes of allergy. To detect the allergy-causing substance, i.e. allergen, Machine Learning (ML) and Deep Learning (DL) based ensemble approaches are discussed. These approaches are used to build a model to classify the protein sequences in the allergen and non-allergen categories, further into the plant and animal food allergen categories. A stochastic method is also used for the prediction of allergy prevalence. The chapter concludes with the overall contribution of this research work and its organization.

1.1 Allergy

The term "allergy" refers to an exaggerated immune system response to a foreign substance, known as an allergen [3]. The production of specific antibodies, such as immunoglobulin E (IgE), occurs in response to the presence of an allergen. In the context of subsequent exposure to the identical allergen, it has been observed that the antibodies previously generated by the immune system exhibit a propensity to adhere to the allergen. This interaction subsequently initiates a cascade of events leading to the liberation of histamine and various other chemical mediators. These mediators are known to induce inflammation and elicit the manifestation of allergic symptoms, as shown in Figure 1.1. The symptoms associated with this condition exhibit a broad range of manifestations, encompassing skin rashes, hives, itching, sneezing, nasal congestion, wheezing, difficulty breathing, and in more severe instances, anaphylaxis [4].

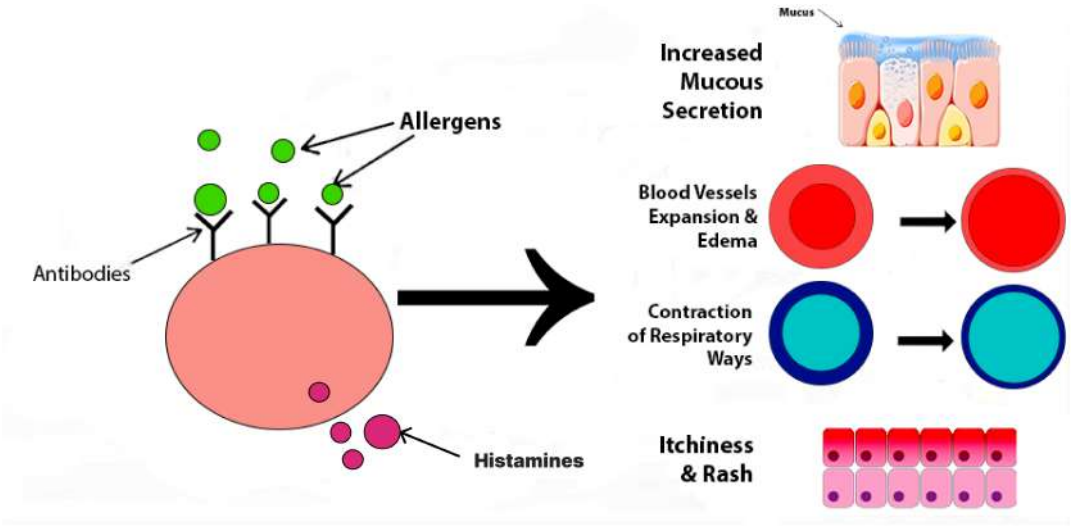


Figure 1.1: Allergic Reaction Mechanism

Allergy is a condition that can be initiated by a wide variety of allergens, encompassing pollen, dust mites, animal dander, specific food items, and medications. The field of allergy and immunology encompasses the investigation of allergy and its underlying mechanisms. This area of research focuses on comprehending how the immune system reacts to foreign substances and the subsequent development of allergic reactions.

1.1.1 The Background of Allergies

Historical evidence of allergy in the human species is found to be from ancient times. The exact term allergy was coined in the 20th century. Starting from ancient times till current times, the spread and advancement in the area of allergy are shown in Figure 1.2.

1.1.1.1 Ancient Times to 18th Century

According to historical evidence, the origins of allergies can be traced back to Ancient Egypt during approximately 1500 B.C. [5]. The first recorded instance of a potential allergic reaction was documented in the "Ebers Papyrus", an ancient Egyptian medical text. This historical document highlighted an observed response to a particular plant, indicating the presence of an allergic reaction. The renowned Greek physician *Hippocrates (460–370 B.C.)* also made note of reactions that were similar to allergies. He discovered that certain people's health might deteriorate after eating particular meals. Later on, in the Roman era, the philosopher *Lucretius* observed that "what is food to one person may be bitter poison to others".

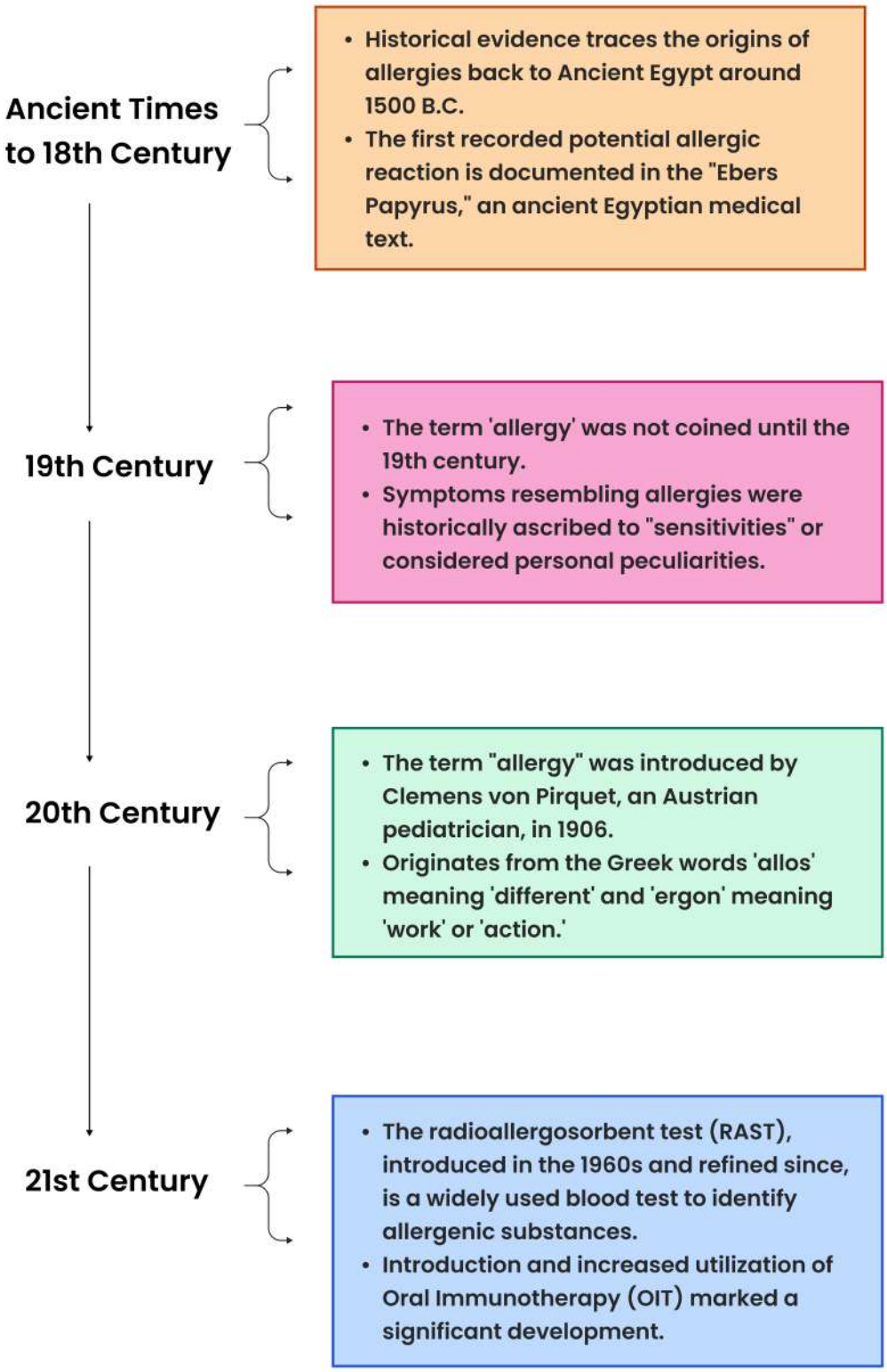


Figure 1.2: Key historical aspects of Allergy

1.1.1.2 19th Century

The term "allergy" was not coined until the 19^{th} century. Historically, symptoms resembling allergies were frequently ascribed to "sensitivities" or regarded as personal peculiarities. The accurate description of hay fever as a disease was first documented by British physician John Bostock in the year 1819. The individual meticulously documented their personal experiences with nasal symptoms that varied throughout different seasons. Through careful investigation, they identified rose pollen as the underlying cause of their symptoms [6].

1.1.1.3 20th Century

The term "allergy" was initially introduced by Clemens von Pirquet, an Austrian paediatrician, in the year 1906. The term was coined by the individual in question subsequent to their observation that certain symptoms exhibited by their patients appeared to be triggered by external factors such as food, dander, or pollen. The term allergy originates from the Greek words allos, which translates to different, and ergon, which translates to work or action [7].

During the 1930s and 1940s, a significant breakthrough in the field of allergy research occurred with the identification of the connection between the immune system and allergic reactions [7]. Pioneer allergists such as Arthur Coca played a pivotal role in this discovery. Their investigations shed light on the intricate mechanisms underlying the immune response to allergens, leading to a better understanding of allergic diseases. By elucidating the relationship between the immune system and allergy, the individual in question pioneered the creation of the initial skin testing panels designed for the purpose of allergy detection. These panels played a crucial role in facilitating the identification of specific allergens to which an individual may be susceptible.

Robert A. Cooke, an allergist, proposed the theory in 1943 that the immune system of the human body might view allergens as potential threats, inducing an immune response [5].

During the 1960s, a significant breakthrough in the field of immunology was made by scientists Kimishige and Teruko Ishizaka. Their research led to the identification of a novel antibody class known as Immunoglobulin E (IgE), which has since been recognised for its pivotal involvement in allergic reactions [8].

1.1.1.4 21st Century

In the 21^{st} century, there have been significant advancements in the field of allergy testing, leading to a more refined and accurate diagnostic process. The radioallergosorbent test (RAST), which was first introduced in the 1960s and has undergone subsequent refinements, is a widely employed blood test for identifying allergenic substances to which an individual may be sensitive. During the early 2000s, the field of diagnostic medicine witnessed the emergence of Component-Resolved Diagnostics (CRD). The utilization of this method enables a more precise diagnostic process and an enhanced assessment of the potential for severe reactions, particularly in cases of food allergies [8].

During the 2010s, the introduction and subsequent increase in utilization of Oral Immunotherapy (OIT) marked a significant development in the field of allergy research and treatment. This is a medical treatment guided by an allergist, which aims to desensitize individuals to allergens by consuming the allergen in small, increasing amounts [9].

In recent years, there has been a notable advancement in the field of medical research on treating severe cases of asthma and other allergic conditions. Specifically, by the year 2020, a novel approach involving the utilization of biological therapies has gained significant attention. These therapies are designed to specifically target molecules that play a crucial role in the allergic response, such as monoclonal antibodies. As a result, they have emerged as a promising treatment [10].

Currently, ongoing research is focused on developing therapies aimed at enhancing the effectiveness and efficiency of treatments for various allergies. The field of molecular allergology and immunotherapy has witnessed significant advancements [11].

The study of allergies has been a subject of extensive research throughout medical history, characterized by significant milestones and profound insights. This ongoing chronicle in the annals of medical history is marked by a rich narrative. The current understanding of immunology has been greatly enhanced by recent advancements in both immunology and technology. These developments have paved the way for a more comprehensive investigation into the intricate immunological mechanisms that are involved.

The continuous pursuit of providing relief to the millions of individuals affected by allergic disorders is marked by each discovery and therapy. As the exploration of the complexities of the immune system progresses, the potential for advancements in allergy research and treatment appears to be on the verge of unprecedented opportunities. The potential of this research is to contribute to the advancement of our understanding, the improvement of treatment approaches, and the ultimate attainment of comprehensive control over these prevalent yet intricate health phenomena.

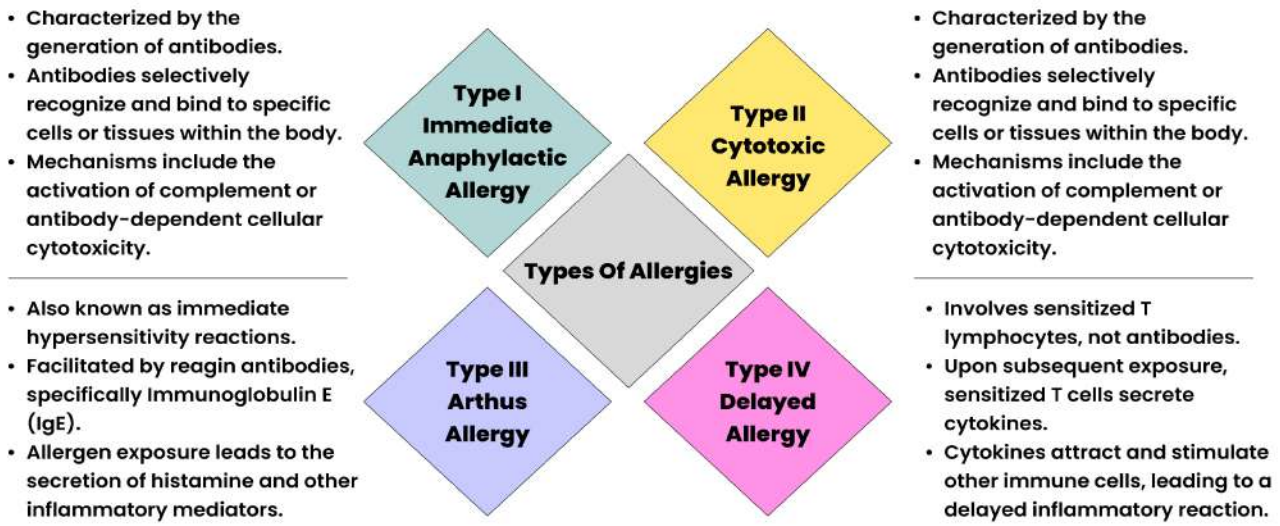


Figure 1.3: Types of Allergies

1.1.2 Types of Allergies

Allergic reactions are intricate immunological responses that can appear in many ways, impacting numerous tissues and organs inside the body. Comprehending many forms of allergic responses is essential for precise diagnosis, efficient management, and focused treatment approaches [12]. The categorization of allergic responses has undergone changes throughout time as shown in Figure 1.3. Based on the primary forms of allergic tissue reactions and their clinical consequences, allergic reactions are classified as follows:

1.1.2.1 Type I Immediate Anaphylactic Allergy

Type I allergic reactions also referred to as immediate hypersensitivity reactions, are distinguished by their swift initiation and are facilitated by reagin antibodies, specifically immunoglobulin E (IgE). When an individual who possesses a genetic susceptibility to Type I allergies comes into contact with an allergen, such as pollen, animal dander, or specific food items, it initiates the secretion of histamine and other inflammatory mediators from mast cells and basophils. The sequence of events described above gives rise to the characteristic manifestations of anaphylaxis, which encompass the presence of hives, swelling, respiratory impairment, and, in more severe instances, cardiovascular failure [13].

1.1.2.2 Type II Cytotoxic Allergy

Type II allergic reactions are characterised by generating antibodies that selectively recognise and bind to particular cells or tissues within the body. Antibodies have been found to

potentially induce cell destruction through diverse mechanisms, including the activation of complement or antibody-dependent cellular cytotoxicity. Type II allergic reactions encompass various conditions, such as autoimmune hemolytic anaemia and specific drug-induced cytopenias. These examples demonstrate the range of manifestations observed in Type II allergies. The comprehension of the fundamental mechanisms involved in Type II allergy is crucial for the advancement of specific therapeutic approaches and the effective control of autoimmune disorders [13].

1.1.2.3 Type III Arthus Allergy

Type III allergic reactions, also known as immune complex-mediated hypersensitivity reactions, exhibit a distinct pattern characterized by the presence of immune complexes composed of antigens and antibodies. These immune complexes have the ability to accumulate in different tissues and subsequently trigger the activation of the complement system. This activation, in turn, results in tissue damage and initiating an inflammatory response. The association of this particular reaction has been observed in various conditions, including serum sickness and specific types of vasculitis. The clinical presentations associated with Type III allergy exhibit considerable heterogeneity, encompassing a broad spectrum of symptoms that can manifest in various ways. These manifestations may include but are not limited to skin rashes, as well as more severe systemic complications such as vasculitis [14].

1.1.2.4 Type IV Delayed Allergy

Type IV allergic reactions, commonly referred to as delayed-type hypersensitivity reactions, are characterized by the involvement of sensitized T lymphocytes instead of antibodies. Upon subsequent exposure to the allergen, the sensitized T cells secrete cytokines, which serve as chemical messengers that attract and stimulate other immune cells. This process ultimately results in a delayed inflammatory reaction. The phenomenon described is frequently observed in various medical conditions, including contact dermatitis, tuberculosis skin tests, and specific drug hypersensitivity reactions. The comprehension of the mechanisms underlying Type IV allergies is of utmost importance in the identification and treatment of these disorders [14].

Various types of allergic reactions pose distinct clinical challenges and necessitate customized therapeutic strategies. In instances of Type I allergic reactions, it is frequently imperative to promptly administer epinephrine and antihistamines as a means to counteract the consequences of histamine release and mitigate the occurrence of anaphylaxis. In contrast, it has been observed that Type IV allergic reactions often necessitate the

implementation of specific immunosuppressive therapies in order to effectively regulate the inflammatory response mediated by T cells [15].

Furthermore, it is imperative to identify the precise allergens accountable for each distinct allergic response in order to effectively implement strategies for allergen avoidance and formulate personalized immunotherapy regimens. To effectively diagnose and manage allergic diseases, healthcare professionals must possess a thorough comprehension of the various types of allergic reactions [15].

The elucidation of immunological mechanisms and clinical manifestations associated with different types of allergic reactions is crucial for advancing our understanding in this field. This knowledge can serve as a valuable guide for the development of innovative therapeutic interventions aimed at improving patient outcomes. In the field of immunology and allergy, there is a continuous advancement in research. This progress requires us to constantly re-evaluate and refine the classification of allergic reactions. By doing so, we can further enhance our understanding of the intricate immune responses involved in these reactions.

1.2 Allergy Prevalence

The prevalence of allergies is a measure that quantifies the frequency or proportion of individuals within a population who experience allergic reactions to particular substances or allergens. Allergies refer to immunological responses that manifest as hypersensitivity reactions towards typically innocuous environmental substances. The quantification of allergies is commonly represented in epidemiological studies through the use of percentages or ratios, which indicate the frequency or occurrence of allergies within a specific population [16].

1.2.1 Factors Contributing in Increasing Prevalence of Allergic Diseases

Allergic conditions, such as asthma, atopic dermatitis, allergic rhinitis, and food allergy, have been increasing in prevalence in recent decades. The rising incidence of these illnesses can be attributed to a complex interaction between genetic, environmental, and lifestyle factors [17]. Individuals may be predisposed to acquire allergy disorders due to genetic sensitivity, which can be impacted by family history and genetic differences.

Environmental variables, including allergen exposure, pollution, and climate change, can potentially contribute to the onset of allergies. Lifestyle variables, such as dietary choices,

physical exercise, and stress levels, may also have an impact. An instance of this would be a diet rich in processed foods and lacking an adequate amount of fruits and vegetables, potentially elevating the likelihood of acquiring allergies. The "hygiene hypothesis" posits that reduced exposure to pathogenic microorganisms during early life may also have a role in the onset of allergy conditions. In general, the rising occurrence of allergy disorders can be attributed to a combination of these causes, and more investigation is necessary to comprehensively comprehend their intricate interplay.

1.3 Global Agencies involved in Allergy Prevention, Detection and Control

Allergies have emerged as a significant global health issue, prompting the establishment of various international and regional organizations dedicated to the detection, research, and management of allergic conditions as shown in Figure 1.4. These organizations are crucial in their contribution to the provision of education, resources, and guidelines for the management of allergies, while also actively promoting research in this particular field. Some of these organizations are discussed in this section.

1.3.1 World Health Organization (WHO)

WHO plays a crucial role in addressing allergies on a global scale, with a primary emphasis on the identification and management of allergies, as well as their prevention and treatment. The organization plays various roles in order to support and advance its mission [18], which are as follows:

- **Setting International Standards:** The formulation of evidence-based policies, guidelines, and standards for diagnosing and managing allergies is typically undertaken by authoritative bodies or organizations with expertise in the field. These entities play a crucial role in synthesizing the available scientific evidence and translating it into actionable recommendations for healthcare professionals and policymakers. By employing rigorous methodologies and systematic reviews of the literature, these organizations strive to ensure the policies undergo a collaborative development process involving experts from various parts of the world. They are consistently revised to align with the most current scientific knowledge [18].
- **Research Support:** WHO funds and coordinates studies that seek to increase scientific understanding of allergies, refine the diagnosis procedure, and provide more efficient therapies, from analyses of allergy prevalence in different populations

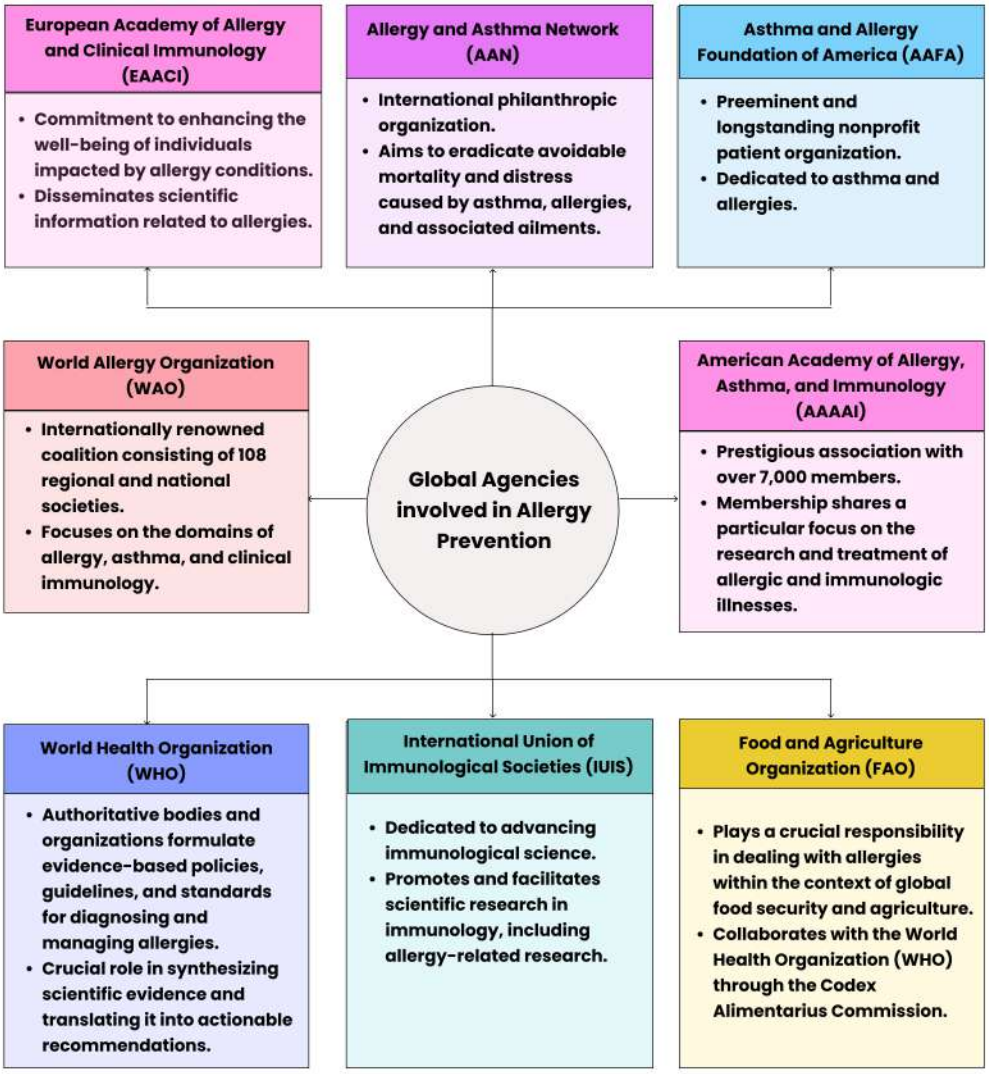


Figure 1.4: Agencies related to Allergy Prevention, Detection and Control

to examinations of the molecular processes underlying allergic responses.

- **Promoting Access to Diagnosis and Treatment:** WHO collaborates closely with its member states to facilitate equitable access to essential resources for the diagnosis and management of allergies on a global scale. The promotion of access to essential medicines and the provision of up-to-date training for healthcare providers in allergy diagnostic and treatment methods are crucial aspects of ensuring comprehensive healthcare services [18].
- **Raising Awareness:** WHO is actively involved in coordinating and implementing global public health campaigns with the primary objective of increasing awareness regarding allergies. These campaigns are designed to address the significant impact of allergies on individuals and communities worldwide. By organizing and lead-

ing these initiatives, WHO aims to promote a better understanding of allergies, their causes, symptoms, and available prevention and treatment options. Through these efforts, WHO strives to enhance public knowledge, foster informed decision-making, and ultimately improve the overall management and control of allergies on a global scale. The primary objective of these campaigns is to disseminate knowledge among the general public regarding the underlying causes and various symptoms associated with allergies. Additionally, they aim to emphasize the significance of promptly seeking professional medical assistance for accurate diagnosis and effective treatment [19].

- **Data Collection:** WHO gathers and examines global data on allergies. This aids the organization in monitoring patterns in the occurrence of allergies, comprehending the worldwide consequences of allergies, and pinpointing regions that want supplementary resources. The data also facilitates the strategic development and assessment of allergy programmes and policies.
- **Building Partnerships:** WHO collaborates with governments, non-profit organizations, and private sector enterprises to combat allergies. These agreements facilitate the integration of allergy detection and control methods into wider public health initiatives [18].

WHO plays a crucial role in improving the worldwide ability to identify and manage allergies, aiming to reduce the impact of allergic disorders and boost the well-being of those affected by allergies.

1.3.2 International Union of Immunological Societies (IUIS)

The International Union of Immunological Societies (IUIS) is an organization dedicated to advancing the field of immunological science and its practical use in medicine, particularly the study and treatment of allergies. Through a range of activities, it plays a prominent role in shaping the global allergy landscape [20], which are as follows:

- **Research Promotion:** IUIS promotes and facilitates scientific research in the field of immunology, including allergy-related research. The platform facilitates the exchange of knowledge and promotes collaboration among researchers, thereby potentially contributing to advancements in the field of allergy detection and control.
- **Education and Training:** IUIS arranges seminars, courses, and conferences, including the International Congress of Immunology, intending to spread the most recent research discoveries and instruct healthcare professionals on the newest knowledge and methods in allergy diagnosis and management.

- **Standardising Nomenclature:** The assignment of systematic names to newly discovered allergens is the responsibility of the IUIS Allergen Nomenclature Sub-Committee. The utilization of standardized nomenclature in the field of allergy research and clinical practice plays a crucial role in promoting effective communication among researchers, clinicians, and patients. By establishing a common language and terminology, standardized nomenclature enhances the accuracy of diagnosing and treating allergies. This, in turn, contributes to improved patient outcomes and facilitates advancements in the field of allergy research.
- **Committees and Task Forces:** Multiple IUIS committees and task groups, such as the Clinical Immunology Committee, work towards the common objective of improving immunology as a field. These groups frequently engage in projects that are closely associated with the detection and management of allergies.
- **Global Networking:** By acting as a link between different national and local immunology organizations, IUIS facilitates the exchange of knowledge and best practices in the field of allergies.

IUIS emphasizes the significance of immunological research and its various applications within the healthcare industry. This includes the field of allergy detection and control, where advancements in immunology have played a crucial role. IUIS plays a significant role in the field of allergy detection and control. It achieves this by actively promoting scientific research in immunology, facilitating education and training initiatives, establishing standardized allergen nomenclature, fostering global collaboration among researchers and professionals, and advocating for the recognition and significance of immunological science in addressing allergies.

1.3.3 Food and Agriculture Organization (FAO)

The Food and Agriculture Organisation (FAO) has a crucial responsibility in dealing with allergies in the context of global food security and agriculture. The FAO collaborates with the World Health Organisation (WHO) via the Codex Alimentarius Commission to develop essential worldwide standards and recommendations for managing allergens. This includes implementing accurate Labelling methods that enable consumers to effectively assess possible dangers [21].

In addition to legislative frameworks, the FAO actively advocates for sustainable and safe agricultural practices, thus reducing the prevalence of allergens in food production. The organization's research endeavours focus on comprehending the prevalence of allergens and assisting in the creation of specific tactics. The FAO's involvement in educational

initiatives increases stakeholders' knowledge, promoting a proactive stance towards preventing allergens. Moreover, its active participation in emergency responses, namely in the context of food recalls, demonstrates its dedication to reducing health hazards related to allergies. The FAO's main responsibility is to address many aspects of food security. However, its detailed contributions have a considerable impact on the prevention and control of allergies worldwide, guaranteeing the safety and welfare of consumers. It is recommended to refer to the newest publications and resources from the FAO for the most up-to-date information [22].

1.3.4 World Allergy Organization (WAO)

The World Allergy Organisation (WAO) is an internationally renowned coalition consisting of 108 regional and national societies that are committed to the domains of allergy, asthma, and clinical immunology. WAO is committed to provide extensive resources and educational materials related to allergies. In addition, the association actively promotes health policies that efficiently tackle allergy-related issues [23]. In addition, WAO plays a vital role in promoting research programmes and encouraging collaboration among allergy practitioners worldwide. Every year, an organization takes the lead in organizing World Allergy Week, a worldwide event with the goal of raising awareness of allergic illnesses and related problems.

1.3.5 American Academy of Allergy, Asthma, and Immunology (AAAAI)

The AAAAI is a prestigious association including over 7,000 members from the United States, Canada, and 72 other nations. This membership encompasses allergists, as well as other medical experts, allied health professionals, and associated healthcare workers, all of whom possess a particular focus on the research and treatment of allergic and immunologic illnesses [24].

1.3.6 European Academy of Allergy and Clinical Immunology (EAACI)

The EAACI is a consortium including medical practitioners, scientists, and other healthcare experts committed to enhancing the well-being of individuals impacted by allergy conditions. They encourage allergy research, disseminate scientific information, and provide advanced training through collaboration with over 50 National Allergy Societies from Europe and other regions [25].

1.3.7 Allergy and Asthma Network (AAN)

The AAN is an international philanthropic organization that seeks to eradicate avoidable mortality and distress caused by asthma, allergies, and associated ailments through outreach, education, advocacy, and research. Their role encompasses the provision of resources, implementation of patient education programmes, and active promotion of policies that prioritize patient satisfaction [26].

1.3.8 Asthma and Allergy Foundation of America (AAFA)

The AAFA is the preeminent and most longstanding nonprofit patient organization committed to asthma and allergies. Their services encompass the provision of pragmatic information, community-oriented assistance, educational support, and active promotion of public policy and research funding [27].

1.4 Allergen

An allergen is primarily characterized as a protein or peptide that has the potential to stimulate an allergic immune response in individuals who are susceptible to allergies. Allergens are substances that can elicit an allergic reaction in individuals. They can originate from diverse sources, including but not limited to pollen, dust mites, animal dander, specific food items, and insect venom. When an allergen is encountered, it has the potential to trigger the production of allergen-specific IgE antibodies and the subsequent release of inflammatory mediators. This process ultimately results in the occurrence of allergic reactions [28]. The complexity and multifactorial nature of allergenicity at the molecular level have been extensively studied. Numerous factors, including the quantity and duration of exposure to the immune system, environmental conditions, and immune-modulating components present in allergenic sources, have been identified as influential in this process.

The categorization of allergens can be achieved by grouping them together according to their common structural, biochemical, and functional characteristics. This approach enables researchers to analyze the potential for cross-reactivity and evaluate the allergenic properties of these substances. The investigation of the relationship between allergens and the immune system has revealed that various allergic conditions can arise as a consequence. These conditions encompass allergic rhinitis, asthma, atopic dermatitis, and food allergies. Hence, it is understood that allergens, typically in the form of proteins or peptides, can trigger an immune response in individuals who are prone to allergies. This

immune response then results in the appearance of clinical symptoms that are commonly associated with allergic diseases [29].

1.4.1 Allergen Types

The origin and features of allergens allow for their classification into distinct kinds as shown in Figure 1.5. Allergens can originate from a variety of sources. The following are examples of typical allergens:

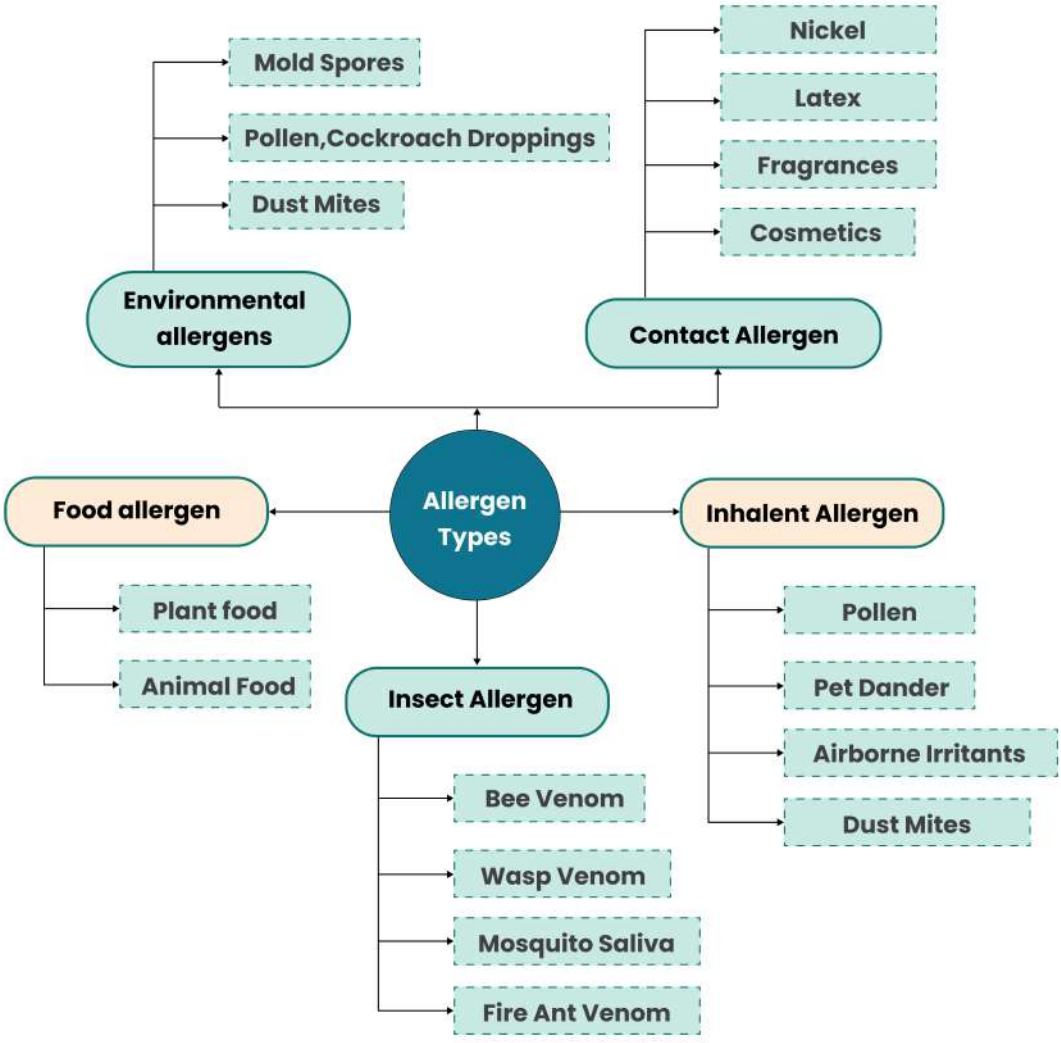


Figure 1.5: Allergen Types

1.4.1.1 Environmental allergens

The presence of allergens in the environment is a well-documented phenomenon. Common examples of these allergens include pollen, dust mites, animal dander, and mould spores.

These substances have been extensively studied due to their potential to trigger allergic reactions in susceptible individuals [30].

1.4.1.2 Food Allergen

A food allergen refers to a specific substance that is found in certain types of food and has the potential to trigger an allergic reaction in individuals who are susceptible to allergies. The occurrence of these reactions is attributed to the immune system's recognition of certain proteins in food as potentially harmful, leading to the initiation of an immune response targeting these proteins. Food allergies encompass a spectrum of severity, ranging from mild manifestations to more severe forms that can potentially result in life-threatening episodes referred to as anaphylaxis [31].

Food allergy symptoms exhibit a broad range of manifestations encompassing various bodily systems. These symptoms encompass dermatological reactions such as the appearance of hives and itching, gastrointestinal symptoms like nausea, vomiting, and diarrhoea, as well as respiratory symptoms including wheezing and coughing. In severe instances, food allergies can lead to anaphylaxis, a life-threatening condition characterised by a sudden decline in blood pressure, respiratory distress, and loss of consciousness [32].

The management of food allergies necessitates the implementation of rigorous avoidance strategies for allergenic foods, meticulous scrutiny of food labels, and the readiness to administer appropriate medications, such as antihistamines or epinephrine, in cases of severe allergic reactions. The process of diagnosing allergies typically includes the utilisation of allergen-specific testing methods, such as skin prick tests or blood tests. The development of personalised management plans is a common practice for individuals with food allergies, as they collaborate closely with healthcare providers [32].

1.4.1.3 Insect Allergen

Insect allergens refer to substances that are derived from insects and possess the capability to induce allergic reactions in individuals who are sensitive to them. The allergens commonly associated with insects are proteins found in different body components, including saliva, urine, faeces, and body parts. The potential for exposure to allergens exists through various means, including bites or stings from insects, inhalation of particles suspended in the air, or direct contact with surfaces that have been contaminated. These allergens can be found in insects such as bees, wasps, and mosquitoes [33].

1.4.1.4 Inhalant Allergen

An inhalant allergen refers to a specific type of substance that has the potential to elicit an allergic response upon inhalation. Allergic reactions manifest as a result of an immune system response to a typically innocuous substance, perceiving it as a potential danger. In the context of inhalant allergens, it is important to note that these substances primarily exist as airborne particles that are commonly inhaled into the respiratory system. This inhalation can subsequently result in the manifestation of diverse symptoms [34]. Common inhalant allergens are categorized as follows:

- **Pollen:** Trees, grasses, and weeds emit these microscopic particles as they reproduce. Seasonal allergic rhinitis, also known as hay fever, is frequently linked to pollen allergies.
- **Pet Dander:** Cats, dogs, birds, and rodents are examples of creatures that shed skin, hair, or feathers, which are then carried through the air by minute particles.
- **Airborne Irritants:** It is possible for certain individuals to experience allergic responses and irritation of the respiratory system when they are exposed to certain substances, such as smoking, pollution, and strong odours.
- **Dust Mites:** Creatures that are microscopic and exist in dust, bedding, and upholstery and flourish there. Allergies might be triggered by the waste particles and body pieces that they produce.

1.4.1.5 Contact Allergen

A contact allergen refers to a substance that has the potential to induce an allergic response when it comes into direct contact with the skin or mucous membranes. In contrast to inhalant allergens, which are introduced into the respiratory system through inhalation, contact allergens elicit immune reactions upon direct contact with the skin or other external surfaces of the body. The allergic reaction commonly exhibits a localized manifestation limited to the specific contact site [35]. Some of the contact allergens are described as follows:

- **Cosmetics:** Contact dermatitis, a common skin condition, can be triggered by certain ingredients found in various personal care products such as makeup, lotions, perfumes, and hair dyes. This condition specifically affects individuals with heightened sensitivity to these substances.
- **Fragrances:** Research has shown that certain fragrances commonly found in perfumes, soaps, and lotions have the potential to induce allergic reactions in individ-

uals.

- **Latex:** Latex, a natural rubber material, is widely utilised in various common products including gloves, balloons, and specific garments. Latex allergies can manifest in certain individuals following direct contact with the skin.
- **Nickel:** Nickel, a widely encountered metal, is frequently found in various items such as jewellery, belt buckles, and metal buttons. It has been identified as a prevalent trigger for contact dermatitis, a condition characterized by skin inflammation upon contact with certain substances.

1.4.2 Allergen Nomenclature

The process of systematically identifying and categorizing compounds that might cause allergic responses in vulnerable people is called allergen nomenclature. To ensure that allergens are consistently and clearly identified and communicated, the World Health Organisation and the International Union for Immunotherapy have developed the WHO/IUIS Allergen Nomenclature [36]. The ability to precisely identify and characterize allergens through the use of this nomenclature is crucial for the diagnosis, treatment, and study of allergies. It also allows for better communication and collaboration among researchers, healthcare providers, and regulatory agencies. Efforts to better understand and treat allergic disorders are supported by the standardized naming system, which helps to minimize misconceptions and confusion among the scientific, clinical, and regulatory groups. The significance of allergen nomenclature [37] can be described as:

- **Clarity and Consistency:** Implementing standardized allergen nomenclature is crucial in maintaining a consistent and unambiguous approach to naming allergens. This practice plays a vital role in preventing any potential confusion or misunderstandings within the scientific, clinical, and regulatory communities. By adhering to a standardized system, researchers and professionals can effectively communicate and exchange information regarding allergens, ultimately enhancing accuracy and reliability.
- **Communication and Collaboration:** Using a common language facilitates efficient communication and collaboration among researchers, healthcare professionals, and regulatory bodies. This common language enables these stakeholders to accurately and efficiently share and interpret data, leading to improved outcomes and advancements in the field.
- **Allergy Diagnosis and Treatment:** Accurate allergy diagnosis and therapy need standardized terminology. It enables the accurate identification and characteriza-

tion of allergens, essential for developing diagnostic tools and treatment materials for individuals with allergies.

- **Research and Development:** A standardized system for naming allergens facilitate research endeavors by establishing a structure for organizing and categorizing data pertaining to allergens. Standardized terminology is crucial in handling and comprehending the vast amount of "Big Data" generated by new technologies, such as various "omics" approaches.

1.5 Computational Intelligent Approaches for Allergen Detection

Computational intelligence is the ability of a computer to learn from experience, i.e. data. The role of computational intelligent approaches in allergen detection is significant, as these approaches utilize advanced algorithms and techniques to analyze data and identify potential allergens [38]. The implementation of these approaches is crucial in order to effectively tackle the intricate and wide-ranging nature of allergens found in different environments. Modern computationally intelligent approaches used in this study include machine learning, deep learning, ensemble learning, and stochastic mathematical model based approaches.

1.5.1 Machine Learning based Approach

Machine learning plays a crucial role in allergen identification [39]. Machine learning algorithms can be trained using allergen data to predict and detect allergenic proteins based on their sequence or structure. These algorithms can extract amino acid composition, physicochemical features, and secondary or tertiary structural components from protein sequences and structures, which help to categorize allergic and non-allergenic proteins.

Machine learning can also predict cross-reactivity when an allergic person reacts to structurally or functionally identical allergenic proteins. This can assist in the design of hypoallergenic foods and allergy treatments by discovering cross-reactive protein families or motifs. Additionally, machine learning can help identify allergenic protein sources (plant, animal, fungal) based on sequence or structural data, which can help discover allergies in new food or environmental samples.

Finally, machine learning can help personalize allergy diagnosis, treatment, and management by combining patient data like genetics and allergen exposure history.

1.5.2 Deep Learning based Approach

In allergen identification, deep learning is crucial because it can sift through vast data sets in search of detailed patterns and produce reliable predictions. Algorithms trained with deep learning use artificial neural networks and can process and analyze amino acid sequences to accurately identify possible allergies [40]. Algorithms like this can classify proteins as either allergic or non-allergenic depending on their intricate properties extracted from protein sequences and structures.

1.5.3 Ensemble Learning based Approach

Ensemble learning is an effective method that strengthens the system's accuracy and resilience by combining the predictions of various machine learning models. To improve the accuracy of protein allergen identification, ensemble learning can be utilized in the context of allergen detection to combine the characteristics of many machine learning models [41].

By lowering the possibility of overfitting and raising the system's stability, ensemble learning can assist in overcoming the shortcomings of individual models. Even when individual models do not work well, ensemble learning can still produce a good allergenicity prediction by aggregating their predictions.

1.5.4 Stochastic Mathematical Model based Approach

The Lotka-Volterra predator-prey model is a well-known mathematical framework that is commonly employed in ecological research to analyze the dynamics between two species within a biological community [42]. This model consists of a set of first-order, nonlinear differential equations, which effectively capture the interactions between a predator species and its corresponding prey species (as shown in Eq. 1.1 and 1.2).

$$\frac{dx}{dt} = \alpha x - \beta xy \tag{1.1}$$

$$\frac{dt}{dt} = \delta xy - \gamma y \tag{1.2}$$

Where, x and y represent the populations of prey and predators respectively. α is the natural rate of growth of prey, β is the rate of capturing of prey by predator, δ is the rate of growth of predator and γ is the natural rate of death of predator.

By utilising this model, researchers can gain valuable insights into the population dynamics and coexistence patterns of these species, shedding light on the intricate relationships that exist within ecological systems [43]. Alfred Lotka and Vito Volterra independently formulated the model under consideration during the early 1900s.

Incorporating the Lotka-Volterra predator-prey model into the examination of allergy prevalence entails conceptualizing allergens as the "prey" and individuals prone to allergies as the "predators". The present analogy explores the dynamics of allergen exposure and the subsequent development of allergies by employing a model that examines the interactions between these two fundamental components. In order to accurately represent the prevalence of allergies, it is necessary to modify the existing model equations. By incorporating relevant factors and variables, we can enhance the model's ability to capture the dynamics of allergy prevalence. This research aims to refine the model equations to better reflect the complex nature of allergies and their impact on populations.

1.6 Thesis Contribution

The major contributions of the thesis are:

- Analysis and classification of related research work:
  - The issues related to the use of traditional and currently available computationally intelligent approaches for allergen identification and allergy prevention and control are presented.
  - An overview of allergen detection techniques related to ML, DL, Ensemble Learning and stochastic modelling is provided.
- A Deep Learning based Ensemble Approach for Protein Allergen Classification.
- Plant and animal food allergen classification using extra tree based ensemble model.
- A stochastic mathematical model for the prediction of allergies.
- Thorough testing and validation of each proposed technique.

1.7 Thesis Organization

This Thesis is organized into the following seven chapters:

Chapter 1: Introduction
This chapter discusses the basics of allergy, its type, and various factors involved in its

prevalence. It focuses on various types of allergens causing allergies. Different agencies involved in allergy detection, prevention, and control are discussed. The significance and use of ML, DL, ensemble learning, and stochastic mathematical methods for allergen detection is also discussed in this chapter.

Chapter 2: Literature Review
In this chapter, a detailed survey of research carried out in the area of allergen-causing allergy is discussed. Different Computational intelligent approaches like Machine Learning, Deep Learning, Ensemble Learning, and stochastic modeling are highlighted.

Chapter 3: Research Methodology
Various software, algorithms, methods, and requirements are discussed in details in this chapter. The chapter describes the overall methodology employed for solving the research problem.

Chapter 4: A Deep Learning based Ensemble Approach for Protein Allergen Classification
To detect whether a protein sequence is going to be allergic or non-allergic, a deep learning-based ensemble framework is discussed in this chapter.

Chapter 5: Plant and Animal Food Allergen Classification using Extra tree Based Ensemble Model
In this chapter, an ensemble framework based on the extra tree algorithm is discussed and the performance of the model is compared with several other algorithms to detect the category of allergen is plant food or animal food.

Chapter 6: Stochastic Mathematical Model for Allergy Prevalence Prediction
This chapter discusses the formulation of a stochastic mathematical model to solve the differential equations depending on various factors involved in the prevalence of allergies. The solution of the differential equation is compared with the trained artificial neural network.

Chapter 7: Conclusion
This chapter summarises all the research findings, concludes the thesis, and provides future research directions.

Chapter 2

Literature Survey

In this chapter, a detailed survey of research carried out in the area of allergen-causing allergy is discussed. The chapter describes in detail the motivation of the research work by highlighting the research problems in the field. Each related work is critically analyzed by discussing the main contributions, algorithms used, results, limitations and future directions. The literature described is divided into two main components: machine learning and deep learning approaches used for allergy detection, prevention, and control. It also discusses various ensemble learning and mathematical models developed in the literature for allergen detection. The chapter also describes various tools which are being used for the design and development of allergen detection frameworks. After a thorough analysis of the literature, the chapter concludes by describing the research problem and objectives.

2.1 Computational Methods for Allergy Detection

This section is divided into two parts: machine learning methods for allergy detection and Deep learning methods for allergy detection. Both methods are discussed in detail in the forthcoming sections.

2.1.1 Machine Learning methods for Allergy Detection, Prevention, and Control

Bin et al. [44] introduced a machine learning algorithm that employs multi-feature fusion to accurately predict allergenic proteins. Initially, they curated a benchmark dataset comprising protein sequences categorized as either allergic or non-allergenic. Subsequently, a preliminary test was conducted on this dataset utilizing a machine learning framework. Three optimal approaches for feature extraction—amino acid composition (AAC), dipeptide composition (DPC), and composition of k-spaced amino acid pairs (CKSAAP), were identified. These features were combined and refined using the Pearson Correlation Coefficient (PCC) and Principal Component Analysis (PCA).

The most significant features were then selected to develop the best predictor using the Random Forest (RF) method. Performance assessment through 5-fold cross-validation

demonstrated that the final model exhibited high accuracy in distinguishing between allergenic and non-allergenic proteins.

Xin et al. [45] developed Quantitative Structure-Activity Relationship (QSAR) models to predict the binding affinity between protein epitopes and IgE. This involved the utilization of four algorithms in tandem with specifically chosen chemical descriptors. The model accurately estimated epitope binding affinities to IgE, achieving an R2 value of 0.7494 and a Root Mean Squared Error (RMSE) of 0.2375.

The model's accuracy was validated using an Enzyme-Linked Immunosorbent Assay (ELISA). Results demonstrated the QSAR model's high effectiveness and precision in forecasting allergic responses to dietary protein epitopes. The predicted outcomes from the model aligned closely with experimental results, showcasing strong consistency with a Pearson correlation coefficient of 0.8956. Both the QSAR model and in vitro tests confirmed that the amino acid sequence 116–130 represents a newly identified epitope of $\beta - lactoglobulin$ ($\beta - LG$) that binds to IgE.

Chinna et al. [46] introduced a data-driven approach utilizing machine learning to detect previously undetected patterns associated with allergens within amino acid sequences. Their methodology facilitates a comprehensive exploration of amino acid sub-sequences displaying heightened statistical prevalence in allergenic proteins.

In a proof-of-concept study, the methodology was applied to a database containing 21,154 proteins, each with known allergy status. This analysis revealed Allergen-Specific Peptide sequences (ASPs) consistent with established biological findings. Employing these ASP-based allergenicity prediction methods demonstrated improved performance compared to current approaches, suggesting the potential value of this method in assessing the viability of synthetic meals and proteins.

Justin et al. [47] proposed the application of machine learning methodologies to predict the outcomes of Oral Food Challenges (OFCs). Authors conducted a retrospective study using data obtained from 1,112 patients who underwent a total of 1,284 OFCs. The dataset encompassed various clinical characteristics, including serum-specific IgE levels, total IgE levels, results of skin prick tests (SPTs), presence of comorbidities, gender, and age. Leveraging these features, multiple machine learning models were developed to forecast the outcomes of OFCs for three prevalent allergens: peanut, egg, and milk.

The most effective model for each allergen was either an ensemble of RF models for eggs or a Learning Using Concave and Convex Kernels (LUCCK) model for peanut and milk. These models demonstrated notable performance, achieving an impressive Area under the Curve (AUC) of 0.91, 0.96, and 0.94, respectively, in accurately predicting the results

of OFCs for peanut, egg, and milk allergens.

Nedyalkova et al. [48] give a detailed research endeavour that sought to create a proficient computational framework for forecasting the allergenic properties of proteins through the utilisation of diverse descriptors and machine learning algorithms. The research employed a dataset comprising proteins that are allergenic and non-allergenic, and conducted feature selection to determine the most pertinent descriptors.

The performance of classifiers was assessed through cross-validation and diverse performance metrics. The findings indicate that KNN exhibited superior performance compared to alternative classifiers in distinguishing between allergenic and non-allergenic sequences. Conversely, SVM demonstrated better performance when utilising a limited number of descriptors. The study found that the chosen descriptors, which comprised of amino acid composition, evolutionary, and AAindex-based features, had a notable impact on the classification performance. The study showcased the potential of employing computational methodologies for the prediction of protein allergenicity. It also emphasised the significance of feature selection in enhancing the classification model's efficacy.

Pinar et al. [49] introduced a hybrid machine learning approach aimed at predicting SARS-CoV-2 epitopes and identifying antigen peptides suitable for vaccine development that are nonallergenic and nontoxic. The study evaluated various prediction approaches, such as RF, SVM, logistic regression (LR), bagging with DT, KNN, and DT, to forecast epitopes.

To counter the imbalance between samples from the epitope and nonepitope classes, the study employed the Synthetic Minority Oversampling Technique (SMOTE) to rebalance the datasets. Subsequently, epitope prediction was performed again. Comparative analysis of the experimental results revealed that the RF technique provided the most effective predictions. Notably, the performance of epitope prediction was significantly enhanced in balanced datasets compared to the original imbalanced datasets.

Ramisetty et al. [50] devised a computer-assisted framework to diagnose allergies by leveraging data from allergy testing clinics in South India. The framework analyzed intradermal skin test results from 878 individuals, highlighting a lack of samples related to concurrent diseases. To address this data imbalance and improve learning algorithms, advanced data sampling methods were employed. Post-hoc explainable artificial intelligence techniques were integrated to ensure transparency in the machine learning models.

Evaluating the system involved testing a trained random forest model on test data, which achieved an 86.39% overall accuracy rate. Notably, it showed a sensitivity of 75% specifically for the Rhinitis-Urticaria comorbid class. The diagnostic system, integrated into a

mobile application, serves as an educational tool, enabling young practitioners to validate their diagnostic predictions.

Yang et al. [51] concentrates on predicting human-virus protein-protein interactions (PPIs) using a random forest classifier based on doc2vec embeddings. This computational framework outperformed combinations of other widely-used machine learning algorithms and sequence encoding schemes, according to the study. The use of feature embedding for protein representation enabled the acquisition of additional context information from protein sequences, thereby enhancing prediction performance.

The authors anticipate that their findings can be used to identify potential interactions between human and viral proteins and to direct experimental efforts to identify proteins implicated in human-virus interactions and their associated functional functions. The authors hypothesise that future advancements in deep learning architectures, protein structural information, and host PPI network topology can enhance the prediction of human-virus PPIs.

Omurca et al. in [52] designed an intelligent diagnostic assistant for predicting the type of an allergic disease across Turkey automatically by using well-known machine learning algorithms such as Decision Tree, Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbour (kNN), and ensemble classifiers. This was accomplished by using Turkey as a case study. An allergic illnesses dataset, which originates from the Kocaeli University Research and Application Hospital, was utilised in the studies that were conducted. As a consequence of this, the highest accuracy rate of 77% was attained with majority vote when recognising 18 distinct allergy diagnoses.

Joost et al. [53] utilized the Random Forest technique to develop an in silico model for predicting a protein's allergenic potential, leveraging its physicochemical and biochemical traits. The final model incorporates twenty-nine variables derived from the protein sequence using ProtParam software and the PSIPred Protein Sequence Analysis tool. Proteins labeled as allergenic were identified based on their presence in the COMPARE database. The outcomes showcase the model's robust performance, achieving sensitivity, specificity, and accuracy all surpassing 85%. As the model solely depends on protein sequence inputs for computations, it offers seamless integration into existing risk assessment methodologies.

Akbar et al. in [54] proposed an ensemble learning approach for the detection of anticancer peptides using genetic algorithm. The author also gives a detailed mathematical representation of the proposed ensemble technique. The authors have also extended the work in [55, 56] for identifying antitubercular peptides and Neuropeptides by combin-

ing various machine learning techniques like Support Vector Machine, Random Forest, K-Nearest Neighbor etc. using ensemble learning.

A detailed conceptual survey of the above-mentioned machine learning methods is shown in Table 2.1.

Table 2.1: Analysis of machine learning models for allergy diagnosis

Year	**Model Used**	**Contribution**	**Results**	**Future Directions**
2023 [44]	PCC, PCA, RF	Presented machine learning framework for prediction of allergenic proteins	RF performed well in accurately identifying the allergenic proteins with 91% AUC.	More physiochemical properties and feature extraction methods will be adopted for better results.
2023 [45]	QSAR, Random Forest (RF), k-Nearest Neighbour (KNN), Partial Least Squares (PLS), Principal Component (PC)	Developed QSAR method to predict binding affinity	Accurately predicted binding affinity with 0.23 RMSE.	Deep learning methods can be applied for better results.
2023 [46]	Machine learning approach	Machine learning is presented to identify allergen proteins	Machine learning accurately identifies the protein allergens.	Deep learning techniques will be applied in future for better results.
2023 [47]	Support Vector Machine (SVM), LUCCK, RF	Presented ensemble of machine learning methods for food allergen prediction	RF outperformed and predicts accurately with 0.96 AUC value	Genomic and immunologic data will be used to extract additional significant factors. External datasets from multiple healthcare systems will validate the models.

2023 [48]	KNN, SVM	SVM and KNN were applied for the detection of allergenic and non-allergenic sequences	SVM outperformed with0.98 AUC and accurately identified the allergenic and non-allergenic sequences.	Deep learning can be applied for better results.
2022 [49]	RF, SVM, LR, bagging with DT, KNN, DT	Presented machine learning algorithms to predict the epitopes.	The presented approach achieves 94% AUC with SVM.	A vaccine for future epidemics might be designed swiftly using machine learning.
2021 [50]	RF, Decision Tree (DT), SVM, Explainable AI	Applied RF, SVM, and DT, to diagnose allergy using skin data	RF outperformed and achieved 86.36% accuracy in allergy diagnosis.	New feature selection and learning models can be applied for better results.
2020 [51]	RF	Used RF for prediction of human-virus PPIs	RF performed best in the prediction of PPIs accurately.	Deep learning architectures and protein-protein structures will be studied for human virus PPI prediction.
2019 [52]	DT, LR, SVM, KNN, ensemble classifier	ML algorithms were applied to design an intelligent allergy-type diagnostic assistant.	Ensemble learning performed best with 77% accuracy.	City-based asthma allergy type prediction and illness variables will be studied.
2019 [53]	RF	Utilized RF for the prediction of allergenicity.	RF performed well and accurately predicted the allergy with 85% accuracy.	Deep learning algorithm and new feature extraction methods will be explored in future.

2017 [54]	Ensemble learning, Genetic Algorithm	Ensemble of SVM, RF, and KNN is presented for detection of anti-cancer peptides.	Ensemble learning accurately diagnosed the anti-cancer peptides with 94.47% accuracy.	Publically accessible web-based tool will be designed for prediction purposes.

2.1.2 Deep Learning methods for Allergy Detection, Prevention, and Control

Zubaida et al. [57] presented a novel hybrid deep learning model that combined a Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) for accurately classifying peptides as either epitopes or non-epitopes.

Various transfer learning procedures were employed, and the fine-tuned approach yielded the most favorable outcome, achieving an AUC of 0.979, a F1 score of 0.902, and an accuracy of 95.1%, surpassing the performance of the model trained from the beginning. The experimental findings demonstrate that this model has greater performance in comparison to previous approaches. The proposed approach aimed to assist scientists in selecting the optimal epitope for the COVID-19 vaccine, resulting in cost and labor reduction and ultimately expediting vaccine manufacturing.

Zhang et al. in [58] presented a mechanism for screening of antihypertensive peptides using deep learning ProrBERT model. The effective results of the proposed deep learning model suggested that similar deep learning models can be used for related tasks involving peptides.

Liang et al. [40] proposed the utilization of a deep learning model called a transformer, which incorporates a self-attention mechanism. Additionally, they employed ensemble learning models, namely Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost), for the purpose of predicting food allergies.

To emphasize the superiority of the proposed innovative strategy, the study also chose several well used machine learning models as the baseline classifiers. The 5-fold cross-validation results demonstrated that the deep model achieved the largest AUC with a value of 0.9578. This performance surpassed that of the ensemble learning and baseline techniques.

The physiochemical and structural characteristics of allergen proteins originating from plants and animals were examined and compared in [59] using *in-silico* analysis and bioinformatics methods. The study attained an accuracy of 88.24% by analysing the attributes of the allergens and applying the PseAAC concept together with deep learning algorithms for categorization.

In terms of their extinction coefficient and secondary structure, plant allergen proteins showed a more randomly coiled shape than animal allergen proteins, according to the investigation. The research shows the promise of bioinformatics-based methods for comparing allergens and elucidating their characteristics.

Ahmad et al. [60] investigated the predictive capacity of Long Short-Term Memory (LSTM) networks for forecasting food allergies in early childhood (0-3 years) based on longitudinal gut microbiome profiles from participants. Employing the DIABIMMUNE dataset, superior predictive capability was evidenced by the LSTM model compared to the Hidden Markov Model, Multi-Layer Perceptron Neural Network, SVM, RF, and LASSO regression.

Additional analysis assessed whether simplifying representations of microbial characteristics could enhance LSTM network training. Apart from conventional feature selection methods like Minimum Redundancy Maximum Relevance (mRMR) and variance, exploration extended to the use of sparse autoencoder for extracting potential latent representations before LSTM network training. The comprehensive study highlighted that LSTM networks utilizing mRMR-chosen features exhibited notably superior performance compared to the other examined machine learning models.

The analysis of the deep learning model to diagnose allergy is given in Table 2.2.

Table 2.2: Analysis of Deep learning models for allergy diagnosis

Year	**Algorithm**	**Contribution**	**Results**	**Future Directions**
2023 [57]	CNN with LSTM	Hybrid Deep learning is proposed to classify peptides as either epitopes or non-epitopes	Presented approach performed well with an AUC value of 97.9%.	Future research involves the use of transfer learning for better results.
2023 [58]	ProrBERT	Deep learning framework is provided to identify antihypertensive peptides	ProrBERT performed best with an AUC value of 97%	To quickly forecast antihypertensive peptides, a user-friendly web server will be created.
2021 [40]	Transformers with self-attention, XGBoost, LightGBM	Deep learning along with machine learning ensemble is presented to detect food allergens.	Transformers with self-attention performed best with 95% AUC value	Wet lab experiments will be done in the future for better results.
2020 [59]	PseAAC with Deep learning	Deep learning with PseAAC is presented to detect food allergens	The hybrid method performed well and attain an accuracy of 88.24% in identifying the allergens.	Other allergies based on protein data will be analyzed.
2019[60]	LSTM, Multi-layer perceptron Neural Network, SVM, RF, LASSO, mRMR	Deep Learning is proposed to predict food allergies in children	LSTM performed best among all algorithms with 67% AUC	Transfer learning will be employed to improve the performance of proposed work.

2.1.3 Tools used for allergy detection, prevention, and control

Zhenjio et al. [61] developed robust models to predict allergenic proteins or peptides utilizing five distinct-sized pLMs (probabilistic language models). These models underwent thorough evaluation and refinement through convolutional neural networks (CNNs). Notably, the pLM4Alg models, a specific subset, exhibited exceptional performance metrics.

The pLM4Alg models showcased outstanding accuracy, Matthews correlation coefficient, and area under the curve (AUC) scores, ranging between 93.4% to 95.1%, 0.869 to 0.902, and 0.981 to 0.990, respectively. This signifies their high precision, robustness, and discriminative power in predicting allergenic proteins or peptides. Furthermore, a notable feature of pLM4Alg is its capability to handle prediction tasks involving residue-missing sequences and sequences containing nonstandard amino acid residues, making it a pioneering model in this domain.

Muhammad et al. [62] introduced DeepBCE, a web server employing deep learning techniques and sequence-based features for the identification of B-Cell Epitopes (BCE) within proteomic sequences. The convolutional neural network (CNN)-based model showcases superior performance compared to previous models, achieving an accuracy (ACC) of 0.878, an F-measure of 0.871, and an area under the receiver operating characteristic curve (AUC) of 0.945. This suggested technique surpasses prior state-of-the-art approaches by an average of 58.7% according to results based on the Mathews Correlation Coefficient.

Chun et al. [63] introduced DeepAlgPro, a deep neural network model, for the purpose of allergen identification. We demonstrated the exceptional precision and suitability of our tool for making forecasts on a broad scale by a comprehensive comparison with other existing tools. Furthermore, we conducted ablation experiments to showcase the crucial significance of the convolutional module in our model. Additionally, additional investigations revealed that the characteristics of the epitope had a role in the model's decision-making process, hence enhancing the model's interpretability.

Ultimately, we discovered that DeepAlgPro has the ability to identify probable novel allergens. DeepAlgPro is a potent software tool for the detection and identification of allergens.

Wajdi et al. [64] introduced LBCEPred, a Python-based web application designed for B-cell epitope prediction. This tool employs a random forest classifier along with sta-

tistical moment-based descriptors to forecast B-cell epitopes within protein sequences. LBCEPred outperforms existing sequence-based models commonly employed for B-cell epitope prediction, achieving an accuracy of 0.868 and an area under the curve of 0.934. Moreover, the proposed models exhibit an average prediction performance that is 56.3% superior to current state-of-the-art models, as assessed by the Mathews Correlation Coefficient.

LBCEPred stands out as a user-friendly tool suitable for novice users due to its ease of use. Its models have demonstrated both stability and reliability, instilling confidence in its significant potential contribution to the research community and the field of bioinformatics.

Minh et al. [65] introduced AllerCatPro 2.0, a novel web server renowned for its heightened accuracy in predicting protein allergenicity compared to existing computational methodologies. AllerCatPro 2.0 leverages both amino acid sequences and projected 3D structures of input proteins to gauge their similarity against expansive databases housing credible proteins associated with allergenicity. The current dataset incorporated within AllerCatPro 2.0 encompasses 4979 protein allergens, 162 proteins characterized by low allergenic potential, and 165 autoimmune allergens.

This extensive repository enables robust comparative analysis to aid in determining the likelihood of a protein's allergenicity. AllerCatPro 2.0 represents a valuable asset for estimating the potential allergenicity of proteins, offering a versatile and reliable approach for researchers and practitioners in the field.

Francisco et al. [66] introduced a novel approach employing Artificial Intelligence (AI) for the prediction of food allergens. This method integrates the use of the FASTA program to extract distinctive features via pairwise sequence alignment. Moreover, it harnesses Restricted Boltzmann Machines, a Deep Learning technique, in conjunction with the Decision Tree method to execute the prediction process effectively. The resultant tool, named ALLERDET, is publicly accessible and outperforms existing state-of-the-art methodologies.

The methodology adopted demonstrated exceptional performance, boasting a sensitivity of 98.46%, specificity of 94.37%, and an accuracy of 97.26% when assessed using a dataset compiled from various publicly accessible sources. This high-performing approach showcases the efficacy and reliability of ALLERDET in accurately predicting food allergens, positioning it as a significant advancement in this domain.

Robin et al. [67] developed DeepFold, an open-source software that combines spatial restrictions generated by multi-task deep residual neural networks with a knowledge-based

energy function to direct its gradient-descent folding simulations. Large-scale benchmark testing shown that DeepFold outperforms traditional folding algorithms and other prominent deep learning methods in terms of accuracy, producing full-length models.

The modeling performance of DeepFold on challenging targets with few homologous sequences is particularly noteworthy. In these cases, DeepFold earned an average TM-score that was 40.3% higher than trRosetta and 44.9% higher than DMPfold. In addition, the folding simulations conducted for DeepFold exhibited a speed that was 262 times greater than that of typical fragment assembly simulations. The results highlight the efficacy of precisely predicted deep learning potentials in enhancing both the precision and efficiency of ab initio protein structure prediction.

Neelam et al. [68] developed AlgPred 2.0, a web server specifically engineered to predict both allergenic proteins and their specific allergenic regions. This version is an updated iteration of the 2006 AlgPred. The training, testing, and validation datasets encompass 10,075 allergens and an equal number of non-allergens. Additionally, 10,451 experimentally confirmed immunoglobulin E (IgE) epitopes were utilized to pinpoint antigenic regions within proteins.

The methodology integrates several approaches: initial identification of allergens through Basic Local Alignment Search Tool (BLAST) similarity searches against known allergens, followed by the detection of IgE epitopes sourced from the IEDB database within proteins. Furthermore, motif-based techniques such as multiple EM for motif elicitation/motif alignment and search tools were employed for allergen prediction. Diverse machine learning methods were utilized to construct predictive models, culminating in an ensemble approach that aggregates prediction scores from multiple techniques. The top model achieved significant performance, boasting a maximum area under the receiver operating characteristic curve (AUC-ROC) of 0.98 and a Matthew's correlation coefficient of 0.85 when assessed on the validation dataset.

Sharma et al. in [69] used molecular descriptors and machine learning methods to create prediction models for chemical compound allergenicity. The PaDEL programme was used to compute the molecular descriptors for the 403 allergenic and 1074 non-allergic substances that were taken from the IEDB database. On a dataset that included 2D, 3D, and FP descriptors, the models were trained and put to the test. The highest performance came from the hybrid descriptor-based Random Forest-based model, which on the validation dataset had an AUC of 0.93 and a maximum accuracy of 83.39%. It was discovered that several chemical fingerprints, such as GraphFP1014 and PubChemFP129, were more prevalent in allergens.

The study also identified pharmaceuticals that are known to produce allergy symptoms and projected the probable allergenicity of FDA-approved medications. ChAlPred, a web server created for the project, enables the prediction and creation of compounds having allergenic qualities.

Sebastian et al. [70] devised AllerCatPro, a tool leveraging the 3D protein structure and amino acid sequence to forecast the allergenic propensity of proteins. This system evaluates these features against a dataset containing 4180 unique allergenic protein sequences to formulate predictions. AllerCatPro demonstrates an impressive overall accuracy of 84%, surpassing alternative contemporary methodologies which typically range from 51% to 73% in accuracy.

In comparison with the FAO/WHO regulations, AllerCatPro showcases superior sensitivity. Notably, AllerCatPro exhibits a noteworthy enhancement in specificity, boasting a substantial 37-fold improvement. This heightened specificity is a significant advancement, setting AllerCatPro apart as a more precise and reliable tool for discerning the allergenic potential of proteins.

Fabio et al. [71] introduced the SPADE software's web server, which utilizes structural similarities among cross-reactive allergens and immunological cross-reactivity data to differentiate potential IgE-binding sites from non-cross-reactive surface patches. The SPADE server boasts a user-friendly interface, an automated pipeline that integrates third-party and in-house developed routines, and provides a comprehensive output page.

The analysis of tools used for allergy detection, prevention, and control is discussed in Table 2.3.

Table 2.3: Tools used for allergy detection, prevention, and control

Year	**Tool Name**	**Technology Used**	**link**
2023 [61]	pLM4Alg	RStudion, python, Cloud, AWS	`https://f6wxpfd3sh.us-east-1.awsapprunner.com`
2023 [62]	DeepBCE	Python	`http://deeplbcepred.pythonanywhere.com`
2023 [63]	DeepAlgPro	Java, python	`https://github.com/chun-he-316/DeepAlgPro.git`
2022 [64]	LBCEpred	Python	`http://lbcepred.pythonanywhere.com/`

2022 [65]	AllerCatPro 2.0	Java, Cloud	https://allercatpro.bii.a-star.edu.sg.
2022 [66]	ALLERDET	FASTA, R Studio	http://allerdet.frangam.com
2022 [67]	DeepFold	Python	https://zhanggroup.org/DeepFold
2021 [68]	AlgPred 2.0	BLAST, Python	https://webs.iiitd.edu.in/raghava/algpred2/
2021 [69]	Chalpred	Responsive html template	https://webs.iiitd.edu.in/raghava/chalpred/
2019 [70]	AllerCatPro	Java, FASTA	https://allercatpro.bii.a-star.edu.sg/
2019 [71]	SPADE	Own Software tools	https://spade.uni-graz.at

2.2 Problem Statement

In a self-learning system, machine learning is used for optimisation, prediction, and drawing inferences. In order to guide future choices, it makes use of past information. How well an ML system learns determines how effective it will be. Machine learning (ML) methods allow for the study and tracking of numerous diseases in computational biology. Medical conditions can cause changes in vital signs, and an ML system can learn from these variations as long as they are within the normal range. This idea is also used in genomic analysis and computer-aided drug design to predict a number of medication-related factors from the actions of individual components.

It is critical to ensure that dietary supplements and pharmaceuticals are not allergic and harmful when they are being developed. Allergens, which are substances or components of substances that can induce allergic reactions in humans, may also be a component of substances that target specific diseases. The kind of allergic reaction that an allergen can cause determines how it is classified. Making sure there are no allergies in the ingredients is a top priority in both medicine development and food production. By analysing the sequence of an allergen, computational methods can determine its type. One way to identify allergies is by looking at the protein sequence and its physicochemical characteristics. These properties can be used to construct a variety of ML models. Once several models have been trained, they can be combined to form an ensemble model. In most cases, an ensemble model will produce more accurate predictions than a system using a single model.

2.3 Research Gaps Identified

Based on the literature survey, the following research gaps are identified.

- An efficient ensemble learning mechanism for allergen detection, which enhances the performance of computationally intelligent model by combining various machine learning models using ensemble learning approach, is required.
- A dense neural network architecture needs to be formulated for effective classification of Allergens.
- The combination of ensemble learning together with the deep learning approach can further enhance the performance of the computationally intelligent mechanism for allergen detection.
- A stochastic mathematical model together with the artificial neural networks is required to be formulated for allergen detection and the prediction of allergy prevalence.

2.4 Research Objectives

Based on the literature review and research gaps, the following objectives are delineated:

1. To collect protein allergen sequences and extract physicochemical properties.
2. To design and implement a computationally intelligent framework for the classification of sequences into allergen or non allergen and to predict the type of allergen.
3. To test and validate the performance of proposed framework on different benchmark datasets and unknown sequences.

Different ML and DL models are ensembled to fulfill the above mentioned objectives. A stochastic mathematical model, together with artificial neural networks, is also proposed to solve the non-linear differential equation for allergy prevalence prediction. The research methodology adopted to formulate these models is discussed in the next chapter.

Chapter 3

Research Methodology

This chapter discusses the methodology to achieve the research objectives of the thesis. The methodology is divided into two parts, i.e. requirement specifications and computational intelligent approaches, as shown in Figure 3.1. Starting from data collection to model building, requirement specifications include a detailed study of various Python programming Language libraries. In ML-based approach, various models were built separately based on ML and DL algorithms and combined as an ensemble model. Stochastic approaches include the solution of allergy-based differential equations by mathematical approach and by using artificial neural network(ANN).

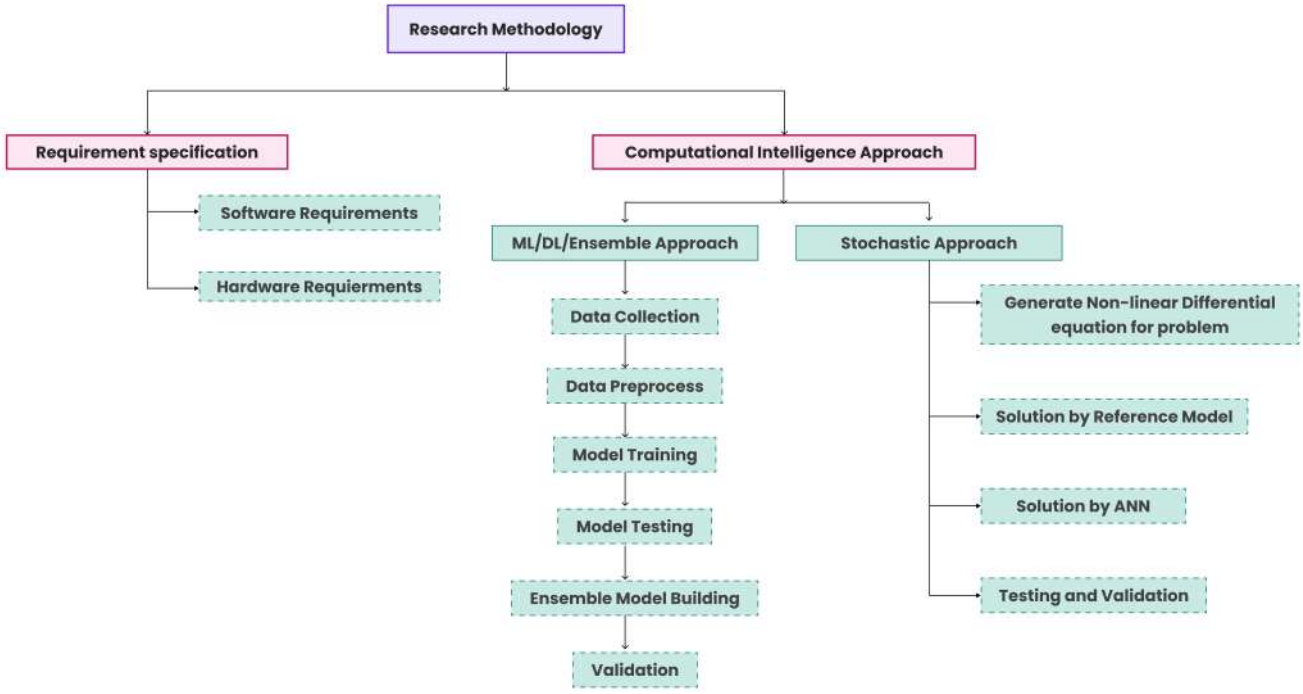


Figure 3.1: Flow of Research Methodology to attain the proposed objectives

3.1 Requirement Specifications

In order to fulfil the stated objectives of the thesis, it is significant to comprehend the necessary specifications for both the hardware and software components. The specifications

are described for the same purpose.

3.1.1 Software Requirements

The following tools and libraries available in various programming languages are used:

- **Peptide package from R:** The package provides a range of functions for computing diverse physicochemical parameters and indexes for amino acid sequences. Understanding the features and behaviours of proteins at a molecular level is highly dependent on this competence.
- **Python Libraries:** Several Python open-source libraries were utilized to train, test and validate the model. They are described as follows:
  - **Pandas:** Commonly utilized for the purpose of data retrieval and initial data manipulation. It provides data structures and functions for manipulating numerical tables and time series.
  - **NumPy:** Crucial for mathematical calculations. It offers a wide range of high-level mathematical functions and support for arrays and matrices.
  - **Matplotlib:** It is utilized for the purpose of data visualisation, which is of utmost importance in comprehending data distributions, trends, and model performance.
  - **Scikit-learn:**This library provides a wide range of methods for machine learning, making it adaptable and flexible. It includes methods for validating models, as well as algorithms for classification, regression, and clustering.
  - **TensorFlow:** TensorFlow is a Google-developed open-source library for deep learning and machine learning. TensorFlow is renowned for its versatility and ability to handle intricate calculations. It is extensively utilized for jobs such as image and speech recognition, as well as natural language processing.
  - **Keras:** An advanced neural networks Application Programming Interface (API) specifically created to facilitate and expedite the process of developing prototypes. It functions as a platform for TensorFlow, streamlining the process of creating deep learning models with its intuitive and modular design.
- **MATLAB Neural Fitting Tool (nftool)** MATLAB's Neural Network Toolbox includes the Neural Fitting, which provides a user-friendly and interactive environment for constructing, training, and simulating neural networks designed for function fitting problems. It streamlines the process of creating neural networks to

mimic intricate nonlinear connections, which is especially valuable in modelling and forecasting data trends. The tool provides structured data entry, network construction, and training processes, making it easily accessible to users of all skill levels in neural network applications.

3.1.2 Hardware Requirements

System's Configuration: The minimum configuration of the system required for the implementation of the research work presented in the thesis is described as:

- Processor: Intel® Core™ i5 processor.
- Memory: 4 GB of RAM, 8 GB recommended.
- Hard Disk: 20 GB of hard disk space required, 10 GB additional hard disk space required for installation.
- Display: 1024 X 768 or higher-resolution display with 24-bit colors.

3.2 Computational Intelligent Approaches

The intelligence of computers, popularly known as Artificial Intelligence, is the field of study where computer machine are trained to learn and act like humans. The whole process revolves around the experience in which the learning is to take place. The process of building the computationally intelligent framework to attain the objectives is discussed as follows:

- **Data Collection:** The dataset used for protein allergen classification consisted of protein sequences collected from various sources. The allergen protein sequences were accessed from the AllergenOnline database, which is a comprehensive database of allergenic proteins. The non-allergen protein sequences were collected from the National Centre for Biotechnology Information, UniProt, and SWISS-PROT databases, which contain a vast collection of protein sequences. The protein sequences were processed for feature extraction, which involved extracting various physical, chemical, and other properties of the protein sequences.
- **Data Pre-processing:** Prior to processing the dataset, it was made certain that all of the feature values were numeric and that they fell inside a specific range. utilizing the labelled encoding method, any values that were missing were substituted with the mean of the feature vector that corresponded to them, and values that were not numeric were converted into numeric values.

- **Features Selection:** From the dataset of protein sequences, the feature selection procedure tried to identify the descriptors that were the most relevant to the problem statement. During this procedure, it was necessary to determine which characteristics had the most important influence on the classification performance of the machine learning models. Feature selection was utilized in order to improve the effectiveness of the classification model and to expand its capacity to differentiate between allergenic sequences and sequences that do not cause allergic reactions.
- **Machine learning Model Building:** ML is a branch of artificial intelligence that focuses on creating algorithms and statistical models. These models allow computer systems to enhance their performance on a particular task by learning from experience or training on data. To identify allergens in a dataset, protein allergen classification uses machine learning methods to analyse and classify protein sequences according to their characteristics and physicochemical properties.
- **Deep Learning Model Building:** Deep learning is a sub-field of machine learning that utilises artificial neural networks with several layers to obtain data representations. Neural networks possess the ability to autonomously acquire the capacity to depict data with numerous degrees of conceptualization. This empowers them to execute complex undertakings like identifying sounds and images, processing natural language, and other similar activities. Deep learning techniques were employed in the field of protein allergen classification to extract intricate patterns and characteristics from protein sequences, enabling accurate identification of allergens.
- **Ensemble Model Building:** Ensemble learning is a technique in machine learning that integrates many models to improve the accuracy and resilience of the overall prediction. Ensemble learning, through the utilisation of multiple models and the combination of their predictions, can frequently outperform individual models and reduce the likelihood of overfitting. Within the field of protein allergen classification, this study employs an ensemble learning model that integrates the results of many machine learning and deep learning models, including Extra Trees, Deep Belief Network (DBN), and CatBoost. By utilizing the combined strengths of several varied models, this strategy improves the overall precision and consistency of allergen identification.
- **Stochastic Mathematical Model Building:** A predictive model for the incidence of allergen sensitization was built by utilizing a system of nonlinear differential equations. The system was categorised into four dynamics based on the human populations, with the final two groups representing the allergens. The system's solution was numerically represented using the stochastic computational artificial

neural network (ANN) and the Levenberg-Marquardt backpropagation (LMBP) technique. The stochastic technique was implemented using the 'nftool', a solver in MATLAB. The Adams scheme's aptitude was utilized to construct networks that generate estimated outcomes for the system, which were subsequently utilized for testing, certification, and training purposes. The ability of the scheme to solve the allergen model was shown by the findings consistently overlapping and the presence of strong and reducible absolute error (AE).

The correctness of the stochastic scheme was evaluated by comparing its performance with the proposed and reference solutions. A minimal, absolute error served as evidence of the scheme's accuracy. The study examined the dependability of the stochastic solver by employing different statistical operators.

In this study, the methodology employed entails constructing a predictive model for the occurrence of allergen sensitization. This is achieved by utilizing a system of nonlinear differential equations and assessing its precision through the application of the stochastic computational artificial neural network (ANN) and the Levenberg-Marquardt backpropagation (LMBP) technique.

- **Testing and Validation:** Testing and validation are essential stages in the development of machine learning models, including those employed for protein allergen classification. Testing and validation entail assessing the model's performance on a distinct dataset that was not utilized during the training phase. This aids in evaluating the model's capacity to generalise and reliably forecast the class labels of novel, unseen data.

  The testing and validation step is essential to verify that the model is not too conforming to the training data and can effectively forecast the class labels of the novel, unseen data. It aids in identifying any deficiencies in the model and offers insights into areas that may necessitate enhancement. The testing and validation procedure assesses the model's performance on a distinct dataset, ensuring its robustness and ability to generalise effectively to new data.

  Protein allergen classification often involves dividing the dataset into two subsets: a training set and a testing set. The training set is utilized to instruct the model, whereas the testing set is employed to assess the model's performance. The testing set should be a true reflection of the entire dataset and should include an equal proportion of protein sequences that contain allergens and those that do not.

  During the testing and validation step, the model is evaluated on the testing set, and the predicted class labels are compared to the actual class labels. The model's

performance is subsequently assessed using a range of measures, including accuracy, precision, recall, F1-score, and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

3.3 Summary

This chapter provides an overview of the hardware, software, tools, and methodologies utilized. In addition, a comprehensive analysis was conducted on the progression of machine learning (ML), deep learning (DL), and ensemble methods, with each stage being thoroughly examined. In this chapter, we provide a comprehensive explanation of the stochastic mathematical model, including its implementation specifics. Additionally, several performance criteria were offered to evaluate the performance during training, testing, and validation.

Chapter 4

A Deep Learning based Ensemble Approach for Protein Allergen Classification

This chapter presents a deep learning based ensemble approach to identify protein allergens using Extra Tree, Deep Belief Network (DBN), and CatBoost models. The proposed ensemble model achieved higher detection accuracy by combining the prediction results of the three models using majority voting. The evaluation of the proposed model was carried out on the benchmark protein allergen dataset, and the performance analysis revealed that the proposed model outperforms the other state-of-the-art literature techniques with a protein allergen detection accuracy of 89.16%.

4.1 Overview

The prevalence of allergic reactions is a serious public health problem that may cause diseases like high fever, rhinitis, asthma, dermatitis etc., and affects a sizeable fraction of the world's population[36]. Currently, the methods that can be used to cure allergies are not completely understood, and the only strategy that is known for preventing allergies is to abstain from substances that contain allergens [72]. However, the inaccurate detection of allergens may result in excessive dietary restrictions, which can then lead to nutritional issues [73]. Hence, it is clear that an efficient mechanism which can amalgamate multiple parameters together for accurate allergen detection is necessary for allergy control. The computational methods combined with bioinformatics have the ability to analyze multiple characteristics together for better allergen detection accuracy [74].

Protein allergens are a key contributor to the development of allergic reactions; hence, locating and describing these allergens is essential to the research and development of efficient diagnostic and therapeutic methods [75]. Bioinformatics techniques may be used to conduct an analysis of protein allergen sequences in order to locate potentially allergenic areas because they contain carefully curated information on the sequences and structures of protein allergens [45]. However, traditional techniques of allergen prediction take a lot of time and frequently depend on methods of experimentation that require a lot of manual resources [76].

Machine learning and deep learning algorithms have only very recently emerged as potentially useful techniques for allergen prediction and classification [53, 77]. These algorithms are meant to analyse enormous volumes of data in order to recognise patterns that human analysts, depending on their level of experience, can miss. These intelligent algorithms have been utilised in a number of studies for the purpose of allergen prediction and categorization.

Individuals who are afflicted by allergies may benefit from a better diagnosis and treatment if researchers and clinicians include intelligent mechanisms in their work. So, in this paper, we have described various essential factors/properties of protein sequences that can be used for the analysis and detection of different types of allergens.

The researchers have worked on various machine learning techniques for allergy detection and control. However, this paper proposes a novel combination of ensemble learning, machine learning, and deep learning approaches for enhancing overall performance. Therefore, in this research, a deep learning based ensemble of three models: extra trees, DBN, and catboost has been proposed for the detection of protein allergens. The performance has been analyzed using different parameters like accuracy, F1-score, False Alarm Rate, Specificity, and Matthews Correlation Coefficient (MCC) etc.

4.2 Proposed Mechanism for Allergen Detection

In this section, we describe the details of the proposed model for the detection of protein allergens. The literature suggests that combining the outcomes of different machine learning models using ensemble learning techniques can enhance the overall accuracy of predictions. This work uses this concept and, hence, proposes an ensemble learning model to detect protein allergens in a given dataset. The system model for the proposed mechanism is shown in Figure 4.1.

In the initial phase, we processed each protein sequence collected from different sources. The allergen protein sequences were accessed from AllergenOnline [78], and non-allergen protein sequences were collected from National Center for Biotechnology Information [79], UniProt [80] and SWISS-PROT [81]. These protein sequences were then processed for extracting features, that could be used to train different machine learning and deep learning models. In Figure 4.1, we have mentioned a few protein sequences in the "Protein Sequence" block. The features were extracted using the Peptide package available on CRAN. The physical, chemical and other properties, as described in Table-4.1, were extracted from each protein sequence.

Table 4.1: Description of Dataset Features

Feature Descriptor	**Category of Descriptor**	**Method used for extraction of feature**	**Description**
Amino acid indices	Physicochemical properties	aIndex(sequence)	The indices of amino acids are numerical values that stand for the physicochemical properties of certain amino acids. Some examples of these indices include hydrophobicity, polarity, and charge [82].
Hydrophobicity scales	Physicochemical properties	Hydrophobicity (sequence, scale)	The hydrophobicity of amino acids is quantified using hydrophobicity scales. Optional prefix indicating whether hydrophobicity scale (KD, H, HW, MF, SH, or WS) is to be employed [83].
kideraFactors	Physicochemical properties	kideraFactors (sequence)	Amino acids are assigned numerical values called kidera factors that reflect their physicochemical traits. It returns the average of 10 kidera factors [84].
Amino acid property scales	Physicochemical properties	stScales (sequence)	Amino acid property scales include ALIPHATIC, AROMATIC, BASIC, CYSTEINE, HYDROXYL, NEGATIVE, POSITIVE, SMALL, and TINY [85]. ALIPHATIC rates alanine, valine, leucine, and isoleucine by aliphatic side chain size and shape. The AROMATIC scale ranks phenylalanine, tyrosine, and tryptophan based on their side chain aromatic rings. Because of their basic amino groups, lysine, arginine, and histidine have greater BASIC scores. All amino acids except cysteine are 0. Whether serine and threonine have hydroxyl groups (-OH) on their side chains affects HYDROXYL values. Aspartic acid and glutamic acid have negative ratings because their side chains include carboxyl groups. If lysine, arginine, or histidine possesses a positively charged amino group in its side chain, values are positive. SMALL scale values glycine, alanine, serine, threonine, and cysteine greater. Glycine and alanine, small amino acids, provide this scale's eighth value.
VHSE scales	Physicochemical properties	vhseScales (sequence)	VHSE-scales (principal components score Vectors of Hydrophobic, Steric, and Electronic properties), are included in a total of 50 physicochemical variables of 20 coded amino acids and are derived from principal components analysis (PCA) on independent families of 18 hydrophobic properties, 17 steric properties, and 15 electronic properties, respectively [86].
Z scales	Physicochemical properties	zScales (sequence)	The physicochemical characteristics of the amino acids, such as NMR data and Thin-Layer chromatography (TLC) data, serve as the foundation for the Z-scales. Its value range from z1 to z4, reprsenting Lipophilicity, Steric, Electronic and other properties. It calculats mean of the Z-scales for each of the amino acids that make up the relevant peptide sequence [87].

Feature Descriptor	Category of Descriptor	Method used for extraction of feature	Description
Auto-correlation coefficients	Molecular descriptors	autoCorrelation (sequence, lag, property)	The autoCorrelation index is computed for a lag 'd' across a sequence of length 'L' using a descriptor 'f' (centred) [88].
Autocovariance coefficients	Molecular descriptors	autoCovariance (sequence, lag, property)	The autoCovariance index is computed for a lag 'd' across a sequence of length 'L' using a descriptor 'f' (centred) [88].
Cross-covariance coefficients	Molecular descriptors	crossCovariance (sequence, lag, property1, property2)	For a sequence of length 'L', the lagged crossCovariance index is computed for a lag 'd' using two descriptors, 'f1' and 'f2' [88].
Cruciani molecular descriptors	Molecular descriptors	crucianiProperties (sequence)	Using the scaled principal component scores that summarize a wide range of descriptors computed from the interaction of each amino acid residue with several chemical groups (or 'probes') like charged ions, methyl, hydroxyl groups, and so on, this function determines the Cruciani properties of an amino-acids sequence, which include: Polarity, Hydrophobicity, and Hydrogen Bonding [88].
WHIM molecular descriptors	Molecular descriptors	mswhimScores (sequence)	Twenty of the naturally occurring amino acids were used to generate 36 electrostatic potential attributes used to calculate TMS-WHIM scores [89].
BLOSUM matrix scores	Sequence similarity	blosumIndices (sequence)	BLOSUM indices were created by employing an alignment matrix of the 20 natural amino acids and physicochemical attributes that had undergone VARIMAX investigations [90].
Boman index	Antimicrobial activity	boman (sequence)	This function calculates Boman's protein interaction index from a protein's amino acid sequence. The index is the total of the solubility values for all residues in a sequence and can indicate a peptide's capacity to attach to membranes or other proteins as receptors. This function predicts peptide-protein interactions [91].
Net charge at a given pH	Charge	charge (sequence, pH = 7, pKscale)	Using the Henderson-Hasselbalch equation presented by D. S. Moore, this function determines the net charge of a protein sequence. There are 9 different pKa scales that may be used to determine the net charge at a given pH, including Bjellqvist, Dawson, EMBOSS, Lehninger, Murray, Rodwell, Sillero, Solomon, and Stryer [92, 93, 94, 95, 96, 97, 98, 99, 100].
Isoelectric point	charge	pI (sequence, pKscale)	When a certain molecule or surface has zero net electrical charge, the pH is said to be at its isoelectric point (pI). It is a factor that influences the solubility of the peptides at specific pH levels. A protein's biological activity is typically lost when the pH of the solvent is the same as the protein's pI pKscale – A string indicating which pK scale is to be used; valid values are 'Bjellqvist', 'EMBOSS', 'Murray', 'Sillero', 'Solomon', 'Stryer', 'Lehninger', 'Dawson', and 'Rodwell'. [92, 93, 94, 95, 96, 97, 98, 99, 100].

Feature Descriptor	**Category of Descriptor**	**Method used for extraction of feature**	**Description**
Peptide length	length	lengthpep (seq)	The number of amino acids in a given protein sequence may be determined by using this function. Amino acids create long chains that are linked together by chemical compounds called peptide bonds to produce proteins [84].
Mass shift due to modifications	Mass spectrometry	massShift (sequence, label , aaShift, monoisotopic)	Chemical changes or heavy isotope labelling change peptide mass. Function arguments: label – Label heavy isotopes. Accepts "none", "silac_13c", "silac_13c15n", and "15n". Replaces aaShift input. aaShift – Name the amino acid mass differential in Dalton vector. Names and values are amino acid one-letter codes and Dalton mass shifts. 'Nterm' and 'Cterm' define N- and C-terminal modifications. monoisotopic: A logical value 'TRUE' or 'FALSE' indicating whether monoisotopic weights of amino acids should be utilised [101].
Molecular weight	Mass spectrometry	mw (seq, monoisotopic, avgScale, label , aaShift)	It estimates protein sequence molecular weight. The scale on Compute pI/Mw tool calculates it as the total of each amino acid's mass [101]. It calculates protein masses using predefined or custom stable isotope mass labels.
Mass-to-charge ratio	Mass spectrometry	mz (seq, charge, label, aaShift, cysteins)	Peptide mass-to-charge ratios (m/z) from mass spectrometry are determined using this function [101].
Protein fingerprint	Structural features	protFP (seq)	Descriptors for the 20 naturally occurring amino acids were drawn from the AAindex database to form the basis of the ProtFP descriptor set [102].

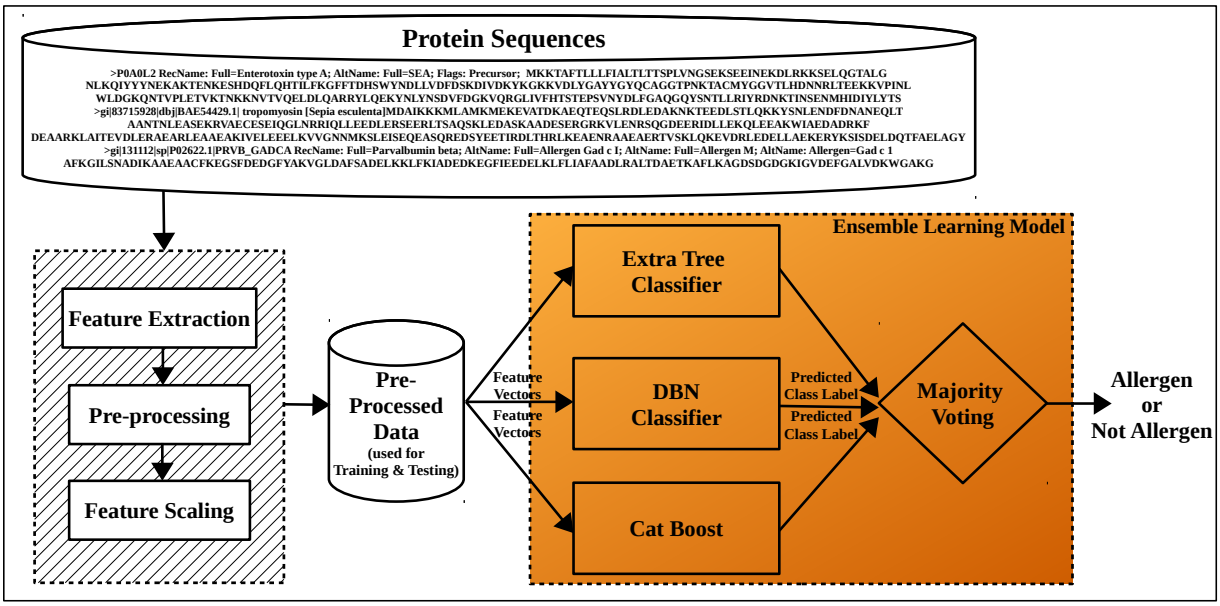


Figure 4.1: System Model

The performance of any machine learning algorithm depends on the values contained in the dataset, and the prediction results are effective if all the feature values are numeric and lie in a particular range. In order to achieve this, we have replaced any missing value with the mean of the corresponding feature vector and encoded non-numeric values to numeric values using the labelled encoding technique.

The updated dataset was then passed to the feature scaling phase, where we used a standard scalar to transform all the values of feature vectors. If x is the feature vector and $\bar{x}$ and σ are the mean and standard deviation of the feature vector, then the transformed feature vector $\hat{x}$ is calculated using standard scalar as shown in equation-4.1.

$$\hat{x} = \frac{(x - \bar{x})}{\sigma} \tag{4.1}$$

The completion of feature extraction, pre-processing and feature scaling results in a pre-processed dataset containing 128 features and 4854 entries (among which 2427 represent allergens and the rest represent non-allergens), which can then be used for training and testing the machine learning models.

We have selected three models for constructing the proposed ensemble learning model for detecting protein allergens. The selection was based on a thorough performance analysis of different machine learning and deep learning models on the dataset.

The proposed ensemble learning model contains a combination of Extra Tree, Deep Belief

Network (DBN) and Cat Boost models, the details of which are given as:

- **Extra Tree:** Decision tree algorithms are the basis of the Extra Trees (Extremely Randomised Trees) classifier, another ensemble learning approach. Training numerous decision trees and aggregating their predictions results in a more accurate and robust categorization. The Extra Trees classifier adds more randomness than the standard Random Forest algorithm. This creates more diverse trees and better generalisation performance [103].

  The Extra Trees classifier generates a collection of decision trees, each of which is constructed using a different subset of the training data and a different set of characteristics at each node. Due to the algorithm's lack of optimisation of the splitting criterion, it is both quicker and more random than conventional decision tree approaches like Random Forest, from which it borrows its name. By adding some randomness to the tree-building process, we may generate a more varied set of trees, which helps lower the model's variance and boosts its generalisation performance. Rather of relying on just one tree's predictions, the classifier aggregates the results from all of the trees in the ensemble, typically by majority vote, to provide more reliable and accurate classifications.

- **DBN**: To quickly and effectively create hierarchical representations of data, a DBN blends unsupervised and supervised learning approaches. DBNs are constructed from a number of layered Restricted Boltzmann Machines (RBMs). They are taught greedily, layer after layer, using unsupervised pre-training to set their starting weights and then supervised learning to fine-tune them. DBNs have been successful in a number of contexts, including those involving image and speech recognition, NLP, and dimensionality reduction [104].

- **Cat Boost:** A gradient boosting technique designed specifically for categorical features is called CatBoost. In addition to providing effective management of big datasets containing categorical variables, it is optimised for high-performance prediction [105].

  Agradient-boostingg technique designed specifically for categorical features is called CatBoost. In addition to providing effective management of large datasets containing categorical variables, it is optimised for high-performance prediction. It uses ordered boosting to reduce target leakage and overfitting. This method reorders the training examples using a random permutation and then uses the examples that came before them in the permutation to figure out the target statistics for each category.

As discussed earlier in this section, we have selected three models based on the performance analysis of various machine learning and deep learning models, to create an ensemble model. The other machine learning models used in the study include Logistic Regression (with maximum iterations of 300 on L2 regularization and limited-memory BFGS optimizer), k-Nearest Neighbour (with $k = 10$), Support Vector Machines (with three kernel variants: Linear, polynomial of degree 3 and Radial Basis Function), Random Forest (with size of base estimator as 100 and criterion for node splitting as Gini index), Ada Boost, Gradient boosting and XG Boost [106].

The deep learning models include: Multi-layer perceptron and DBN, the architecture of which is shown in Figure 4.2. The number of neurons in the input layer of each deep learning model equals the number of available features. The first hidden layer has the same number of neurons as that of the input layer, but the second hidden layer has 50% of neurons in the first hidden layer (i.e., 64 neurons). The third hidden layer has 13 neurons that are responsible for pattern recognition for major feature descriptors which include Hydrophobicity, kideraFactors, Amino acid property scales, VHSE scales, Z scales, Cruciani molecular descriptors, WHIM molecular descriptors, BLOSUM matrix scores, Boman index, charge, Molecular weight, Peptide length and Protein fingerprint.

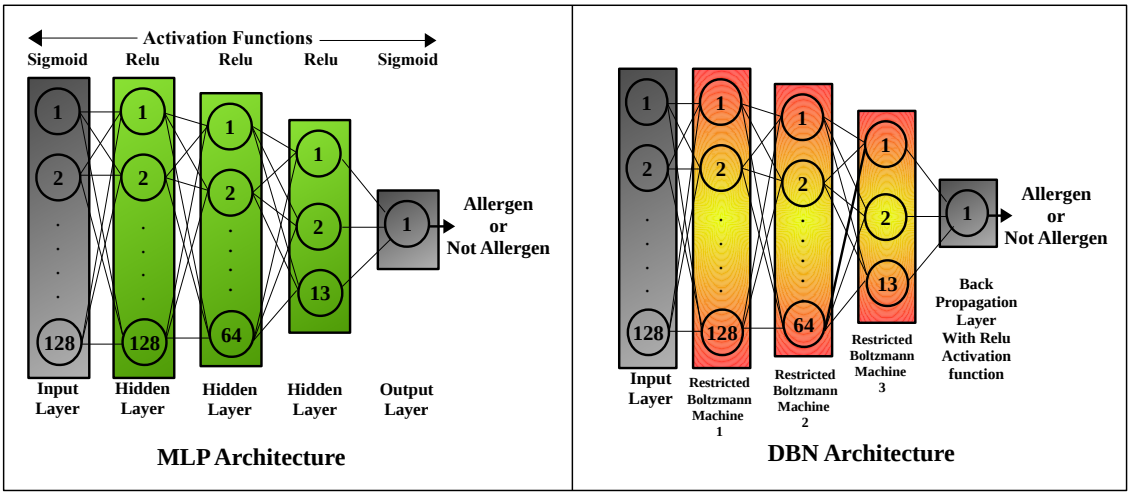


Figure 4.2: Architecture of Neural Networks used in the study

Each of the models in the proposed ensemble learning approach receives feature values concurrently and predicts the class label (allergen or not). The prediction results of the base models are then combined using majority voting, the output of which gives the final predicted class label for the input values. In majority voting, the class label that receives the most votes becomes the ensemble's prediction for that instance. We have used majority voting because it lowers the likelihood of overfitting, boosts model stability,

and increases prediction accuracy [107].

4.3 Results and Discussion

The performance analysis of the proposed model and other machine learning and deep learning models on the dataset was implemented using python 3.10 with scikit-learn, PyTorch and other related packages on 1.6 GHz Intel Core i5 with 12 GB of RAM. The dataset was divided into training and testing sets by a random selection of feature values. The training and testing datasets contained 3347 and 1457, respectively, randomly selected entries.

4.3.1 Performance Parameters

In order to evaluate the performance of each learning model, we have trained each model on the training dataset and tested it on the testing dataset. The test results were calculated using the confusion matrix (as shown in Table-4.2).

Table 4.2: Confusion Matrix

Actual Value	**Predicted Value**	
	Allergen	**Non-Allergen**
Allergen	True Positive (TP)	False Negative (FN)
Non-Allergen	False Positive (FP)	True Negative (TN)

Once the confusion matrix for each model on the testing dataset was evaluated, then the below-mentioned parameters were calculated to further analyze the performance:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.2}$$

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} \tag{4.3}$$

$$FalseAlarmRate = \frac{FP}{TN + FP} \tag{4.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN + FN)}} \tag{4.6}$$

Apart from the above-mentioned parameters, we have also plotted the Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve and calculated the Area Under the Curves (AUC). The training and testing time was also noted during the analysis.

4.3.2 Result Analysis

In the study, we have also used deep learning models (MLP and DBN) for allergen classification. The loss that occurred during the training of these models with varying epochs is shown in Figure 4.3. The figure shows that as the number of epochs increases, the training loss for both the models decreases and attains a minimum value of 5% (approx.). This reduced value of training loss indicates the effective training of deep learning models.

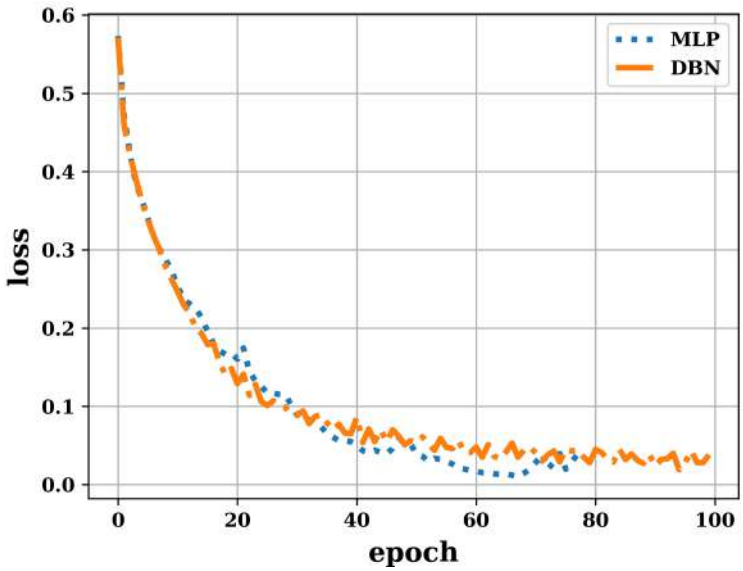


Figure 4.3: Training Loss for Deep Learning Models

The test results of various machine learning, deep learning and proposed model are presented in table-4.3 in the form of True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP), Training Time and Testing Time (in Seconds).

The lower values of False Positives and False Negatives for DBN, Extra Tree and Cat Boost models, as shown in Table-4.3, resulted into the formation of proposed ensemble model, with an objective to further enhance the overall classification performance. The proposed model has reduced the number of FP and FN, as show in the table. The

Table 4.3: Classification Results of Various Learning Models on the Test Dataset

Model	**TN**	**FP**	**FN**	**TP**
Logistic Regression (LR)	553	158	148	598
k-NN	590	121	85	661
Linear SVM (LS)	558	153	148	598
Polynomial SVM (PS)	609	102	176	570
SVM with RBF (RS)	588	123	109	637
Random Forest (RF)	624	87	97	649
Adaboost (AB)	586	125	116	630
Gradient Boosting (GB)	595	116	105	641
MLP	664	47	180	566
DBN	622	100	84	651
Extra Tree (ET)	627	84	102	644
XG Boost (XB)	619	92	85	661
Cat Boost (CB)	622	89	83	663
Proposed Model	634	77	81	665

test results of each model are then used to compute the value of other performance parameters.

Figure 4.4 shows the value of allergen classification accuracy of different models studied on the test dataset. Accuracy offers a gauge of how effectively the model can generate accurate assumptions about fresh and unseen data. A high accuracy rating means the model can generalize effectively and make precise assumptions about new data. The figure shows that the proposed model achieves maximum accuracy when compared with other models. The false alarm rate of each learning model is shown in Figure 4.5. The figure shows that the proposed model has least false alarm rate of 10.83% than the other studied models (except MLP).

The F1 score is a statistic that integrates recall and precision into one performance parameter for the model. It offers a fair assessment of the model's capacity to recognize both positive and negative instances. When the cost of false positives and false negatives is not equal, this is especially crucial. The F1-score of various studied models is shown in Figure 4.6. It is evident from the figure that the proposed model has an F1-score of 89%, which is the maximum when compared with other studied models.

Specificity quantifies the percentage of true negatives that the model properly detects, i.e., it tells us how well the model is able to detect undesirable circumstances. Figure 4.7

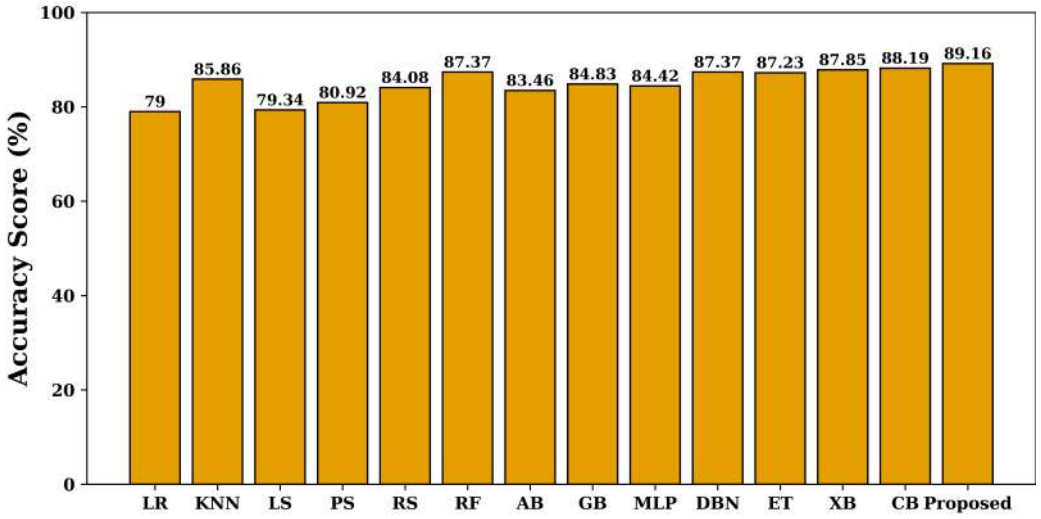


Figure 4.4: Allergen Classification Accuracy of various models

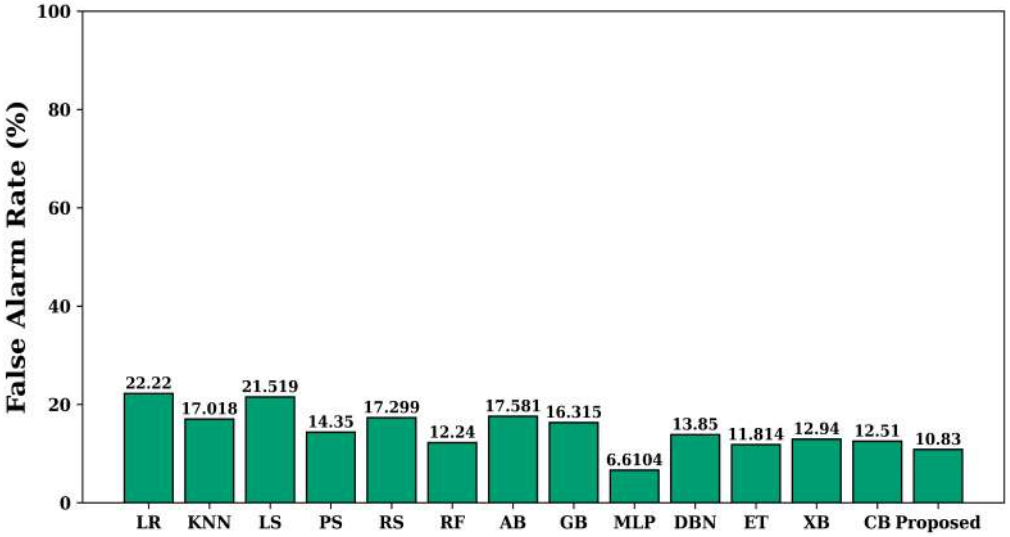


Figure 4.5: False Alarm Rate of various models

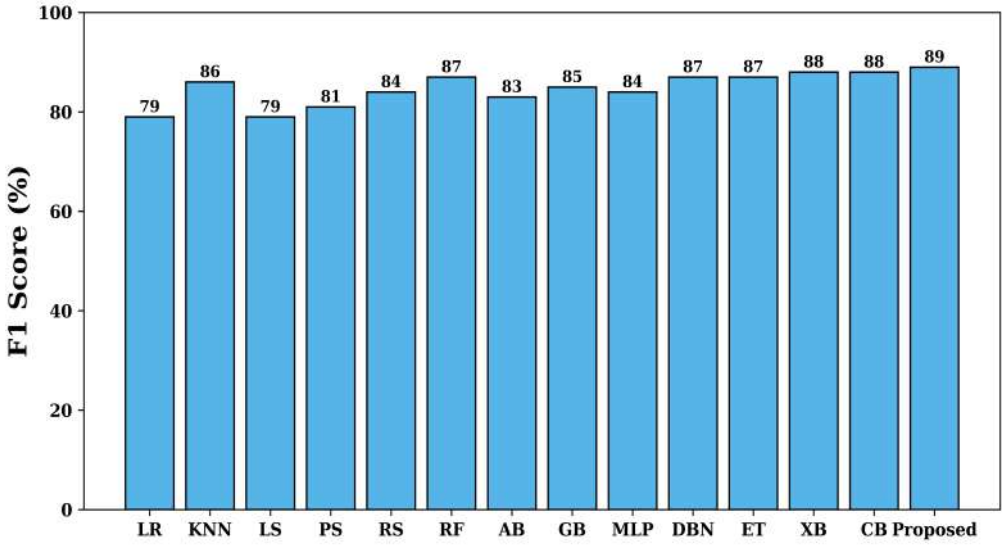


Figure 4.6: F1-Score of various models

shows this specificity value for each studied model. Here again, the proposed model has a maximum specificity value of 89.17% than the other studied models (except MLP). The MLP model has a maximum specificity value due to the lower value of the false alarm rate. Besides MLP, the proposed model has a maximum value of specificity than the other studied models.

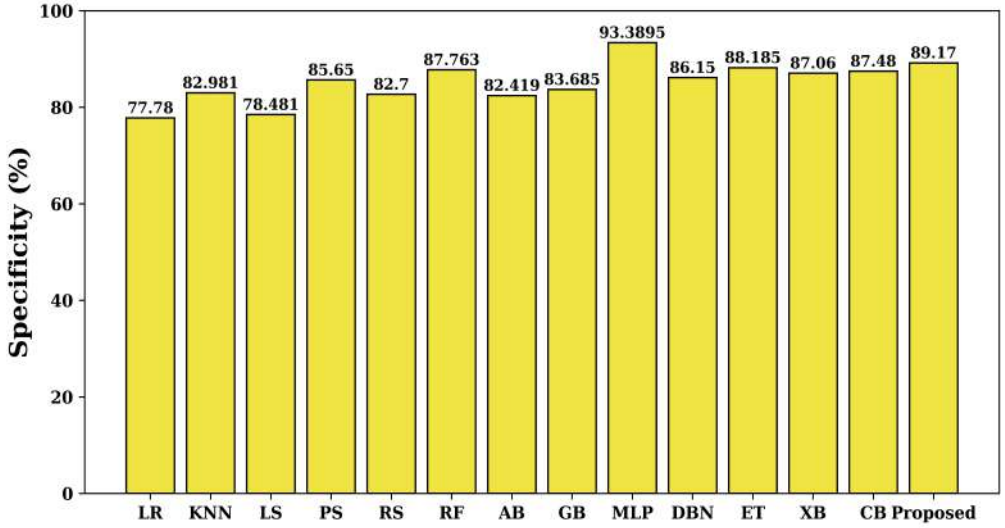


Figure 4.7: Specificity of various models

The significance of the Matthews Correlation Coefficient (MCC) is that it does not depend on unbalanced datasets and gives a fair assessment of the model's performance, accounting for both true positives and true negatives. The MCC runs from -1 to +1, with +1 denoting

perfect agreement, 0 denoting predictions that are no better than random, and -1 denoting utter disagreement between the predicted and the actual values. The MCC value of each model on the test dataset is shown in Figure 4.8. Here again, the proposed model has a maximum MCC value of 78.3%, which shows the effectiveness of the proposed model for allergen classification.

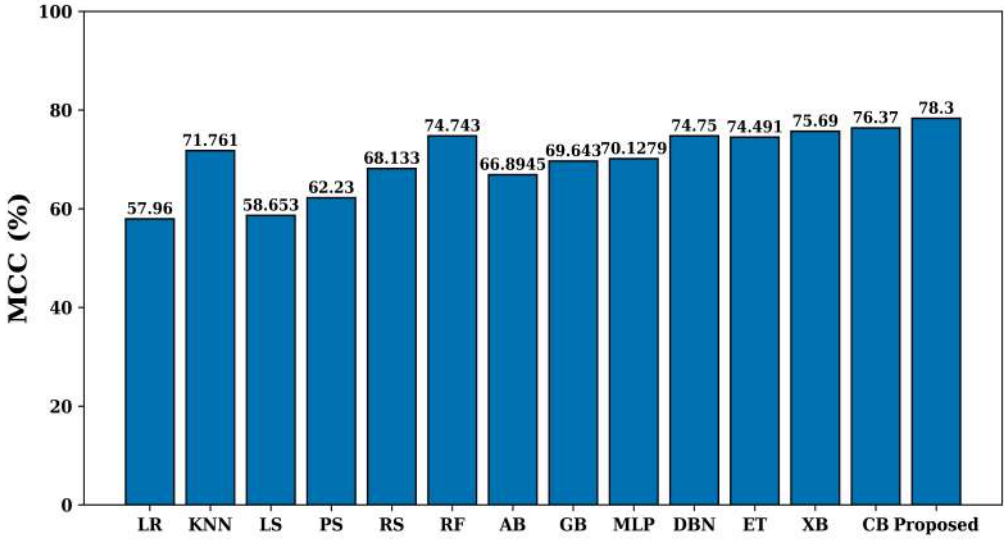


Figure 4.8: MCC of various models

Apart from the already described performance parameters, we have also plotted ROC and PR curves for various models on the test dataset. Since the number of models are more, that is why we have presented the ROC and PR curves in two parts. The ROC curves are shown in Figures- 4.9 and Figures- 4.10, and the PR curves are shown in Figure 4.11 and Figures- 4.12. We have also calculated the Area Under each Curve (AUC) and is mentioned in each figure. Closer the value of AUC to 100%, more efficient is the classification model. After analyzing the figures, it is clear that the proposed model has an area under the ROC curve value of 89.2% and area under the PR curve value of 92.1%, which is maximum when compared with the other studied models.

The combined analysis of all the presented results clearly reveals that the proposed model effectively detects allergen and non-allergen protein sequences. The proposed model has maximum value of detection accuracy, F1-score, MCC and area under the ROC and PR curves.

The model was also validated on a validation data set. The validation data set was created using 50 randomly selected protein sequences, which were initially kept separate, and were not included in training or testing data sets. During validation, it was observed that the model had same performance, when compared with the results evaluated in

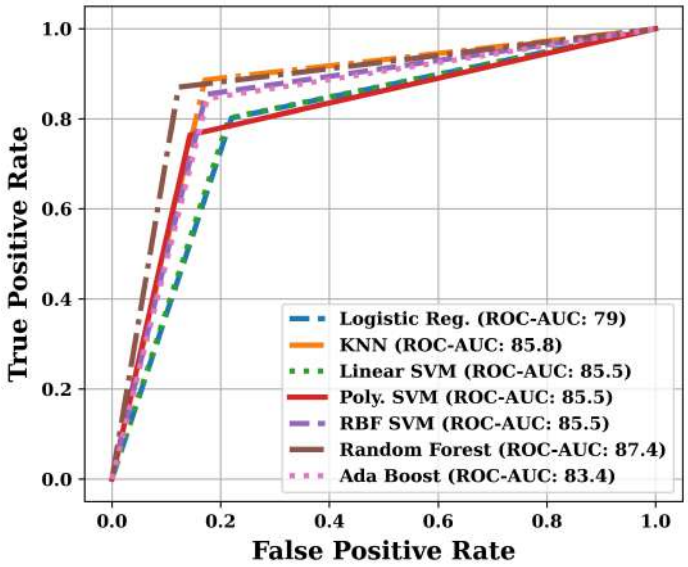


Figure 4.9: ROC curves of various models (Part-I)

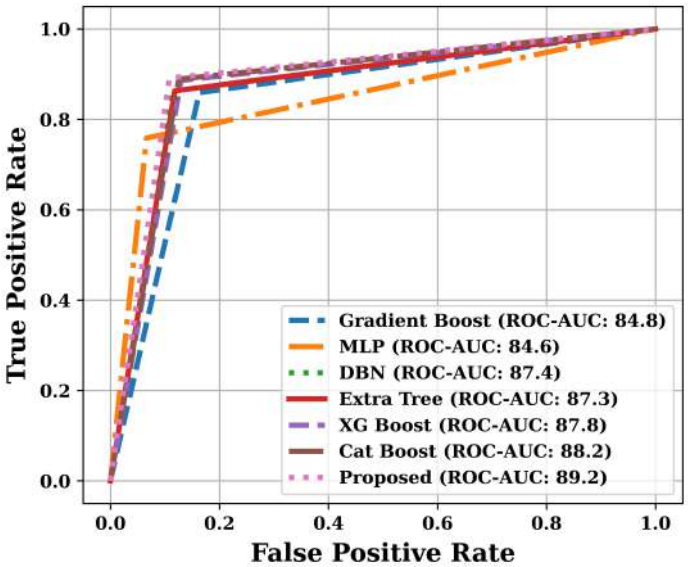


Figure 4.10: ROC curves of various models (Part-II)

testing phase. The overlapping of testing and validation results concludes that the model is stable and can produce same results when tested on similar data sets. The validation results of the proposed model are shown in the form of confusion matrix, in Figure 4.13. The analysis of the values of confusion matrix clearly evaluate the model accuracy as 90% (approx.), which is almost same as evaluated during the testing phase.

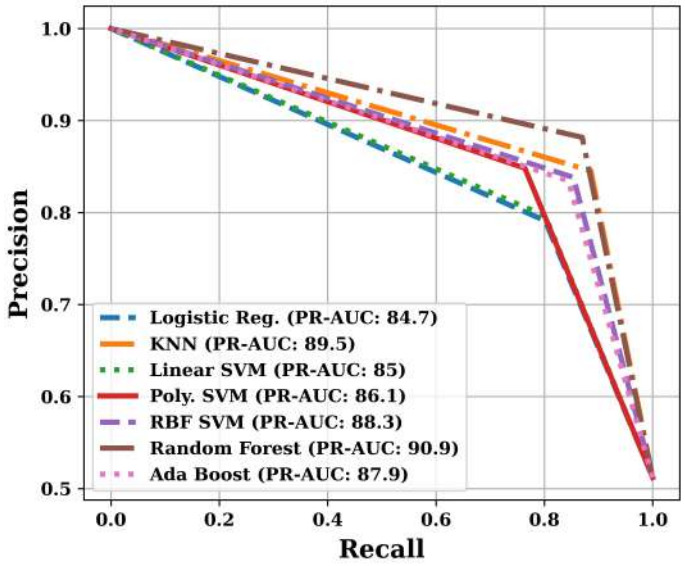


Figure 4.11: PR curves of various models (Part-I)

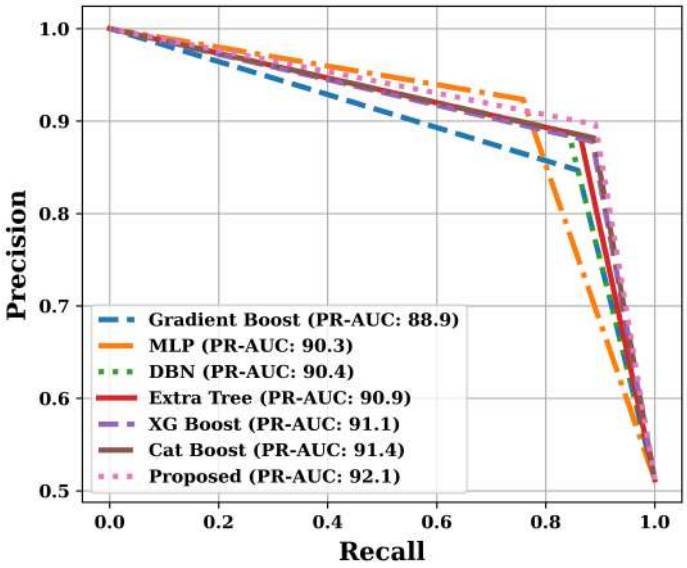


Figure 4.12: PR curves of various models (Part-II)

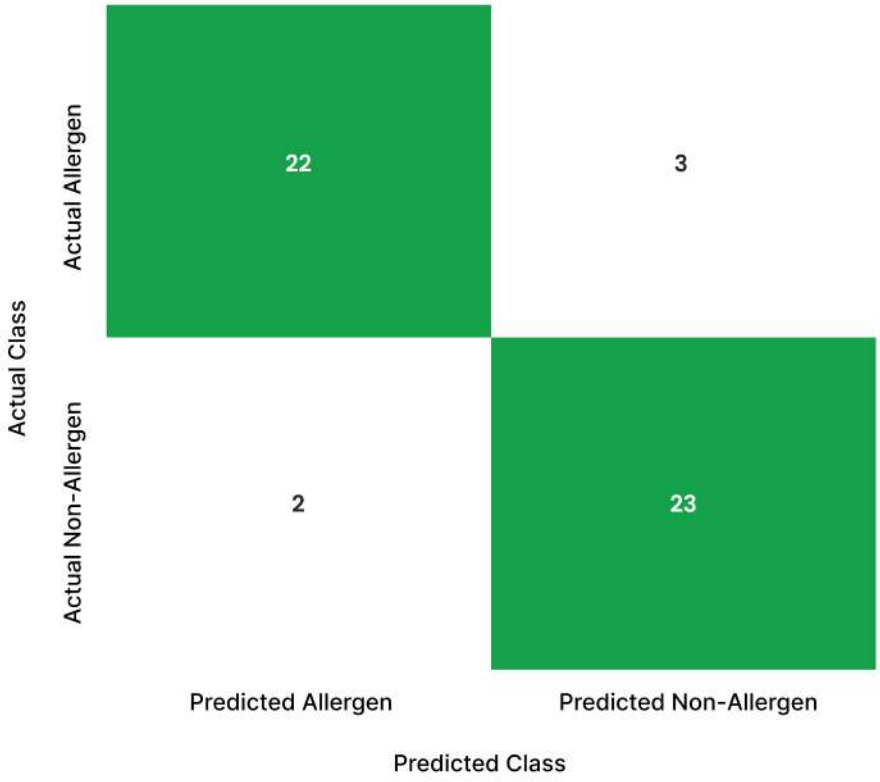


Figure 4.13: Confusion matrix representing model validation results

4.3.3 Performance Comparison

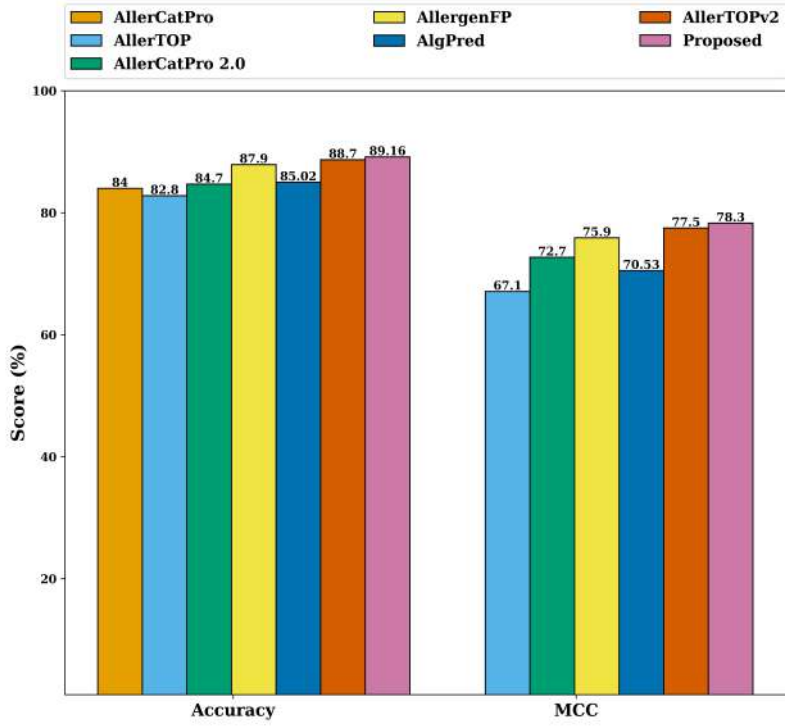


Figure 4.14: Comparison of the proposed model with related work

To further specify the significance of the proposed model, we have compared its performance with AllerTOP [108], AlgPred [109], AllergenFP [110], AllerTOPv2 [75], AllerCatPro [70] and AllerCatPro 2.0 [65]. We have evaluated the performance of the proposed model on different datasets (independent from the one described in previous section of the chapter) used in AllerTOP, AlgPred, AllergenFP, AllerTOPv2, AllerCatPro, and AllerCatPro 2.0 respectively. The performance comparison is shown in Figure 4.14, where it is clear that the proposed model outperforms the specified state-of-the-art literature techniques for allergen classification. The Figure 4.14 shows that the proposed model has an average allergen detection accuracy of 89.16% and MCC of 78.3%, which is highest when compared with the other state-of-the-art literature techniques. The comparative study also shows that the model is free from overfitting issues and is stable.

4.4 Summary

This chapter presented the analysis of various machine learning and deep learning techniques to propose an ensemble learning based approach for the identification of protein allergens. The chapter also highlighted various entities and concepts related to allergy control. In this chapter, we also described various protein sequence properties like Physicochemical, Molecular, Antimicrobial and other characteristics to extract the feature vectors, giving rise to the dataset used for training and testing of intelligent techniques. The performance evaluation of various machine learning, deep learning techniques and the proposed mechanism revealed that the proposed mechanism outperformed the other discussed techniques and achieved an allergen detection accuracy of 89.16%. The performance of the proposed mechanism was also compared with the state-of-the-art literature techniques, which used a similar dataset as used in this study, and the results prove that the proposed mechanism outperforms the other techniques.

Chapter 5

Plant and Animal Food Allergen Classification using Extra Tree based Ensemble Model

This chapter presents the development of a food allergen ensemble classification model based on the extra tree classifier ensemble approach using physicochemical properties. The study used 738 allergen sequences collected from various plants and animals. The study revealed that the proposed ensemble framework has outperformed traditional and various ensemble machine learning models. The performance of the proposed model has been evaluated using various parameters such as accuracy, ROC, sensitivity, and specificity. The model achieved the accuracy of 94.23% with 10 fold cross-validation.

5.1 Overview

An allergen is a protein that can cause a specific allergy associated with it. The dead format of an allergen can induce the antibody, which can target the same type of disease in the future. The immune system's overreaction on an older protein, i.e., allergen, may result in allergy. These overreactions can be in the form of skin rash, swelling of mucous membranes, sneezing or wheezing, or other abnormal conditions. In the human body, allergy can be innate or acquired, and there can be many reasons for allergic reactions in the human body. It can be due to drugs, food, demographic conditions, etc. [111]. Commercial food products are manufactured with modified proteins, leading to allergic reactions. Manufacturers and consumers must identify the types of allergen associated with the original or altered protein [112]. Generally, in all these modified proteins, there is some change in the physicochemical properties of the original protein. The United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) have developed guidelines for evaluating the potential allergenicity of novel food proteins [113].

There are two types of bioinformatics-based allergen prediction. The first approach follows FAO/WHO Codex Alimentarius guidelines and searches for sequence similarity. The second approach is based on identifying conserved allergenicity-related linear motifs [114]. In this study, the physicochemical properties of animal and plant food allergens have been

analyzed. Machine Learning models have been trained and tested on the experimentally verified allergen from the "Allergen Online Database". Based on extracted features, various machine learning models have been built to predict whether the allergen sequence is in the category of a plant or animal food allergen. The proposed ensemble framework based on boosting ensemble technique achieved the highest accuracy of 94.23%.

5.2 Methodology

The description of the dataset, pre-processing and machine learning techniques used in the study are described in this section.

5.2.1 Data set and its Features

The dataset for allergen classification has been extracted from the AllergenOnline Database. This dataset consists of a peer-reviewed list of 2,233 allergen sequences [115]. 738 Experimentally validated animal and plant food allergen sequences have been considered for this study. Animal food allergens were the category whose source was animals, and plants were where an allergen was extracted from plants.

5.2.2 Features Extraction

Various packages are available with R (an open-source GPL), which has been used for feature extraction. Physicochemical properties from these allergen sequences have been extracted with "peptide" package. Various built-in functions have been used to extract the physicochemical properties of particular sequences, which later worked as features for machine learning algorithms. These features have been described in Table- 5.1.

5.2.3 Data Cleansing

The process of finding and fixing mistakes or inconsistencies in datasets is known as data cleaning. Enhancing the quality of data to make it more accurate, comprehensive, and dependable for analysis is the aim of data cleaning. The effectiveness of machine learning models is strongly impacted by the calibre of the input data. Subpar or erroneous models may result from noisy, erroneous, or missing values in the training data. Certain machine learning algorithms exhibit sensitivity towards anomalies, missing values, or discrepancies within the dataset. By making the dataset more stable and dependable, data cleansing helps lower the likelihood that these problems would impair model performance. Various data cleansing techniques employed in the study are described as:

Table 5.1: Physicochemical Properties Extracted from Allergen Sequences

Feature Name	**Description**
boman	Computes the Boman (Potential Protein Interaction) index
hmoment	Homent index
charge	Computes the theoretical net charge of a protein sequence
instaIndex	Computes the instability index of a protein sequence
ppeptide	Probability detection of a peptide sequence
getNofNeighbors	Computes the number of neighbor of degree one for a set of peptide sequences
hydrophobicity	Computes the hydrophobicity index of a protein sequence
mw	Computes the Molecular Weight of protein sequence
pI	Computes the PI value of protein sequence
kideraFactors	Computes the Kidera factors of a protein sequence
aaComp	Computes the amino acid composition of a protein sequence

5.2.3.1 Imbalance Handling

Data imbalance is one of the core challenges in applied machine learning [116]. When dealing with a binary or multi-class classification problem, equivalency in data is essential so that the model can be free from bias [117]. It has been observed that the model can achieve high accuracy with the imbalanced data but that the model will fail in testing with other parameters. In this study, data imbalance has been handled with the Synthetic Minority Oversampling Technique (SMOTE). This technique follows a four-step approach for imbalance handling. which are as follows:

- Choose the minority class; here in this study, the animal allergen is the minority class.
- To each data point available in the minority animal allergen class k- nearest neighbour has been found. 'k' is the number of neighbours around one single data point.
- Place one synthetic point on the line joining the point in consideration and its neighbours.
- Repeat the above steps until there is balancing in the Data-set in consideration.

5.2.3.2 Normalization

Normalization is the process of bringing all the feature set values on the scale between [0,1], keeping the distance scale and information intact. Normalization ensures equal opportunity for each feature to participate in the training of a model [118]. In this study, the MinMax normalization approach has been used for normalizing the values. MinMax normalization approach linearly rescales all the feature values on the scale of [0,1]. Re-scaling is achieved by shifting the minimal value to 0 and dividing it by the new maximal value. The value of every feature vector is normalized by using formulas, as shown in equation 5.1.

$$z = \frac{(originalValue(x) - minimum(x))}{(maximum(x) - minimum(x))} \tag{5.1}$$

In the equation 5.1 Z stands for normalized feature column value. x stands for the original value of the feature set.

5.2.3.3 Outliers Removal

Outliers in the data are the values placed at an unexpected location far from the range of that class data points. The Data-set which has as outliers can also be termed as bad data [119]. Identification and removal of outliers are one of the critical phases of model building. In this study, the data points found away on the scale of 0.05 from the original class label were treated as outliers.

5.2.3.4 Multi-collinearity Removal

Multi-collinearity is a measure of the correlation between independent variables/features. Keeping more than one highly correlated variable will be unnecessary for a prediction model. Highly multi-collinear data may lead to instability in the prediction model [119]. In this study, more than one independent variable with a correlation coefficient of 0.85 has been removed.

5.2.3.5 Transformation

Transformation changes the shape of data so that any suitable distribution can represent it. In this study, the yeo-johnson transformation has been used for data transformation, which transforms the feature set to follow Gaussian-like or normal distribution [120].

5.2.4 Model Building

A binary classification model is built in the model-building phase, which later works as an allergen prediction model. The process of model building is divided into two parts, which are as follows:

- **Model Training:** In the training phase, 70% of the whole data set is bound to the classification algorithm along with features vector and target value. The respective algorithm learns from the available set of feature vectors.
- **Model Testing:** In the testing phase, 30% of the whole data set is used for testing the trained model. This 30% data consisted of only the features vector. Target values are kept reserved for comparison with values predicted by the model. The trained model predicted the value for each feature, and the exact value was compared with the original value to evaluate model performance.

5.2.5 Ensemble Model Building

Ensembling is a collective combination of its participating models. There are three building blocks of the Ensemble Machine Learning approach. viz:

1. Data Sampling/Selection
2. Training Member Classifier
3. Combining Classifiers

To cover all the possible domains of the data set, we need a combination of ensemble members who are diverse in their decisions. Less correlated and independent member classifiers are best suited for effective ensembling. To achieve diversity in the judgement of member classifiers, different subsets of the same data have been used for training purposes. Mainly, there are four types of approaches to train and combine member classifiers, which are described in the following subsections.

5.2.5.1 Bagging

In this approach, independent member classifiers are trained on a data sample with replacement. After the successful training of each member classifier, they are combined through majority voting. The member classifier of bagging is a bootstrap replica of training data, which means each instance has an equal chance to be part of training data.

5.2.5.2 Boosting

Boosting is a robust classifier of its weak member. It is an iterative approach to generate a strong classifier by reducing the training error in binary class problems. It also uses the concept of majority voting but differs from bagging the training data-set for each next classifier focused on those instances that were wrongly classified by the previous one [121]. It works on the principle of three weak classifiers, where the first two were trained on a different subset of the same data, and the third one is trained only on those instances which were wrongly classified by both of them. While working in binary class problem, the least we can expect from each member classifier is their error should be less than 0.5, and the overall error for boosting is expressed as follows:

$$g(e) < 3e^2 - 2e^3 \tag{5.2}$$

where e is the error of any of three classifier.

5.2.5.3 Stack Organization

In the previous approaches, weights were fixed once for all at the time of training. When combining these classifiers, non-trainable combiners combine them, Which does not allow us to determine which member classifier has learnt better on which part of feature space. Stack organization uses a trainable classifier (as shown in Table- 5.2) to overcome this problem, and it combines member classifiers accordingly.

Table 5.2: Stack Generation Member classifier learning

F1	F2	F3	F4	T1 may better learn on this feature set
F5	F6	F7	F8	T2 may better learn on this feature set

5.2.5.4 Mixture of Experts (MoE)

MoE is a classification selection algorithm. In this algorithm, Tier 1 classifiers are trained with appropriate sampling and combined with weighted majority voting. MoE differ from stack organization, here each classifier acts as an expert of their domain. At the time of combination most appropriate classifier got selected through a combination rule. Combining the member classifiers is the last and vital phase of building an ensemble classifier. In this phase, each member classifier learning on data instances is combined to cover the maximum possible data domain. The concept of majority voting has been used to combine the results of the member classifier. The majority voting works as follows.

- Either all members classifier unanimously agree (means all are predicting output for the same class).
- If more than half of the classifiers are agreed on the same class, then that particular class will be treated as the final prediction.
- If some class does not receive more than half of the classifiers vote but still receive the highest vote, then that particular class will be treated as final output. This concept is mostly applicable in multi-class problems. It can also be termed as plurality voting based on the formula as shown in equation 5.3. On each decision d by its member classifiers t, the final classification c will be decided on the basis of maximum agreement on same value.

$$\sum_{t=1}^{T} d_{tc^*} = max_c \sum_{t=1}^{T} d_{tc} \tag{5.3}$$

In case of majority voting accuracy of ensemble is governed by Binomial distribution. With the T member classifiers and k classed accuracy of P_e is given as equation 5.4.

$$P_e = \sum_{k=\frac{T}{2}+1}^{T} \binom{T}{k} p^k (1-p)^{T-k} \tag{5.4}$$

5.2.6 Methodology

Evaluation of the model's performance is done by various parameters such as accuracy, specificity and sensitivity, etc. Testing of correctness of multiple models has been done by repeated k-fold cross-validation. Comparison of performance on these parameters was initially carried out on 14 classification algorithms, including traditional Machine Learning algorithms and ensemble algorithms. Model-based on extra tree classifier has outperformed on all the parameters initially without applying any cleansing approach. Later on, with data cleaning operations viz. imbalance handling, outlier removal, multi-collinearity removal, normalization and transformation, ensemble extra tree classifier has again outperformed all the other models. In this study, various algorithms have been trained and tested, which are mentioned as follows:

- **Decision Tree:** Decision Tree is a supervised machine learning algorithm that works for both classification and regression. For the building of decision trees, the main algorithm called ID3 by J. R. Quinlan is used, which employs a top-down, greedy search through each of the possible branches without any backtracking [122].

Node decision in the decision tree is to be decided by either information gain of Ginni index.

- **Support Vector Machine(SVM):** In the beginning SVM was originally developed to solve binary class problem [123]. The improved version of SVM also deals with the multi-class problem; it first reduces it to the suitable binary class problem after that classification task is performed [124].
- **Random Forest Classifier:** Random forest is an aggregate classifier that contains many decision trees [125]. The RF model generates n data sets with duplicate samples from the original data set. For each dataset, it establishes an unpruned classification or regression tree. For the regression problem, the final result is the average of all the tree results. However, for the classification problem, the final classification result is selected by all trees voting [126].
- **Light Gradient Boosting Machine (LGBM):** LGBM is a boosting based ensemble approach and the framework of this algorithm work on the decision tree. LGBM adds a maximum depth limit to the top of the leaf to prevent overfitting while ensuring high efficiency.
- **Extreme Gradient Boosting(xgboost):** xgboost is the efficient and scalable implementation of gradient boosting framework [127]. The working of xgboost is faster in comparison to other boosting algorithms.
- **CatBoost Classifier:** CatBoost stands for "Category" and "Boosting" algorithm. It is also based on a gradient boosting algorithm that builds a strong classifier from the combination of weak decision trees.
- **Ada Boost Classifier:** Ada Boost stands for adaptive boosting. A meta estimator is built that fitting the original data set and then fits extra copies of the classifier on the same dataset. Weights of wrongly classified records are adjusted such that different classifiers can target complex cases.
- **k-nearest neighbours (knn):** knn is supervised machine learning algorithm can be used for both classification and regression problem. The concept that Similar things exist in close proximity is the foundation of knn. The data points which are near kept in one label and decision of other class is decided based on distance.
- **Ridge Classifier:** Ridge classifier is based on the concept of ridge regression. Class labels are first converted in the range of [-1,1], and then the classification problem is solved as a regression problem. Based on the regression output value final predicted label is decided. In the case of a multi-class problem final label is determined on

the basis of the output regression approach.

- **Linear Discriminant Analysis(LDA):** LDA is based on Bayes' rule. The decision boundary for class separation is decided by fitting class densities on the data set. Gaussian density is fitted to each class based on the assumption that the same covariance matrix is followed in all classes. Fitted LDA can also be used for dimension reduction by projecting the input to the most discriminative directions through the transform method.
- **Quadratic Discriminant Analysis (QDA):** QDA is a version of LDA. Prior covariance matrix knowledge about every class is used in QDA for the prediction model. Dimensionality reduction can not be performed through QDA. This is one of the significant disadvantages of QDA.
- **Logistic Regression (LR):** Logistic regression is an algorithm for classification. It is based on a logistic function which is also known as the sigmoid function. The final prediction in LR is predicted in the form of probability, and values are plotted on the scale of 0 to 1. The class which have more than 50% probability is considered as the final prediction of LR based model.
- **Naive Bayes(NB):** Group of classifiers based on the Bayes' Theorem are called Naive Bayes classifiers. NB classifiers are based on the assumption that each feature in the data set is independent of others and has equal participation in training a model.
- **Extra Tree Classifier:** Extra Tree classifier is the advanced version of random forest classifier. A meta estimator is implemented on various sub-samples of the original data set to fit the different randomized trees.

5.3 Proposed Framework for Food Allergen Classification

The proposed ensemble framework based on the approach described in Section 5.2 is shown in Figure 5.1. A known or unknown food allergen sequences are the required input for the proposed framework. Physicochemical properties for each sequence is calculated as described in Section 5.2.2. These physicochemical properties serve as a feature set for the machine learning models. Once these features are ready data pre-processing phase starts, and the same dataset goes through the cleaning steps as described in Section 5.2.3.

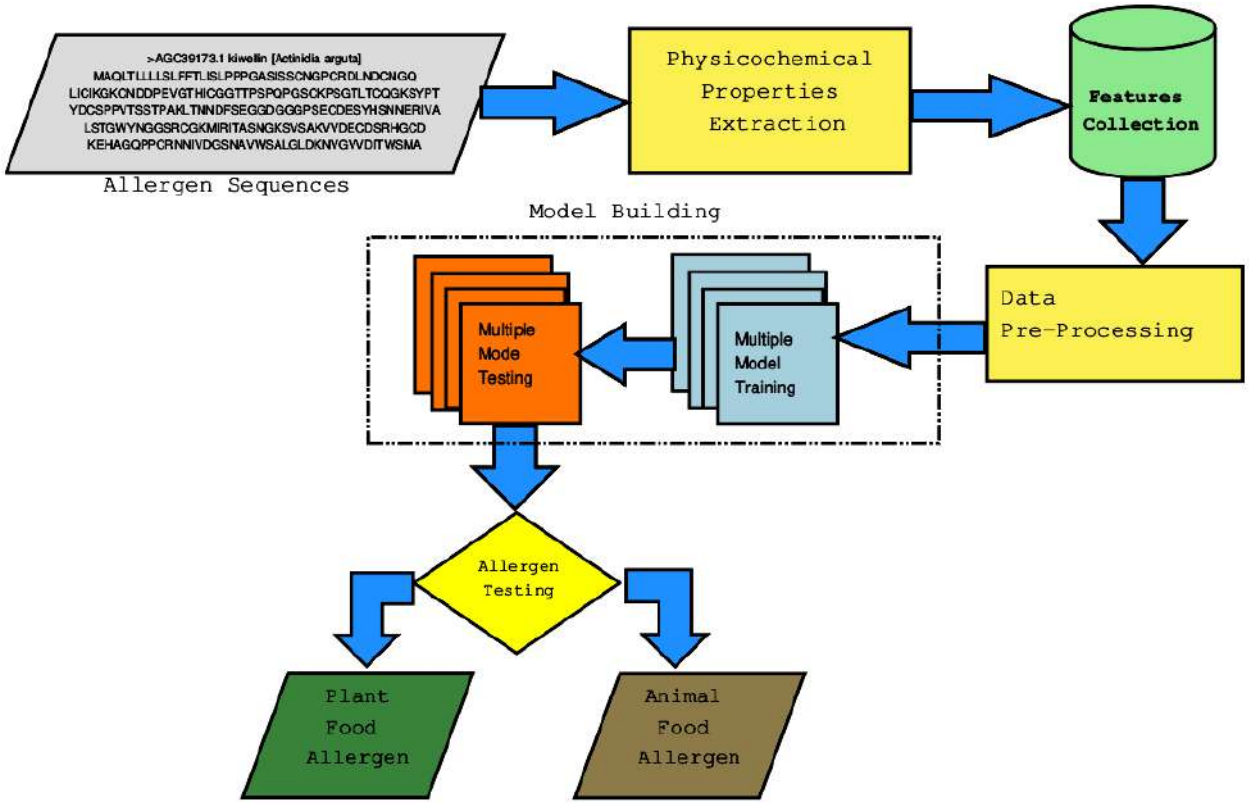


Figure 5.1: System Model

The cleaned data free from all kinds of noise is supplied to the algorithms for model building, which completes the process in two phases, i.e. training and testing. The ensemble model is built on the cleaned data set as described in Section 5.2.5. In the last phase, the result are shown as Plant or Animal Food allergen.

5.4 Model Evaluation Parameters

The performance of a model signifies the stable and efficient working of a model. The efficiency of a classification model refers to the correct prediction of a model concerning all the labels. The accurate prediction should not be confused with the overall correctness, i.e. accuracy. However, accuracy is a widely used parameter for model performance evaluation. It is a necessary parameter, not a sufficient parameter for the overall stability of the model.

Along with the correct prediction, consistency with bias and variance trade-off is also equally important. For the effectiveness and stability measurement of a prediction model in this study, performance has been evaluated on various parameters along with accuracy viz. Sensitivity, Specificity, F1 Score, Receiver operating characteristic (ROC) and

Matthews Correlation Coefficient (MCC).

- Accuracy: It is calculated to measure the overall correctness of classifier which is a ratio of total correct prediction for all the classes to the complete data. From the prediction results of a model, accuracy can be calculated as shown in equation-5.5. Problem-related to overfit and underfit are not addressed by the accuracy parameter value. A model can have good overall accuracy, but still, there is a possibility of high bias and high variance in prediction.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{5.5}$$

  where, TP is True positive, FP is False positive, TN is True negative and FN is False negative.

- Sensitivity/True Positive Rate/Recall: Sensitivity is also known as recall, or true positive rate. It is the ratio of actual positives which are correctly identified as positives by the classifier, to the overall positive and negative observations, and is computed as:

$$Sensitivity = \frac{TP}{TP + TN} * 100 \tag{5.6}$$

- Specificity: It is also known as true negative rate. It relates to the ability of the classifier to identify negative results and is computed as:

$$Specificity = \frac{TN}{TN + FP} * 100 \tag{5.7}$$

- F1 Score/ F Score/ F-Measure: F1 Score is a measure of balance between precision and recall. Weighted average of Precision and Recall can be interpreted as F1 Score.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.8}$$

  The value of the F1 score, as evaluated by the above equation, will vary from 0 to 1. where the zero value of the F1 score signifies the worst performance of the model, and one signifies the best-performing model with balanced precision and recall.

- Receiver operating characteristic (ROC)/ Area under the Curve (AUC): ROC is a probability curve that is plotted with True Positive Rate against False Positives. Whereas, Area under the curve is a degree of measure of separability in classification. It signifies how better the classes are separated in the classification model.

- Matthews Correlation Coefficient (MCC): MCC is mainly used for binary class

classification. When there is an unequal distribution of data among classes, overall accuracy can not signify the performance of a model. MCC is a measure of how better both the classes are classified.

5.5 Result Analysis and Discussion

Experimentally validated 738 allergen sequences have been used for building various machine learning models. 14 Machine Learning models have been trained and tested on the physicochemical properties extracted from these allergen sequences. Performance comparison of these models based on various parameters is shown in Table- 5.3.

Table 5.3: Performance Comparison of various Model on Accuracy, F1 Score and MCC Parameters

	Without Data Preprocessing			**With Data Preprocessing**		
Model	**Accuracy (%)**	**F1 Score**	**MCC**	**Accuracy (%)**	**F1 Score**	**MCC**
Extra Trees Classifier	93.75	0.9568	0.8476	94.23	0.9581	0.8686
Random Forest Classifier	92.96	0.9508	0.8303	93.21	0.9507	0.8463
Light Gradient Boosting Machine	92.59	0.9481	0.821	93.82	0.955	0.8604
Extreme Gradient Boosting	92.38	0.9462	0.8171	92.59	0.945	0.8359
CatBoost Classifier	92.19	0.946	0.8109	93.63	0.9528	0.8599
Decision Tree Classifier	88.47	0.9181	0.7257	88.69	0.9156	0.7476
Ada Boost Classifier	86.13	0.9046	0.6622	86.86	0.9026	0.7034
Quadratic Discriminant Analysis	82.61	0.8803	0.5692	72.84	0.789	0.4171
knn	81.25	0.8718	0.5328	90.32	0.9298	0.7755
Naive Bayes	79.87	0.864	0.4946	71.62	0.7674	0.4317
Linear Discriminant Analysis	78.71	0.8642	0.4352	73.67	0.7906	0.4547
Logistic Regression	78.51	0.8643	0.4301	72.86	0.7815	0.4439
Ridge Classifier	76.95	0.8565	0.3782	73.47	0.7884	0.452
SVM - Linear Kernel	50	0.4142	0	70.99	0.7622	0.4043

Overall prediction accuracy of a Machine Learning Model is one of the widely used performance parameters. Finalizing the model as best based on accuracy is not a practical approach. It does not address the various issues related to data such as class imbalance, underfitting, overfitting. Bias and variance play a vital role in the performance evaluation

of any model. The model with overall high accuracy can still be highly biased due to underfitting, and if there is a high variance in model performance parameter, that will leads toward overfitting. For a stable model, the variance and bias trade-off is an essential part. Keeping bias and variance trade-off in consideration, performance has been evaluated on various parameters. The learning curve and validation curve have been piloted for further performance analysis.

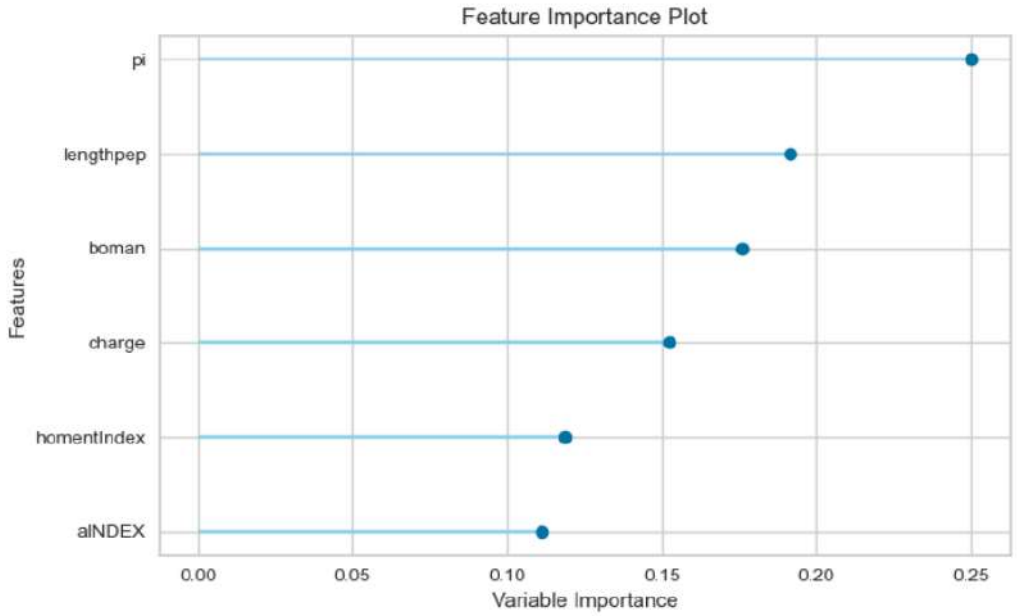


Figure 5.2: Features Importance in Extra Tree Classification Model

All the 14 models have been compared on with and without data pre processing. It is found that Extra tree classifier outperformed with the mean accuracy of 94.23% with 10 fold cross validation. Features that have more impact on the target were considered as more important features. Figure 5.2 shows the set of features considered in Extra Tree Classifier model building. 'pi' is one of the feature which has a high impact on the final prediction model of the allergen category decision system as shown in Figure 5.2. Confusion matrix is shown in Figure 5.3, which is a measure of how correctly the model is behaving for both classes. Rows in Confusion matrix represent Actual value and Columns shows predicted value. The confusion matrix consists of True Positive, True Negative, False Positive and False Negative prediction values, as shown in Table-5.4 also which will further act as a base for calculating other performance parameters.

Decision boundary is the line with two different classes. The boundary of separation by extra tree classifier is shown in Figure 5.4. ROC curve is the trade-off plot between true positive rate and false-positive rate. It plots the area under the curve. In this study, the overall ROC curve and considering each class as positive separately has been plotted as

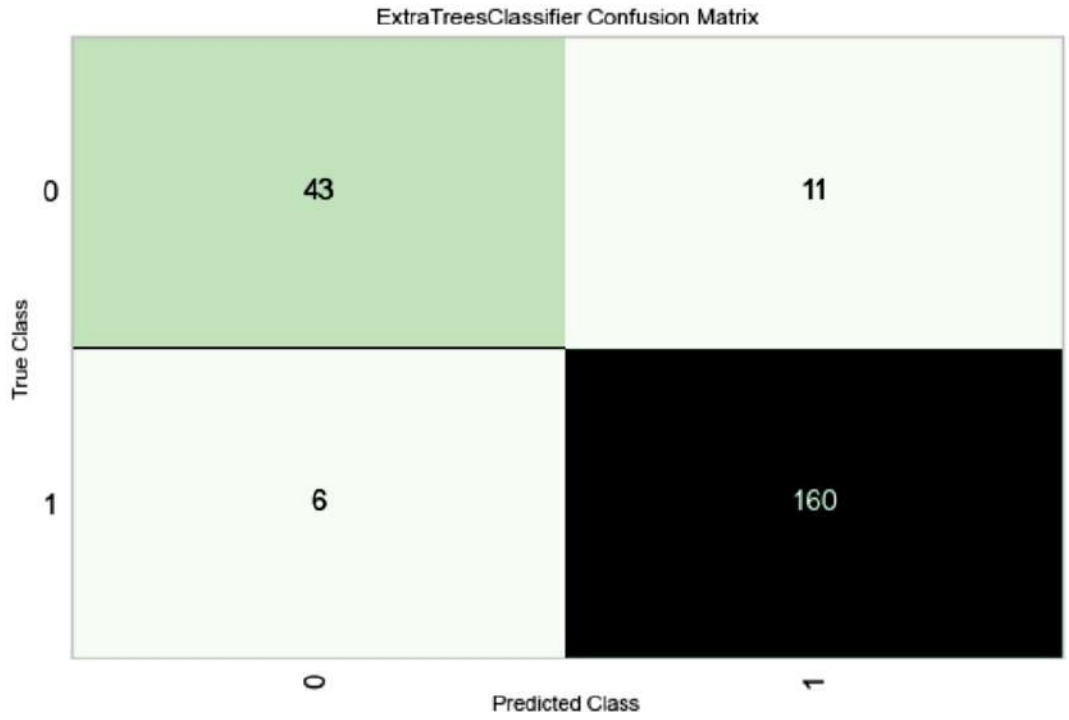


Figure 5.3: Confusion Matrix for Extra Tree Classifier Model after Data Cleansing Operations

Table 5.4: Description of confusion matrix

Parameter	**Definition**	**Values**
True Negative	Sequences which were actually Animal Food Allergen and Predicted as Animal Food Allergen by Extra Tree classifier model	43
False Positive	Sequences which were actually Animal Food Allergen and Predicted as Plant Food Allergen by Extra Tree classifier model	11
False Negative	Sequences which were actually Plant Food Allergenand Predicted as Animal Food Allergen by Extra Tree classifier model	6
True Positive	Sequences which were actually Plant Food Allergen and predicted as Plant Food Allergenby Extra tree classifier model	160

shown in Figure 5.5. The curves converge at a score of more than 0.96, and micro and macro average ROC is also lying and converging on the same scale.

Error report for both the class labels is shown in Figure 5.6. For both the classes of the allergen, the model is behaving efficiently. The ratio of error is almost the same for both the class, which signifies the model is free from bias.

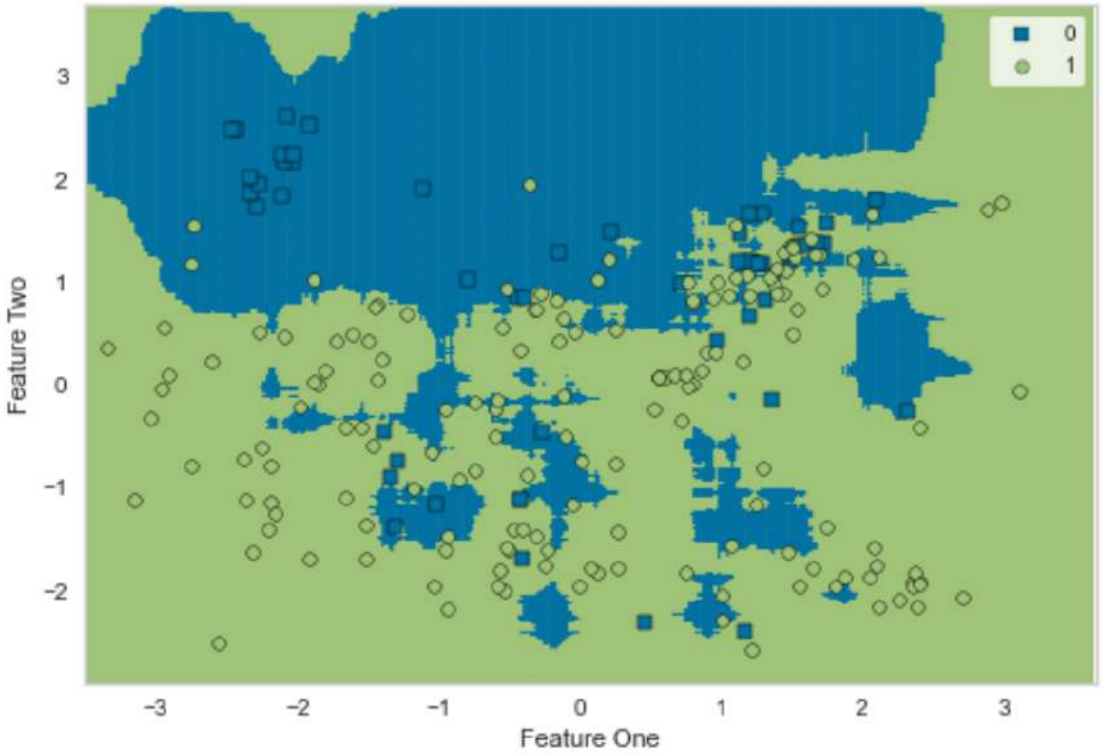


Figure 5.4: Extra Tree classifier Model Boundary

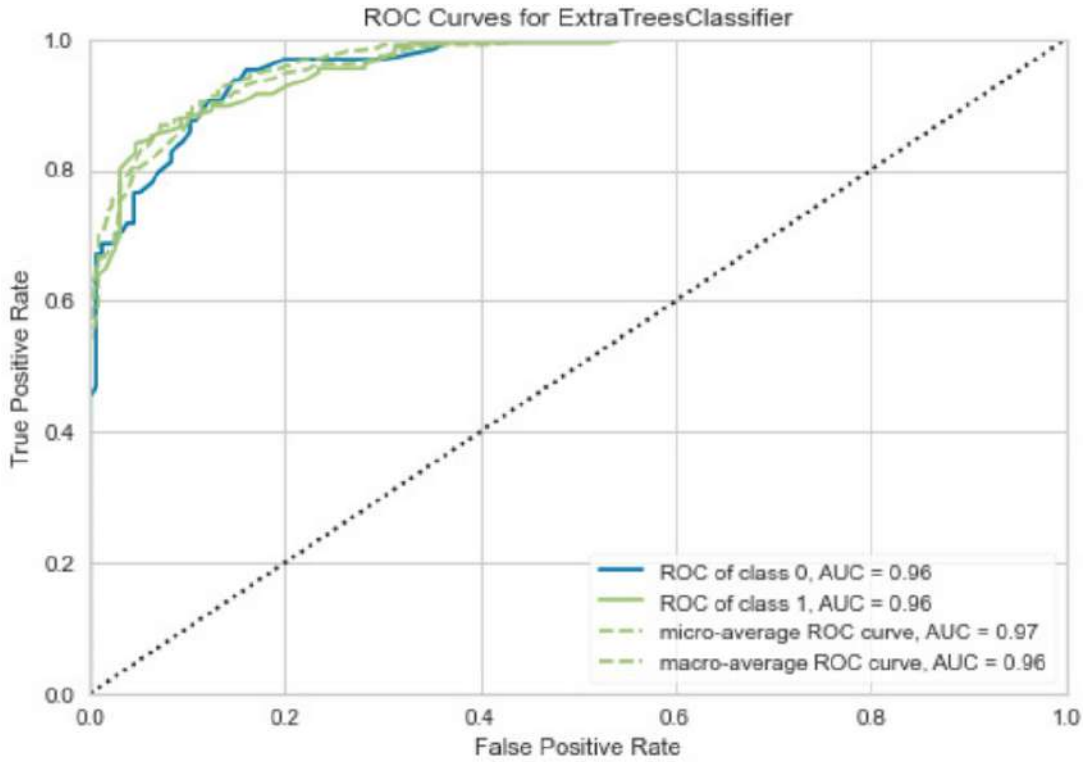


Figure 5.5: Roc for Extra Tree Classifier

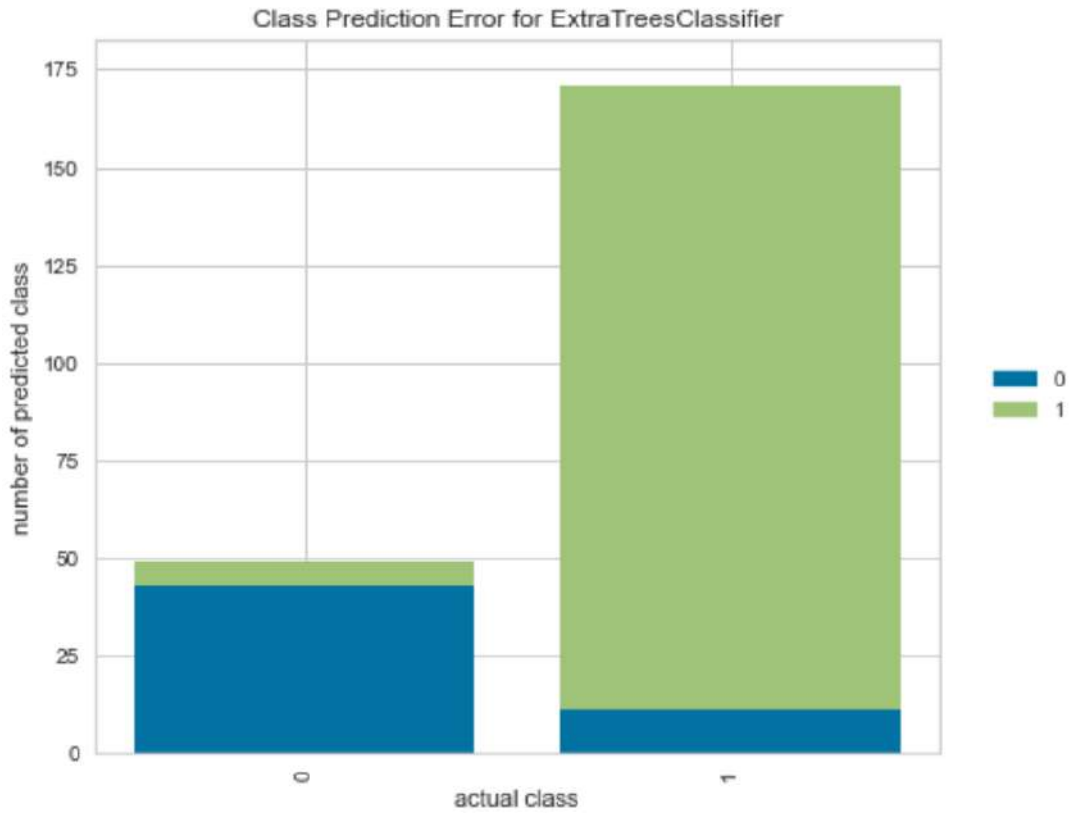


Figure 5.6: Class Error Report for Extra Tree Classifier

5.6 Summary

In this chapter, the physicochemical properties of allergens have been used as features of allergen classification. Based on those features, multiple machine learning models have been built for the classification of food allergen. Among these model Extra Tree classifier ensemble model has outperformed all the existing model with an accuracy of 94.23%. The model was also validated using 10 fold cross-validation approach. A food allergen which can cause an allergy in human body can be predicted by the proposed ensemble classification framework.

Chapter 6

Stochastic Mathematical Model for Allergy Prevalence Prediction

In this chapter, numerical simulations of a six-compartment model were conducted in order to predict the likelihood of developing allergies. The research categorised the division of model into four classes of human population, along with two allergen groups, namely inhalants and food. The study aimed to develop a predictive model for allergen sensitization occurrence using a system of nonlinear differential equations. The numerical presentation of the system's solution was achieved through the utilisation of the stochastic computational artificial neural network (ANN) and the Levenberg-Marquardt backpropagation (LMBP). The neural network process involves presenting three types of statistics: training, testing, and sampling. The allergy occurrence prediction model was trained on 72% and tested and validated on 14% for both. The accuracy of the stochastic scheme was evaluated by comparing its performance with the proposed and reference solutions. A small absolute error indicated the correctness of the proposed scheme. The research observed the reliability of the stochastic solver by utilizing various statistical operators.

6.1 Overview

The allergic situations are considered by using the expression of activating allergen antibodies based on the ingestion or inhalation. The frequent sources of the inhaled allergen are moulds, tree/grass pollen, and house dust mites. The food allergen selection is based on milk proteins, egg whites, fish, wheat, fruits, and nuts. The food allergen supplies by which a person is revealed to influence through the dietary options, alter a population that transfers to rural backgrounds. The dominance of conventional food items is performed in urban settings, where the selections are greater along with the increasing variety. The fauna and flora based on urban settings are subjected to ornamental trees, lawn grass, and weeds, which are considered different to obtain through arid rural settings [128].

There are various studies that have been presented based on the allergic conditions, including the erratic progression behaviour, which is not easy to diagnose due to its

potential stressful impacts and cumulative health care load in various regions of the world [129]. Few expressive investigations have been proposed based on the Sub-Saharan regions like Nigeria, South Africa, Botswana, and Zimbabwe, which shows the increment in the allergy prevalence. Zimbabwe is such a country that has health urgencies, which are biased to the humid contagious diseases [130, 131, 132].

Generally, allergic diseases, along with their effects on life quality, are not highlighted in a region captured by major contagious diseases like AIDS, malaria, HIV, and tuberculosis. Consequently, the important levels of allergy care are based on human resource training, planning, and optimum mitigation arrangements, along with disease management and its diagnosis, which need to be minimized.

The data on allergies is regularly collected by the World Health Organization (WHO) to guide and treat asthma. The periodicity and prevalence rates are performed in various patients, including the diseases of inhalant allergy and mortality rates. The non-transmissible form of allergic diseases is not systemically accumulated nationally, which is not suggested by WHO. The edifying information is not described to damage the care of the allergy patient [133].

The current form of the system is presented using the susceptible, exposed, infected, and recovered, i.e., SEIR model. The modified form of this model is presented by Mushayi et al. [134]. The modelling form of the allergy does not correspond to the modelling of the transferable diseases together with the comparative diseases using the environmental spread trials [135]. A similar mathematical nonlinear system is presented to investigate the asthma spreading due to inhaled chemicals for the industry and tobacco smokers based on the variable size population, which is accessible in Tripathi and Naresha [136].

One of the sensitized forms of the allergen, along with its groups, inhalants, and food. The nonlinear form of the SEIR system has been provided to focus mainly on the spread of the human population together with allergens and represented mathematically by equation-6.1 [134]. Based on the literature, the mathematical model for allergy prevalence prediction represented in equation-6.1, has never been solved before by using the Artificial Neural Network (ANN) techniques together with the Levenberg-Marquardt backpropagation (LMBP) technique.

Therefore, the main aim of this research is to present a stochastic computational procedure for the stochastic analysis of a mathematical model for allergy prevalence prediction (APP). To predict the prevalence of allergies, the dynamics of the APP model are divided into six classes through a system of nonlinear differential equations. Four of them represent the distinct epidemiological classes of the human population, and the other two

represent allergen groups, namely inhalants and food.

The artificial neural network (ANN) techniques have been implemented to solve the mathematical model for APP together with the Levenberg-Marquardt backpropagation (LMBP) technique. The numerical simulations of the APP model are performed efficiently and more accurately by utilizing the proposed stochastic computational procedure of ANN-LMBP technique. The ANN-LMBP process involves three types of statistics: training, testing, and sampling. To test the performance of the APP model the data for the statistical analysis is chosen as 72% for training data and 14% for both authentication and testing data.

The correctness of the proposed stochastic computational procedure of the ANN-LMBP technique is observed in the comparison of the obtained and reference solutions, along with the performances of the absolute error (AE), ranging around 10^{-06} to 10^{-07}. The reliability of the proposed ANN-LMBP technique is validated through multiple statistical performances for the APP model to predict the prevalence of allergen sensitization and presented graphically.

6.2 Designed mathematical model and stochastic framework

This section shows the designed mathematical model and the stochastic computing framework for solving the nonlinear model including the stochastic scheme and the process of its implications. The current form of the system is presented using the susceptible, exposed, infected, and recovered, i.e., SEIR model. The modified form of this model is presented by Mushayi et al. [134]. The modelling form of the allergy is not correspondent to the model for transferable diseases together with the comparative diseases using the environmental spread trails [135].

A similar mathematical nonlinear system is presented to investigate the asthma spreading due to inhaled chemicals for the industry and tobacco smokers based on the variable size population, which is accessible in Tripathi and Naresha [136]. One of the sensitized forms of the allergen, along with its groups, inhalants, and food. The nonlinear form of the SEIR system has been provided to focus mainly on the spread of the human population together with allergens, mathematically shown as [134]:

$$
\begin{cases}
S'(x) = \Lambda_h - (\beta_1 A_f(x) - \beta_2 A_i(x))S(x) - \mu S(x), \\
X'_f(x) = \beta_1 A_f(x) S(x) - (\gamma_1 + \mu) X_f(x), \\
X'_i(x) = \beta_2 X_i(x) S(x) - (\gamma_2 + \mu) X_i(x), \\
X'_T(x) = \gamma_1 X_f(x) + \gamma_2 X_i(x) - \mu X_T(x), \\
A'_f(x) = \Lambda_f - \upsilon_f A_f(x), \\
A'_i(x) = \Lambda_i - \upsilon_i A_i(x)
\end{cases} \tag{6.1}
$$

The above four dynamics of the model represent the human populations and the last two classes represent the Allergens. The compartmental representations are shown in Figure 6.1, whereas the descriptions of each parameter that is used in system 6.1 are given in Table-6.1.

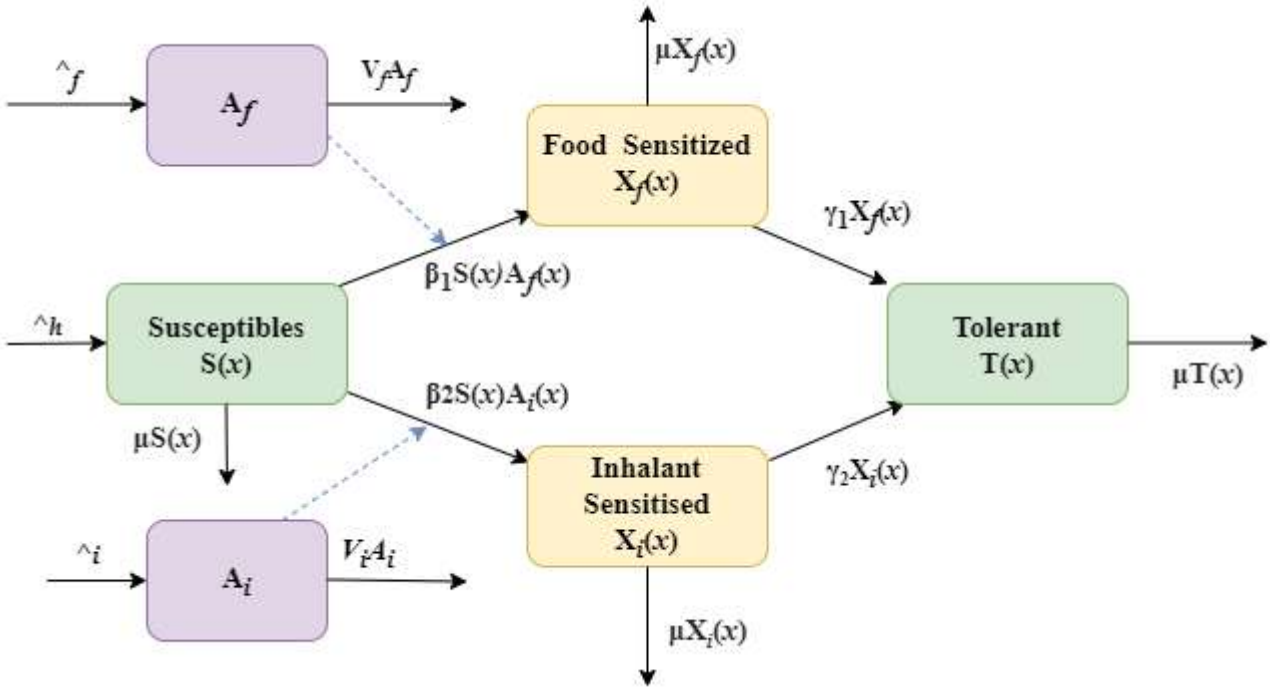


Figure 6.1: Model framework representations

Various differential models based on the nonlinear system of differential equations have been proposed and solved numerically by using different schemes [137, 138]. The current numerical performances have been used for the mathematical model with its six compartments to forecast allergy occurrence. The division of human population is divided into four classes along with two allergen groups (inhalants and food), which predict the allergen sensitization occurrence. These compartments of the allergen groups designate the system of nonlinear differential equations, and the solutions are numerically presented

Table 6.1: Parameter details of the prevalence prediction of allergies model

Parameters	**Description**
β_1 and β_2	Rate of sensitization through the susceptible by cooperating with the allergens based on food and inhalant
υ_f and υ_i	Depletion rate using the food and inhalant allergens
μ	Natural rate of death or birth
γ_1 and γ_2	Developing tolerance rate
Λ_h	Susceptible population's growth rate
Λ_f and Λ_i	Growth rate of food and inhalant allergens
x	Time

through the computational artificial neural network (ANN) and the Levenberg-Marquardt backpropagation (LMBP).

Recently, these solvers have been explored to present the solutions of many dynamical systems, few of them are food chain models [139, 140], nonlinear smoking dynamical system [141], eye surgery systems [142, 143], thermal explosion models [144, 145], and biological differential models [146, 147]. Few pioneering perceptions of the scheme are given as:

- The present differential system is proposed to solve the allergen sensitization prediction model, and these compartments designate the nonlinear system.
- The nonlinear model 6.1 is divided into four dynamics based on the human populations, while the last two classes represent the Allergens.
- The stochastic computing ANN-LMBP performance is used for the numerical solutions of the mathematical system.
- The aptitude of the Adams scheme is employed to build the networks for the approximate results of the system, which are applied for testing, certification, and training.
- The consistent overlapping of the results together with strong and reducible absolute error (AE) represent the competency of the scheme for solving the Allergen model.
- The consistent/exact matching of the solutions, small AE and reliable statistical performances indicate the implication of the scheme.

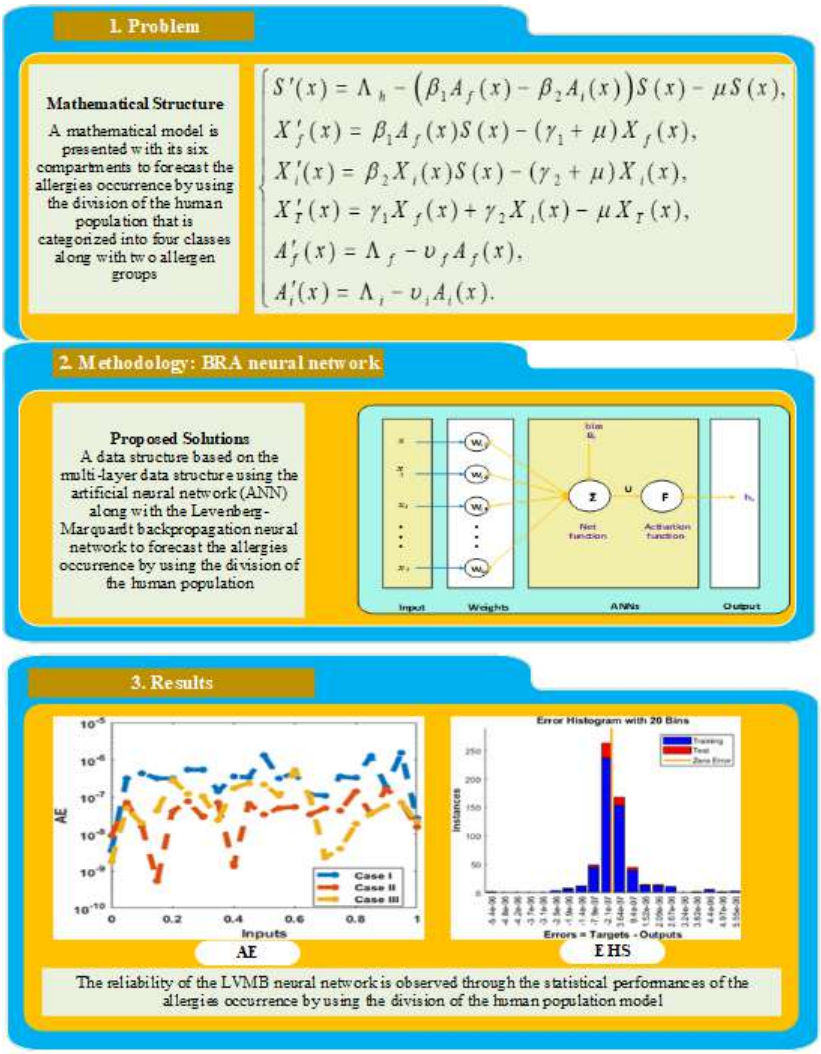


Figure 6.2: Designed procedure for the nonlinear system

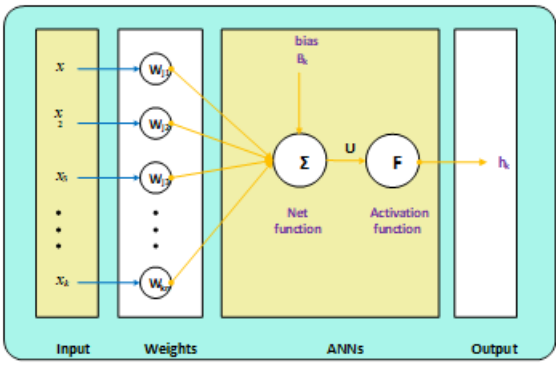


Figure 6.3: A structure of the single neuron

Figure 6.2 represents the proposed computational procedure based on the nonlinear system, which is divided into four dynamics based on the human populations, while the

last two classes represent the Allergens. Figure 6.3 presents the multi-layer procedure through the structure of a single neural network. The stochastic procedure is executed using the 'nftool' a solver in 'MATLAB', validation through n-folded, multiple layers, nineteen neurons, log-sigmoid transfer function, maximum 700 epochs, tolerances, supervised algorithm, step size, and terminating criteria. The training and label data for inputs/targets are obtained using the standard outputs, whereas the statistics is applied as 72% of train data and 14% for both test and validation.

LMBP is one of the popular optimization schemes that is applied to train artificial neural networks, remarkably to solve optimization and nonlinear regression problems. It shows the expansion of the typical backpropagation scheme and is proposed to get the biases and optimal weight sets for a neural network in order to minimize the error function.

LMBP has specific advantages over the conventional gradient descent optimization techniques to train the neural networks. LMBP converges faster and performs less sensitivity to the excellence of the learning cost. LMBP is remarkably helpful for problems based on nonlinear regression and in such conditions where the surface of error can have adapting curvatures. In addition, like other optimization schemes, it can be stuck in local minima, so regularization and initialization approaches are critical to training the neural networks. A few concise points of the LMBP are presented as:

1. *Start:* Initiate with a primary speculation based on the network's biases and weights.
2. *Forward Pass:* Based on the input data set that is used to perform a forward pass in the neural network to calculate the expected productivity.
3. *Error computation:* Find the cost or error function that is used to quantify the difference between the actual target performances and predicted outputs.
4. *Jacobian Matrix:* Calculate the Jacobian matrix that shows the error gradient with respect to the parameter of the network (biases and weights).
5. *Update Biases and Weights:* Based on the Jacobian matrix, the scheme computes the renewal phases for the factors. LMBP combines the Gauss-Newton schemes and gradient descent elements. It modifies the step size using the local curvature based on the surface of the error.
6. *Iterate:* Continue repeating steps 2 to 5 till a stopping process is met. This condition can have a certain iteration, completing a limited error tolerance, or a few other conditions of convergence.

The neural network performances by taking nineteen numbers of neurons have been pro-

vided based on the underfitting/overfitting using statistics as 72% for training and 14% for both test and validation data. For smaller neuron values, premature convergence is noticed, while overfitting behaviour is shown in the case of higher neurons. While over-simplified, this mathematical model emulates some characteristics exhibited by organic neurons. The components of the perceptron, as shown in Figure 6.3 are as follows:

- *Inputs:* The inputs, denoted as the vector $X = [x_1, x_2, \ldots, x_n]$, consist of the data points or characteristics extracted from a provided dataset, which are then supplied to the neuron. The values in question may consist of either raw data or a modified form, with each element of this vector representing a unique dimension or characteristic of the data. In the majority of real-world datasets, the dimensionality, denoted as n, exhibits a wide range, spanning from a small number to several thousand or even greater.

- *Weights:* The characteristics that govern the significance or impact of individual inputs on a neuron's output are commonly denoted as weights. These weights are typically represented as a vector, denoted as $W = [w_1, w_2, \ldots, w_n]$. The primary focus of the learning process in machine learning is centred on the optimisation of these weights. The parameters are initialised, often either randomly or according to a preset plan, and then iteratively changed throughout the training phase in order to minimize the error in predictions.

- *Bias:* Bias, represented by the symbol b, can be likened to the y-intercept in linear equations. The neuron is enabled to provide outputs that are not only determined by the weighted inputs. The introduction of a shift or offset facilitates improved data fitting, particularly in cases where the data does not intersect the origin. The flexibility afforded by this mechanism enables neurons to more correctly reflect both data and functions.

- *Weighted Sum:* Prior to activation, the neuron performs a computation known as the weighted sum, which involves taking a linear combination of its inputs and their respective weights. The mathematical expression for this is represented as $z = w \cdot x + b$.

- *Activation Function:* With the help of an activation function, non-linearity is introduced, which lets a neuron or a network of neurons effectively collect and show complex connections within data. By subjecting the weighted sum to this function, the output of the neuron is not only a linear conversion of its input. Several activation functions are frequently used in various applications.

- *Sigmoid Function:* Defined as $\sigma(z) = 1/(1 + e^z)$. Its S-shaped curve bounds the

output between 0 and 1, making it especially suitable for binary classification tasks.

- *Y hyperbolic Tangent (tanh):* Defined as $tanh(z) = (ez + e - z)/(ez - e - z)$. This function scales outputs between -1 and 1, offering a zero-centered range which can be advantageous during training.
- *Output:* The neuron's output is the result of processing the weighted sum through the activation function: $y = a(z)$ Where y is the neuron's output and $a(z)$ denotes the value produced by the activation function when z is provided as input. Although a single neuron provides a rudimentary depiction, it serves as a symbol of the fundamental principles that form the basis of more advanced neural network structures.

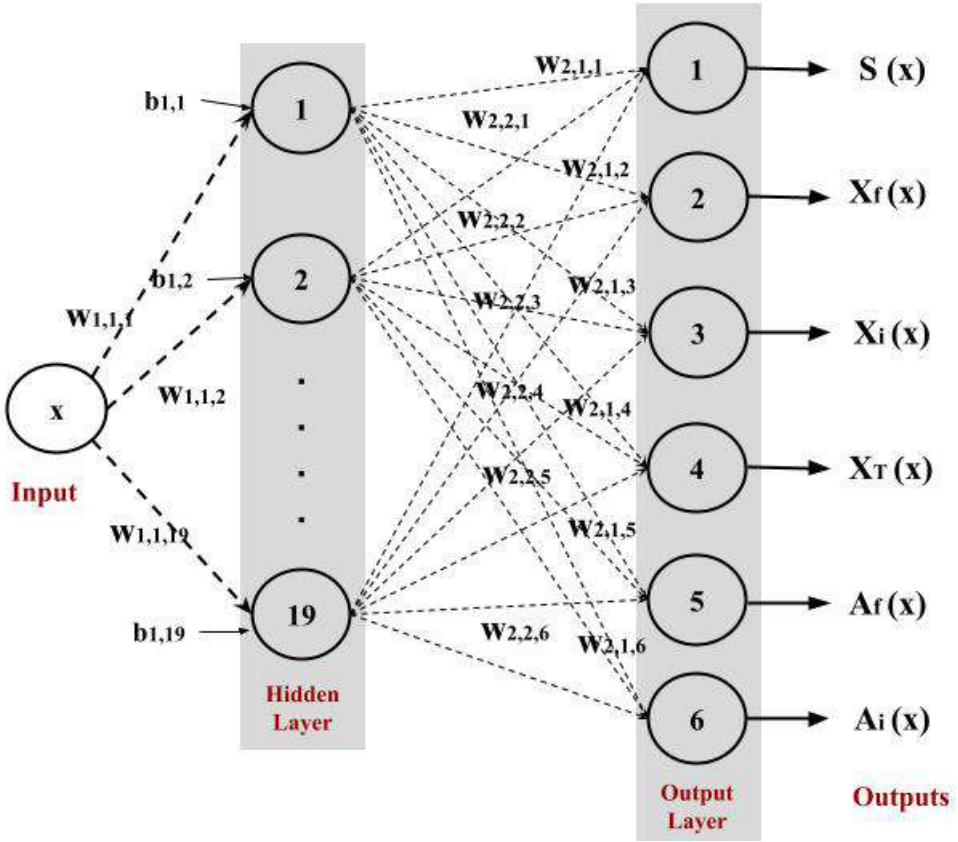


Figure 6.4: Neural structure for solving the mathematical system

The neural structure based on different layers is presented in Figure 6.4. The visualization provided in Figure 6.4 depicts the neural architecture employed in the process of addressing mathematical systems. It effectively highlights the fundamental components that facilitate this computing mechanism are as follows:

- *Input:* The layer depicted on the far-left side of the diagram consists of several

nodes, with each node representing a distinct property or variable within the mathematical system under consideration.

- *Hidden:* Positioned intermediary to the input and output layers, the hidden layers include a multitude of neurons responsible for the transformation of the incoming data. The processing of information by individual neurons involves the consideration of connection weights and biases. Although, for simplicity, only one hidden layer has been illustrated, it is important to note that neural networks have the capability to include several hidden layers. This allows the model to go deeper and possibly enhances its ability to capture complex correlations within the data.
- *Output:* The layer situated on the far-right side represents the conclusion derived by the neural system or the solution obtained for the mathematical system. In instances with regression challenges, the neural network may consist of a solitary neuron; however, in situations involving classification problems, several neurons may be included.
- *Weights:* The connections between neurons across layers are shown by lines or arrows. The intensity and direction of signal transmission between neurons are determined by weights. During the training process, the weights of the network are iteratively modified in order to optimize the network's ability to approximate the solution to the given mathematical problem.
- *Biases:* the extra factors associated with each neuron outside of the input layer, are not necessarily represented directly. As a result of these biases, the network is better able to fine-tune its predictions, giving it greater latitude to capture patterns and complexities in the data.

6.3 Numerical performances of the model

The numerical results of the differential model for three different cases by applying the proposed scheme is presented in this section, given as:

Case 1: Put the parameter values $\beta_1 = 0.066, \beta_2 = 0.2363, \mu = 0.0342, \upsilon_f = 0.347, \upsilon_i = 0.2124, \gamma_1 = 0.0116, \gamma_2 = 0.005, \Lambda_h = 1, \Lambda_f = 0.8945, \Lambda_i = 0.7608$, while the initial conditions (ICs) are $S(0) = 0.1, X_f(0) = 0.2, X_i(0) = 0.3, X_T(0) = 0.4, A_f(0) = 0.5, A_i(0) = 0.6$, in system 6.1. The updated form is achieved as:

$$\begin{cases} S'(x) = 1 - (0.066A_f(x) - 0.2363A_i(x))S(x) - 0.0342S(x); S(0) = 0.1 \\ X'_f(x) = 0.066A_f(x)S(x) - 0.0458X_f(x); X_f(0) = 0.2 \\ X'_i(x) = 0.2363X_i(x)S(x) - 0.0392X_i(x); X_i(0) = 0.3 \\ X'_T(x) = 0.0116X_f(x) + 0.005X_i(x) - 0.0342X_T(x); X_T(0) = 0.4 \\ A'_f(x) = 0.8945 - 0.347A_f(x); A_f(0) = 0.5 \\ A'_i(x) = 0.7608 - 0.2124A_i(x); A_i(0) = 0.6 \end{cases} \tag{6.2}$$

Case 2: Put the parameter values $\beta_1 = 0.066, \beta_2 = 0.2363, \mu = 0.0342, \upsilon_f = 0.347, \upsilon_i = 0.2124, \gamma_1 = 0.0116, \gamma_2 = 0.005, \Lambda_h = 1, \Lambda_f = 0.8945, \Lambda_i = 0.7608$, while the initial conditions (ICs) are $S(0) = 0.2, X_f(0) = 0.3, X_i(0) = 0.4, X_T(0) = 0.5, A_f(0) = 0.6, A_i(0) = 0.7$, in system 6.1. The updated form is achieved as:

$$\begin{cases} S'(x) = 1 - (0.066A_f(x) - 0.2363A_i(x))S(x) - 0.0342S(x); S(0) = 0.2 \\ X'_f(x) = 0.066A_f(x)S(x) - 0.0458X_f(x); X_f(0) = 0.3 \\ X'_i(x) = 0.2363X_i(x)S(x) - 0.0392X_i(x); X_i(0) = 0.4 \\ X'_T(x) = 0.0116X_f(x) + 0.005X_i(x) - 0.0342X_T(x); X_T(0) = 0.5 \\ A'_f(x) = 0.8945 - 0.347A_f(x); A_f(0) = 0.6 \\ A'_i(x) = 0.7608 - 0.2124A_i(x); A_i(0) = 0.7 \end{cases} \tag{6.3}$$

Case 3: Put the parameter values $\beta_1 = 0.066, \beta_2 = 0.2363, \mu = 0.0342, \upsilon_f = 0.347, \upsilon_i = 0.2124, \gamma_1 = 0.0116, \gamma_2 = 0.005, \Lambda_h = 1, \Lambda_f = 0.8945, \Lambda_i = 0.7608$, while the initial conditions (ICs) are $S(0) = 0.3, X_f(0) = 0.4, X_i(0) = 0.5, X_T(0) = 0.6, A_f(0) = 0.7, A_i(0) = 0.8$, in system 6.1. The updated form is achieved as:

$$\begin{cases} S'(x) = 1 - (0.066A_f(x) - 0.2363A_i(x))S(x) - 0.0342S(x); S(0) = 0.3 \\ X'_f(x) = 0.066A_f(x)S(x) - 0.0458X_f(x); X_f(0) = 0.4 \\ X'_i(x) = 0.2363X_i(x)S(x) - 0.0392X_i(x); X_i(0) = 0.5 \\ X'_T(x) = 0.0116X_f(x) + 0.005X_i(x) - 0.0342X_T(x); X_T(0) = 0.6 \\ A'_f(x) = 0.8945 - 0.347A_f(x); A_f(0) = 0.7 \\ A'_i(x) = 0.7608 - 0.2124A_i(x); A_i(0) = 0.8 \end{cases} \tag{6.4}$$

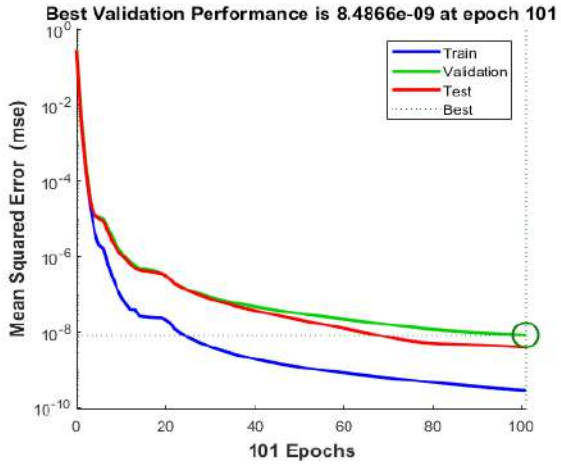
6.3.1 Performance Analysis

The illustrations through the achieved results have been performed in Figure 6.5 to 6.9 for the system 6.1 using the proposed scheme. The capable measures for the system 6.1 based on the best validation and state transitions (STs) has been provided in Figure 6.5. The performance based on the best validation in the perspective of mean square error (MSE) varies on the precise problem to check the nature of the dataset.

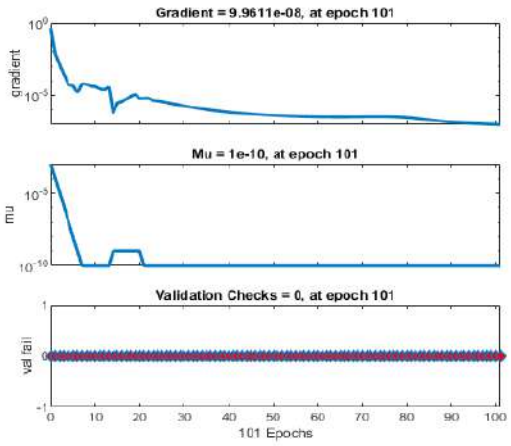
Generally, the objective of the study is to reduce the MSE in the process of validation, as reducible MSE specifies the better analytical exactitude for regression systems. Eventually, to achieve the performances based on the best validation in the process of MSE needs a order of precise model selection, hyperparameter tuning, feature engineering, and a good estimation of the specific dataset and problem.

To present the mathematical form of the model, the MSE values for the train, best curves, substantiation, and test are reported in Figures- 6.5(a–c), while the gradient is provided in Figures- 6.5(d–f). The optimal results are shown at epochs 100, 31 and 561, which are performed as $8.48660 * 10^{-09}$, $4.8848 * 10^{-11}$ and $2.30689 * 10^{-11}$. The gradient measures of the mathematical system are shown as $9.9611 * 10^{-08}$, $9.6982 * 10^{-08}$ and $9.9455 * 10^{-08}$.

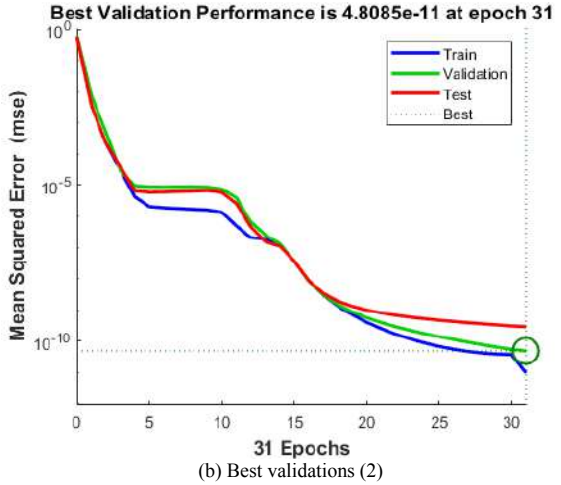
In the perspective of a supervised neural network, “mu”, “gradient”, and “validation checks” are significant theories that are associated with optimizing and training the network. The gradient shows the magnitude and direction of the sharpest by increasing the loss function, validation checks are a central part of improving and monitoring the performances of neural network during the process of training to evaluate it on a distinct authentication dataset, and mu that is also called the learning rate is used to control the step size throughout the optimization of gradient descent. These theories are essential in the process of optimization and training using the supervised neural networks.



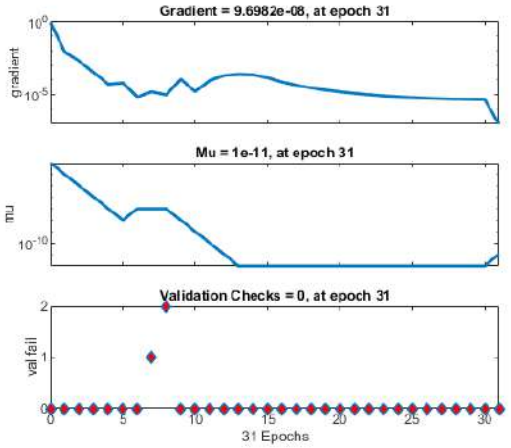
(a) Best validations (1)



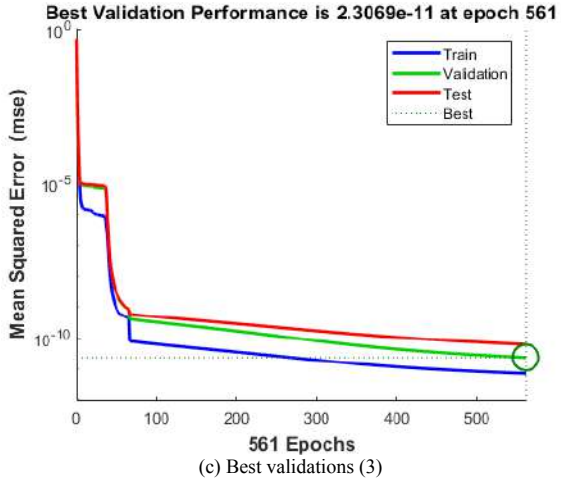
(d) Gradient (1)



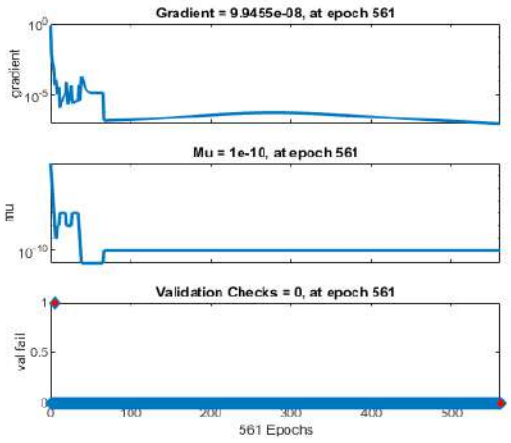
(b) Best validations (2)



(e) Gradient (2)



(c) Best validations (3)

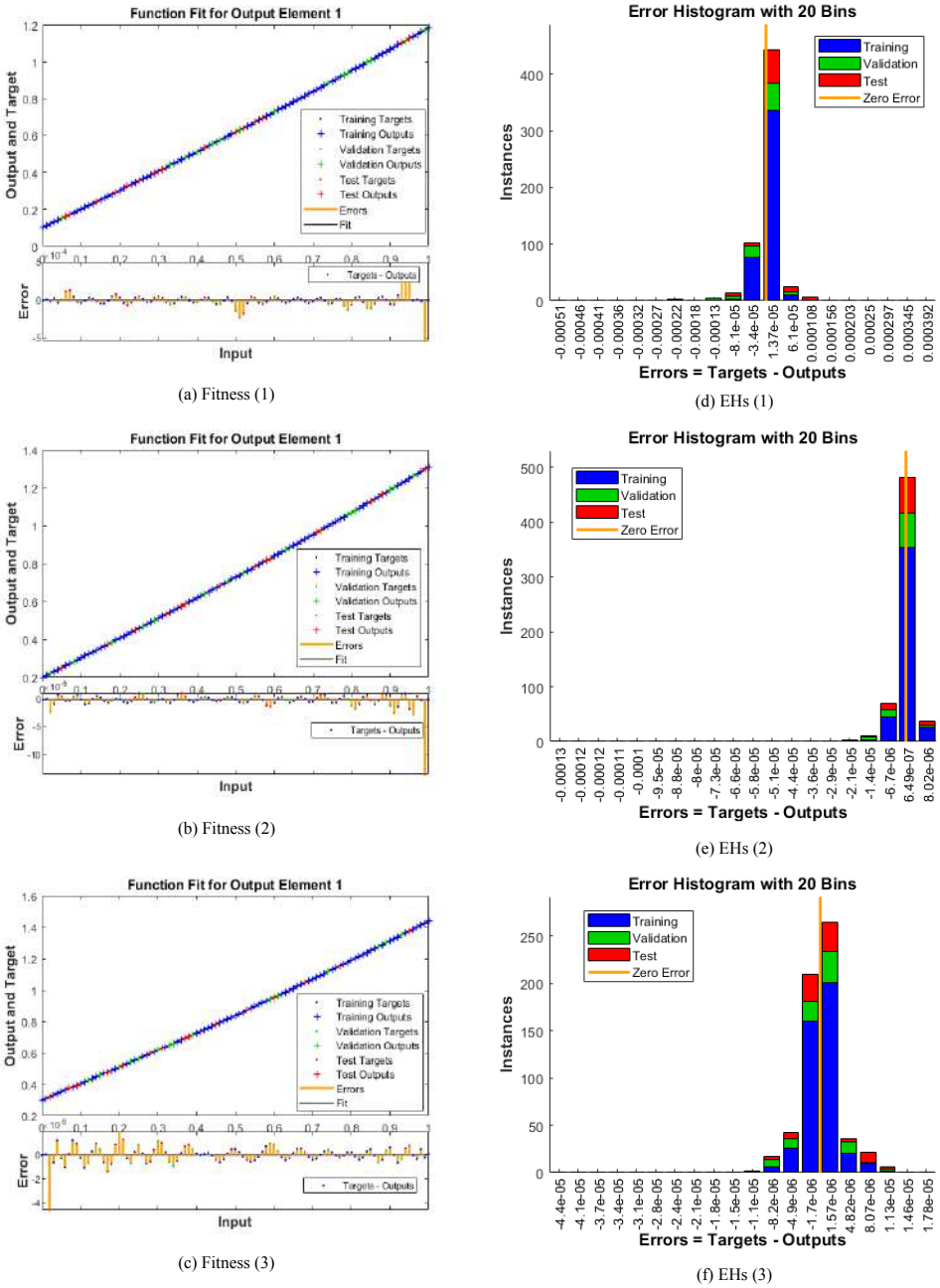


(f) Gradient (3)

Figure 6.5: Best validation and gradient for the nonlinear model

Figures-6.6(a to c) represents the fitting curves performances of the model, which corroborates the comparison of the results. In the process of neural networks, the function

fit usually implies training the network in order to learn a mapping through the input to output data. Figures-6.6(d to f) presents the error histogram (EH), which are performed as $1.37 * 10^{-05}, 6.49 * 10^{-07}$ and 1.5710^{-06} to solve the model. EH is a graphical depiction that presents the residuals or error distribution performed through the model of neural network by using the dataset predictions. EH is an effective vision tool in order to understand the performances of the model and to identify the prospective issues with the guesses. The correlation is performed in Figures-6.7–6.9 for each case of the system.



(a) Fitness (1)

(d) EHs (1)

(b) Fitness (2)

(e) EHs (2)

(c) Fitness (3)

(f) EHs (3)

Figure 6.6: Outputs and EHs for solving the nonlinear model

The correlations using the determination coefficients R2 are provided as 1 for the nonlinear system. The correlation usually provides the connection between the output (or target variable) and input. These plots indicate the test/train and verification data for the correctness of the proposed scheme. The convergence using these data is provided in Table-6.2 to solve the nonlinear model. The complexity is presented to compute the time applied for executing iterations and network training.

Table 6.2: Proposed scheme for the nonlinear model

Case	**MSE**			**Performance**	**Epoch**	**Mu**	**Gradient**	**Time**
	Validation	**Training**	**Testing**					
1	$8.487 * 10^{-09}$	$2.938 * 10^{-11}$	$4.00 * 10^{-09}$	$2.94 * 10^{-11}$	100	$1.00 * 10^{-10}$	$9.96 * 10^{-08}$	2
2	$4.809 * 10^{-11}$	$9.538 * 10^{-12}$	$2.89 * 10^{-10}$	$9.54 * 10^{-12}$	31	$1.00 * 10^{-11}$	$9.70 * 10^{-08}$	1
3	$2.307 * 10^{-11}$	$7.012 * 10^{-12}$	$6.12 * 10^{-11}$	$7.01 * 10^{-12}$	561	$1.00 * 10^{-10}$	$9.95 * 10^{-08}$	4

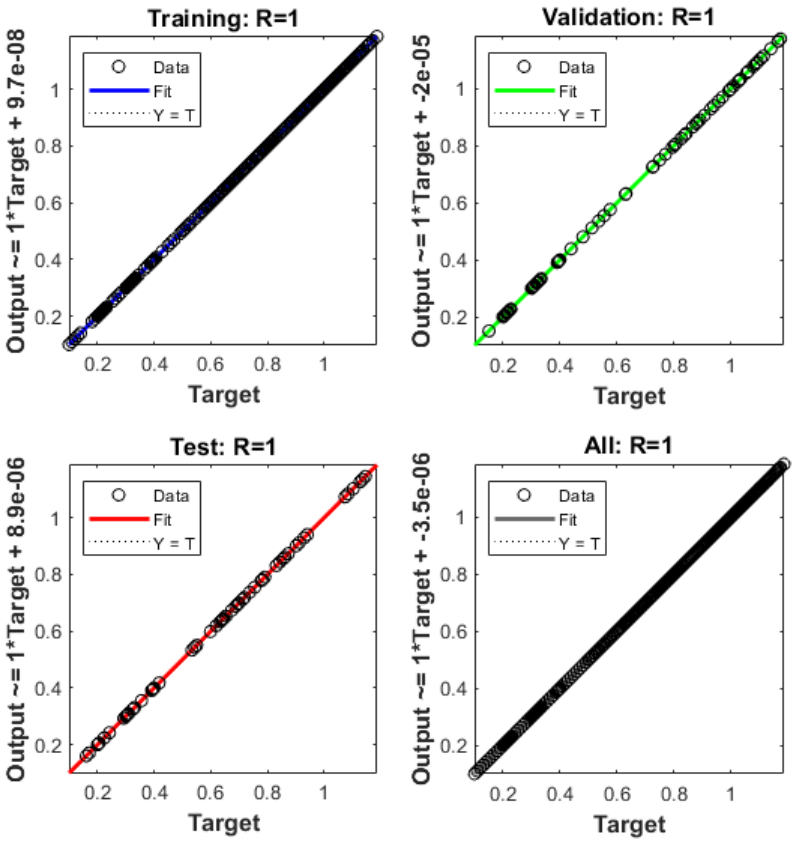


Figure 6.7: Regression of the model (1)

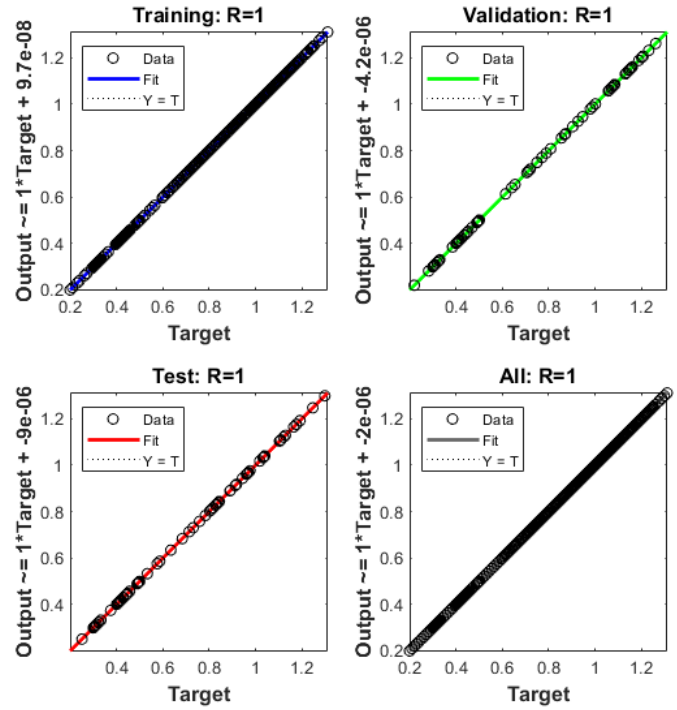


Figure 6.8: Regression of the model (2)

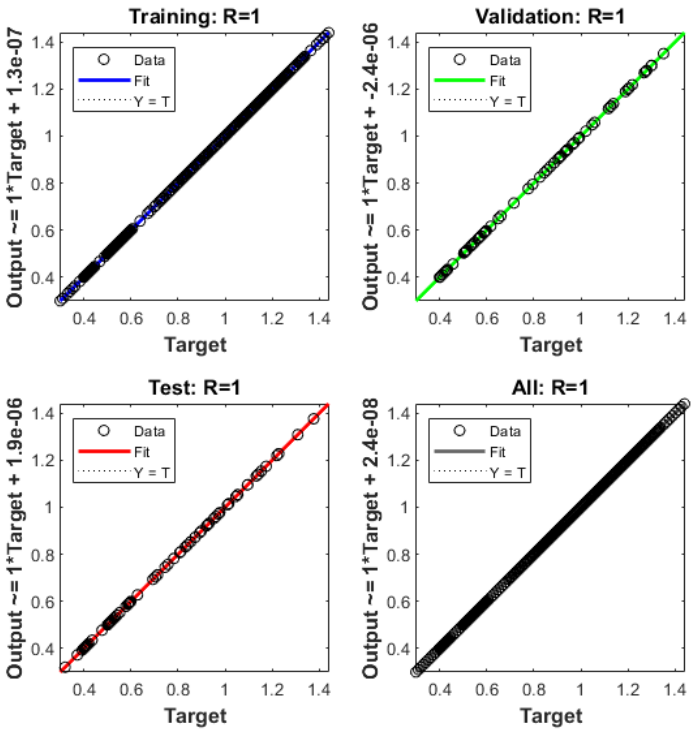
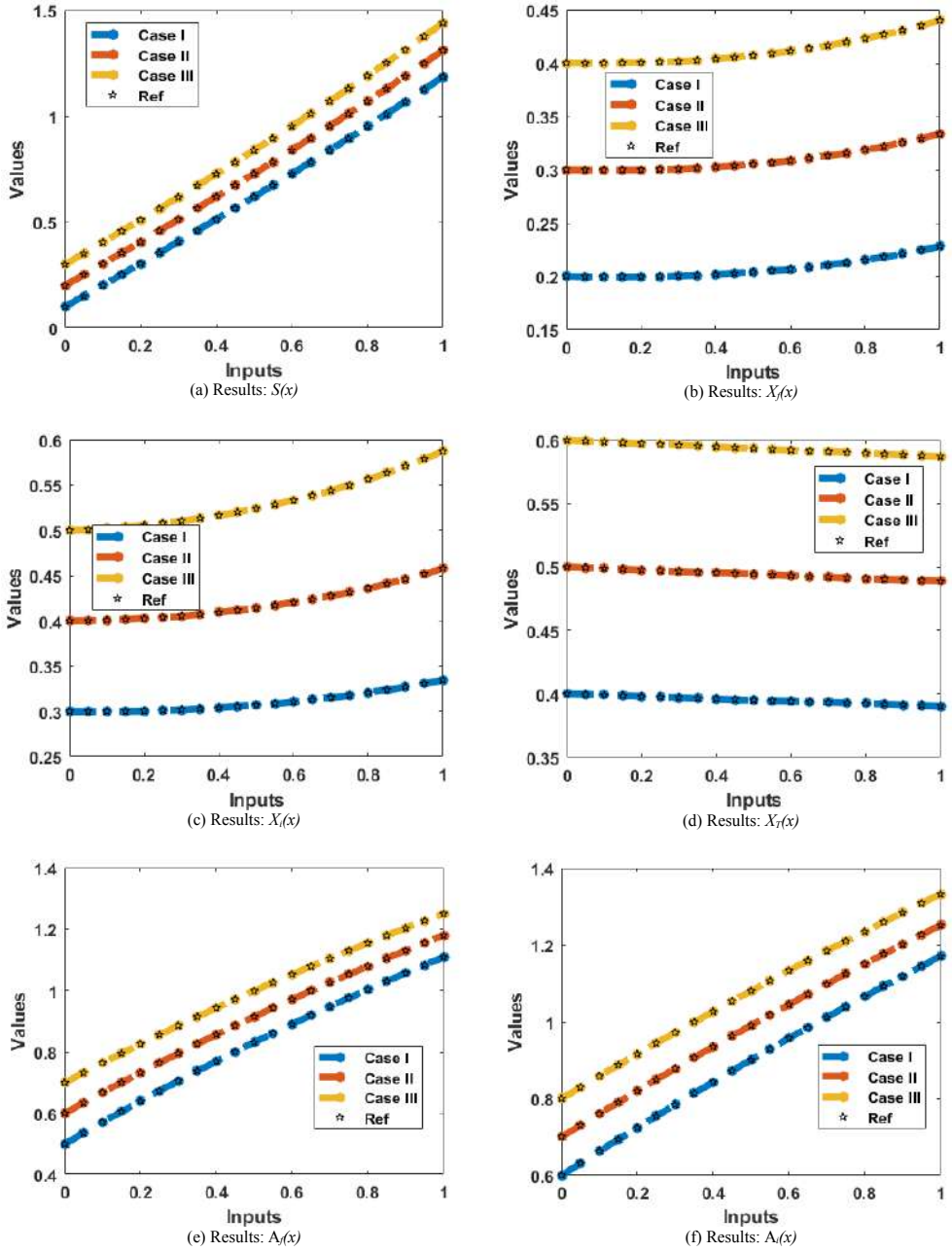


Figure 6.9: Regression of the model (3)

The comparisons of the obtained numerical solutions obtained through the proposed stochastic computational procedure of the ANN-LMBP technique are represented in Figures 6.10. It is shown in Figures-6.10 (a) that the number of susceptible individuals $S(x)$ in the population at time x for each numerical case of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4 overlaps with the obtained reference solution.



(a) Results: $S(x)$ (b) Results: $X_I(x)$

(c) Results: $X_I(x)$ (d) Results: $X_{II}(x)$

(e) Results: $A_I(x)$ (f) Results: $A_{II}(x)$

Figure 6.10: Results comparison for solving the allergen sensitization occurrence model

Similarly, it is also shown in Figure 6.10 (b-f) that the number of individuals sensitized to food allergens $X_f(x)$, amount of food allergens A_f, number of inhalant allergens A_i and proportion of individuals $T(x)$ in the population at time x for each numerical case of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4 are overlapping with the obtained reference solution, respectively. It is also shown that each generation of the nonlinear dynamical APP mathematical model shown by equations-6.2–refc6Eq4 and their allergen sensitization events overlap with the reference and results are carried out.

Finally, the obtained results and their achieved performance, along with the values of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4 along with the performances of absolute error (AE) are represented in Figure 6.11. As it is already observed in Figure 6.10 the overlapping of the solutions is performed for all three numerical cases of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4, respectively. This matching of the results enhances the precision and correctness of the proposed stochastic computational procedure of the ANN-LMBP technique.

The AEs of all three numerical cases are plotted for the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4. It is shown that the number of susceptible individuals $S(x)$ in the population at time x for each numerical case of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4 are overlapping with the obtained reference solution.

Similarly, it is also shown in Figure 6.10 (b-f) that the number of individuals sensitized to food allergens $X_f(x)$, the amount of food allergens A_f, the number of inhalant allergens A_i and the proportion of individuals $T(x)$ in the population at time x for each numerical case of the nonlinear dynamical APP mathematical model represented by equations-6.2–6.4 are performed in the range of 10^{-06} to 10^{-07} and 10^{-06} to 10^{-07}, overlapping with the obtained reference solution respectively for each case are illustrated in Figure 6.11 (a-f).

The AEs' small error ranges almost do not matter, and they prove that the proposed stochastic computing procedure of the ANN-LMBP technique for solving the nonlinear dynamical APP mathematical model works perfectly. Thus, it is observed that the present stochastic computational procedure of the ANN-LMBP technique is specific to solving the nonlinear dynamical APP mathematical model to predict the allergen sensitization occurrence system for each generalized class.

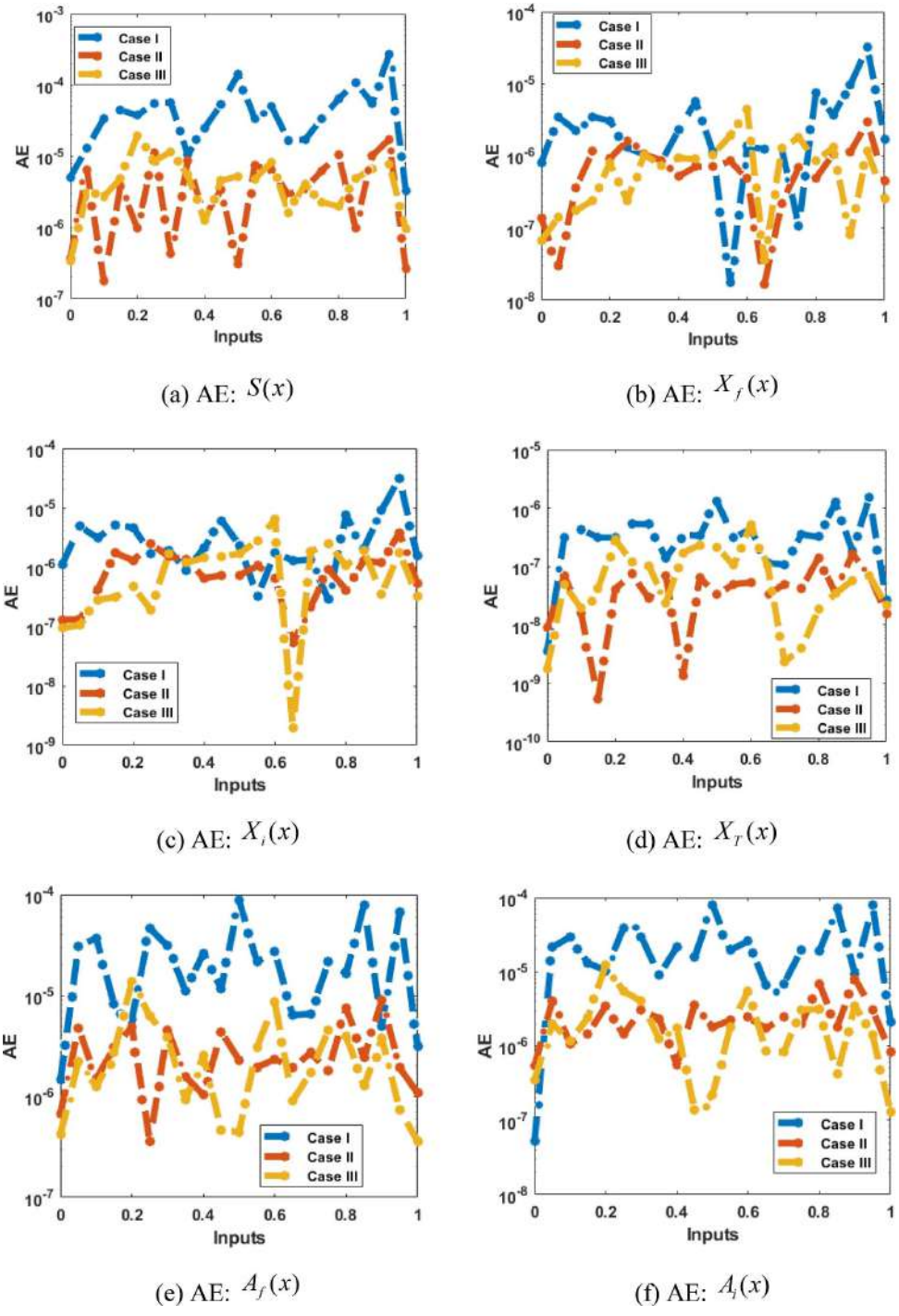


Figure 6.11: Performances of AE for the allergen sensitization occurrence model

6.3.2 Comparative Analysis

A comparative study of various state-of-the-art literature techniques with the proposed mechanism are given in Table-6.3.

Table 6.3: Comparison of Proposed Mechanism with Related Work

Reference	**Dataset**	**Contribution**	**Result**	**Remarks**
U Barman et al. [148]	360 Images acquired using smartphone and torchlight	Artificial Neural Network (ANN) and Linear regression (LR) were used to predict digital chlorophyll meter.	ANN outperformed and performed effectively in predicting the chlorophyll content of citrus leaves.	Only applicable to citrus leaves not on all plants. Moreover, the same size is there. Effect of environmental factors needs to be considered.
TK Dash et al. [149]	A large speech signal dataset	Proposed speech enhancement model using Convolutional Neural Network (CNN)	CNN outperformed the existing state-of-the-art methods effectively.	Complex CNN architectures and different loss optimizations need to be considered.
TK Dash et al. [150]	720 noisy speech files	A six-step evaluation strategy is proposed to investigate the effect of input features on the noise level classification in speech enhancement algorithms.	FSVN performed best and identified important features with highest accuracy.	Future research should focus on exploring the full potential of noise level estimation and classification in speech processing.
SJ white et al. [151]	Whole protein extracts	Utilized mathematical modeling of profiling data for allergen-specific equine IgE in serum against a wide range of putative allergenic proteins.	Identified a wide range of putative allergenic proteins, including novel pollen, bacteria, mold, and arthropod proteins, that are sign.	The methodology and algorithms used in this work provide a valuable framework for future research in this area.
Elisa et al. [152]	Magnetic microbeads and chemiluminescence	Developed a origami paper-based biosensor to detect OVA using a chemiluminescence immunoassay.	The presented biosensor detected OVA in buffer solutions with a limit of 0.1 ng/mL and a linear range of 0.1–10 ng/mL	Can be used for the detection of protein allergens in future.

Reference	Dataset	Contribution	Result	Remarks
TK Dash et al. [153]	Noizeus and spear database	Improved hybrid speech enhancement scheme integrating ANN and firefly algorithm.	The proposed scheme performed well, thus enhancing denoising performance at low SNR conditions and improving speech intelligibility	Other deep learning and image processing techniques can be used for further enhancements.
Ayse Nur Akkilic et al. [154]	Variations of the natural death rate of humans.	Novel approach to improve the accuracy of the monkeypox transmission mathematical model by using a deep neural network process with radial basis activation function and Bayesian regularization optimization.	Improves the accuracy of the monkeypox transmission mathematical model.	Can be used for the detection of protein allergens in future.
Zulqurnain Sabir et al. [155]	105 decision/ optimization variables in the objective function to solve the SEIR based spreading of the Zika virus system.	Proposed a novel computational approach using artificial neural network procedures and swarming procedure to solve the SEIR mathematical model for the spreading of Zika virus.	Swarming neural network computing was proved to be more effective and accurate than conventional numerical approaches, and the results produced using this method were compared to those acquired using these methods. The scientists speculate that expanding this method to additional infectious illnesses might have major public health ramifications.	Further validation and testing of the proposed approach is needed to fully assess its effectiveness and potential for application to other infectious diseases.
Proposed Work	Variations of the allergy causing parameters.	A stochastic computational model for allergy prevalence prediction using ANN+LMBP	The proposed mechanism achieved minimal absolute error, in a range from 10^{-6} to 10^{-7}, and effectively performs allergy prevalence prediction.	The mechanism is first of its kind to solve the nonlinear allergen sensitization prediction problem.

6.4 Summary

The work presented in this chapter highlighted the numerical simulations based on the stochastic paradigms for solving the mathematical model with its six compartments to forecast the allergy occurrence. The human population was divided into four classes along with two allergen classes (food and inhalants). These compartments of the allergen forms was designated to the nonlinear system, which was used to predict the allergen sensitization occurrence. A few concluding remarks on this work are presented as follows:

- The numerical performances of the system was presented through the stochastic computational artificial neural network along with the Levenberg-Marquardt back-propagation.
- Three kinds of data were selected in the process of neural network for the nonlinear system that was chosen as 72%, training, and 14% for both testing and authentication.
- Nineteen numbers of hidden neurons along with log-sigmoid function as an activation was selected in this study.
- The correctness of the stochastic scheme was observed through the comparison of the proposed and reference results, while reducible AE represents the precision of the designed approach.
- The consistency of the proposed approach was observed by performing different statistical operators.

Chapter 7

Conclusions and Future Work

The overall findings of the study conducted in this thesis are presented in this chapter. There is also a discussion of recommendations for potential expansions of the study and future research directions.

7.1 Conclusion

It is critical to ensure that dietary supplements and pharmaceuticals are not allergic and harmful when they are being developed. Allergens, which are substances or components of substances that can induce allergic reactions in humans, may also be a component of substances that target specific diseases. The kind of allergic reaction that an allergen can cause determines how it is classified. Making sure there are no allergies in the ingredients is a top priority in both medicine development and food production. The current methods that can be used to cure allergies are not completely understood, and the only strategy that is known for preventing allergies is to abstain from substances that contain allergens. The work presented in this thesis analysed the allergen protein sequences and their physicochemical characteristics using computational methods and developed an intelligent framework that can detect allergens.

7.1.1 Feature Extraction and Selection

The protein sequences were collected from different sources, which included AllergenOnline, UniProt, SWISS-PROT, etc. The collected protein sequences were then processed for extracting features, that could be used to train different machine learning and deep learning models. The features were extracted using the Peptide package available on CRAN. The physical, chemical and other properties were extracted from each protein sequence. Once the features were extracted, then the identification of the most relevant descriptors was implemented using the feature selection mechanism. The mechanism was utilized in order to improve the effectiveness of the classification model and to expand its capacity to differentiate between allergenic and non-allergenic sequences.

7.1.2 Deep Learning based Ensemble Approach for Protein Allergen Classification

A deep learning based ensemble approach to identify protein allergens, using Extra Tree, Deep Belief Network, and CatBoost models, was proposed. The proposed ensemble model achieved higher detection accuracy by combining the prediction results of the three models using majority voting. The evaluation of the proposed model was carried out on the benchmark protein allergen dataset, and the performance analysis revealed that the proposed model outperforms the other state-of-the-art literature techniques with a protein allergen detection accuracy of 89.16%.

7.1.3 Plant and Animal Food Allergen Classification

A food allergen ensemble classification model based on the extra tree classifier ensemble approach, using physicochemical properties, was also proposed. The study used 738 allergen sequences collected from various plants and animals. The study showed that the proposed ensemble framework outperformed traditional as well as various other ensemble machine learning models. The performance of the proposed model was evaluated on various parameters such as Accuracy, ROC, Sensitivity and Specificity. The proposed model achieved an accuracy of 94.23% with 10 fold cross-validation.

7.1.4 Allergy Prevalence Prediction using Stochastic Mathematical Model

The stochastic mathematical model was developed by utilising a stochastic computational artificial neural network and the Levenberg-Marquardt backpropagation. The neural network process involved presented three types of statistics: training, testing, and sampling. The allergy occurrence prediction model had a preferred statistical accuracy of 72% for training data and 14% for both authentication and testing data. The accuracy of the stochastic scheme was evaluated by comparing its performance with the proposed and reference solutions. A small absolute error indicated the scheme's correctness. The reliability of the stochastic solver was also observed by utilizing various statistical operators.

7.2 Future Work

The work done in this thesis primarily focuses on developing a computationally intelligent framework for the classification of protein sequences into allergen or non-allergen and allergen sequences into animal and plant food allergens. Also, for the prediction of allergy

prevalence, a stochastic model has been developed. Some of the suggestions for future work as follows:

- All the approaches for the classification of sequences into allergen or non-allergen and further classification of allergen into various categories can be combined as a multilevel-multimodel framework.
- The Stochastic model can be further enhanced with various real-time parameters in allergen categories.
- An online tool combining all the approaches can be developed.

List of Publications

SCI Journals

1. Arun Kumar, Prashant Singh Rana, *"A deep learning based ensemble approach for protein allergen classification"*, PeerJ Computer Science 9:e1622. (IF: 3.8)

2. Arun Kumar, Prashant Singh Rana, *"Stochastic Performances of the Mathematical System Based on the Prevalence Prediction of Allergies"*, Arabian Journal for Science and Engineering (2024): 1-18, Springer. (IF:2.9)

3. Arun Kumar, Prashant Singh Rana, *"Extra Tree based Ensembling for Food Allergen Classification based on Physicochemical Properties"*, Multimedia Tools and Applications, Springer. [Under Review](IF:3.6)

References

[1] Onyinye I Iweala, Shailesh K Choudhary, and Scott P Commins. Food allergy. *Current gastroenterology reports*, 20:1–6, 2018.

[2] Graham Roberts and Gideon Lack. Food allergy and asthma—what is the link? *Paediatric respiratory reviews*, 4(3):205–212, 2003.

[3] Hugh A Sampson, Liam O'Mahony, A Wesley Burks, Marshall Plaut, Gideon Lack, and Cezmi A Akdis. Mechanisms of food allergy. *Journal of Allergy and Clinical Immunology*, 141(1):11–19, 2018.

[4] Helen A Brough, Kari C Nadeau, Sayantani B Sindher, Shifaa S Alkotob, Susan Chan, Henry T Bahnson, Donald YM Leung, and Gideon Lack. Epicutaneous sensitization in the development of food allergy: what is the evidence and how can this be prevented? *Allergy*, 75(9):2185–2205, 2020.

[5] JM Igea. The history of the idea of allergy. *Allergy*, 68(8):966–973, 2013.

[6] Mark Jackson. *Allergy: the history of a modern malady*. Reaktion Books, 2007.

[7] Jessica Savage, Scott Sicherer, and Robert Wood. The natural history of food allergy. *The Journal of Allergy and Clinical Immunology: In Practice*, 4(2):196–203, 2016.

[8] Tom Platts-Mills, Allison Worth, and Aziz Sheikh. From the 18th to the 21st century, and beyond... *Landmark Papers in Allergy: Seminal Papers in Allergy with Expert Commentaries*, page 307, 2013.

[9] A Liu, L Fanning, H Chong, J Fernandez, D Sloane, M Sancho-Serra, and M Castells. Desensitization regimens for drug allergy: state of the art in the 21st century. *Clinical & Experimental Allergy*, 41(12):1679–1689, 2011.

[10] Carole Ober and Tsung-Chieh Yao. The genetics of asthma and allergic disease: a 21st century perspective. *Immunological reviews*, 242(1):10–30, 2011.

[11] Arianna Cafarotti, Mattia Giovannini, Philippe Begìn, Helen A Brough, and Stefania Arasi. Management of ige-mediated food allergy in the 21st century. *Clinical & Experimental Allergy*, 53(1):25–38, 2023.

[12] J Pepys. Types of allergic reaction. *Clinical & Experimental Allergy*, 3:491–506, 1973.

[13] B Niggemann and K Beyer. Factors augmenting allergic reactions. *Allergy*, 69(12):1582–1587, 2014.

[14] Frédérique Maindrault-Goebel, Thierry André, Christophe Tournigand, Christophe Louvet, Nathalie Perez-Staub, Nora Zeghib, and Aimery De Gramont. Allergic-type

reactions to oxaliplatin: retrospective analysis of 42 patients. *European Journal of Cancer*, 41(15):2262–2267, 2005.

[15] Brian A Baldo, Nghia H Pham, Brian A Baldo, and Nghia H Pham. Classification and descriptions of allergic reactions to drugs. *Drug allergy: clinical aspects, diagnosis, mechanisms, structure-activity relationships*, pages 17–57, 2021.

[16] Diego G Peroni, Giulia Nuzzi, Irene Trambusti, Maria Elisa Di Cicco, and Pasquale Comberiati. Microbiome composition and its impact on the development of allergic diseases. *Frontiers in immunology*, 11:700, 2020.

[17] Boudewijn JH Dierick, Thys van der Molen, Bertine MJ Flokstra-de Blok, Antonella Muraro, Maarten J Postma, Janwillem WH Kocks, and Job FM van Boven. Burden and socioeconomics of asthma, allergic rhinitis, atopic dermatitis and food allergy. *Expert review of pharmacoeconomics & outcomes research*, 20(5):437–453, 2020.

[18] Scott H Sicherer and Hugh A Sampson. Food allergy: a review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *Journal of Allergy and Clinical Immunology*, 141(1):41–58, 2018.

[19] Engineering National Academies of Sciences, Medicine, et al. Finding a path to safety in food allergy: assessment of the global burden, causes, prevention, management, and public policy. 2017.

[20] Capucine Picard, H Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Yanick J Crow, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, et al. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. *Journal of clinical immunology*, 38:96–128, 2018.

[21] Carolin Anthes and OLIVIER De Schutter. The food and agriculture organization of the united nations. *Human Rights in Global Health: Rights-Based Governance for a Globalizing World*, page 261, 2018.

[22] Dafydd Pilling, Julie Bélanger, and Irene Hoffmann. Declining biodiversity for food and agriculture needs urgent global action. *Nature Food*, 1(3):144–147, 2020.

[23] Victoria Cardona, Ignacio J Ansotegui, Motohiro Ebisawa, Yehia El-Gamal, Montserrat Fernandez Rivas, Stanley Fineman, Mario Geller, Alexei Gonzalez-Estrada, Paul A Greenberger, Mario Sanchez Borges, et al. World allergy organization anaphylaxis guidance 2020. *World allergy organization journal*, 13(10):100472, 2020.

[24] Theresa Bingemann, Hemant Sharma, Anil Nanda, David A Khan, Sharon Markovics, James Sussman, Andrew Murphy, Denise Kalman, and Giselle Mosnaim. Aaaai work group report: physician wellness in allergy and immunology. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(4):1224–1229, 2020.

[25] Christoph Bergmann, Aurélie Poli, Ioana Agache, Rodolfo Bianchini, Heather J Bax, Mariana Castells, Silvia Crescioli, David Dombrowicz, Denisa Ferastraoaru, Edda Fiebiger, et al. Allergooncology: Danger signals in allergology and oncology: A european academy of allergy and clinical immunology (eaaci) position paper. *Allergy*, 77(9):2594–2617, 2022.

[26] Joanna S Zeiger, William S Silvers, Tonya A Winders, Mary K Hart, and Robert S Zeiger. Cannabis attitudes and patterns of use among followers of the allergy & asthma network. *Annals of Allergy, Asthma & Immunology*, 126(4):401–410, 2021.

[27] AK Ellis. Anne k. ellis. *A Practical Guide*, page 123.

[28] S Scheurer, M Toda, and S Vieths. What makes an allergen? *Clinical & Experimental Allergy*, 45(7):1150–1161, 2015.

[29] Ethem Şahin, Sameer Ali Bafaqeeh, Selis Gülseven Güven, Erdem Atalay Çetinkaya, Nuray Bayar Muluk, Zerrin Ozergin Coşkun, Andrey Lopatin, Murat Kar, Mehmet Ozgur Pinarbasli, and Cemal Cingi. Mechanism of action of allergen immunotherapy. *American journal of rhinology & allergy*, 30(5_suppl):S1–S3, 2016.

[30] AB Singh and Pawan Kumar. Common environmental allergens causing respiratory allergy in india. *The Indian Journal of Pediatrics*, 69:245–250, 2002.

[31] Rudolf Valenta, Heidrun Hochwallner, Birgit Linhart, and Sandra Pahr. Food allergies: the basics. *Gastroenterology*, 148(6):1120–1131, 2015.

[32] Heimo Breiteneder and EN Clare Mills. Molecular properties of food allergens. *Journal of Allergy and Clinical Immunology*, 115(1):14–23, 2005.

[33] Steffie de Gier and Kitty Verhoeckx. Insect (food) allergy and allergens. *Molecular immunology*, 100:82–106, 2018.

[34] Adnan Custovic and A Simpson. 1 the role of inhalant allergens in allergic airways disease. *Journal of investigational allergology and clinical immunology*, 22(6):393, 2012.

[35] I Kimber, DA Basketter, M Butler, A Gamer, J-L Garrigue, GF Gerberick, C Newsome, W Steiling, and H-W Vohr. Classification of contact allergens according to potency: proposals. *Food and Chemical Toxicology*, 41(12):1799–1809, 2003.

[36] Anna Pomés, Janet M Davies, Gabriele Gadermaier, Christiane Hilger, Thomas Holzhauser, Jonas Lidholm, Andreas L Lopata, Geoffrey A Mueller, Andreas Nandy, Christian Radauer, et al. Who/iuis allergen nomenclature: Providing a common language. *Molecular immunology*, 100:3–13, 2018.

[37] Martin D Chapman. Allergen nomenclature. *Allergens and allergen immunotherapy*, pages 65–76, 2008.

[38] Giuliana Ferrante, Amelia Licari, Salvatore Fasola, Gian Luigi Marseglia, and Stefania La Grutta. Artificial intelligence in the diagnosis of pediatric allergic diseases.

Pediatric Allergy and Immunology, 32(3):405–413, 2021.

[39] Mattia CF Prosperi, Danielle Belgrave, Iain Buchan, Angela Simpson, and Adnan Custovic. Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach. *Pediatric Allergy and Immunology*, 25(1):71–79, 2014.

[40] Liyang Wang, Dantong Niu, Xinjie Zhao, Xiaoya Wang, Mengzhen Hao, and Huilian Che. A comparative analysis of novel deep learning and ensemble learning models to predict the allergenicity of food proteins. *Foods*, 10(4):809, 2021.

[41] Manju Bhardwaj, Debasis Dash, and Vasudha Bhatnagar. Accurate classification of biological data using ensembles. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1486–1493. IEEE, 2015.

[42] Fernando Vadillo. Comparing stochastic lotka–volterra predator-prey models. *Applied Mathematics and Computation*, 360:181–189, 2019.

[43] Sandra Dedrick, Vaishnavi Warrier, Katherine P Lemon, and Babak Momeni. When does a lotka-volterra model represent microbial interactions? insights from in vitro nasal bacterial communities. *Msystems*, pages e00757–22, 2023.

[44] Bin Liu, Ziman Yang, Qing Liu, Ying Zhang, Hui Ding, Hongyan Lai, and Qun Li. Computational prediction of allergenic proteins based on multi-feature fusion. *Frontiers in Genetics*, 14:1294159, 2023.

[45] Xin-Xin Yu, Meng-Qi Liu, Xiao-Yan Li, Ying-Hua Zhang, and Bing-Jie Tao. Qualitative and quantitative prediction of food allergen epitopes based on machine learning combined with in vitro experimental validation. *Food Chemistry*, 405:134796, 2023.

[46] Chinna babu Mupparaju, Bhavishya Katta, Venkata Chaitanya Kumar Suram, Narsimha Chary Rajampetakasham, and Mounika Nalluri. Food allergy detection with ml: Developing machine learning algorithms to detect and prevent allergens in food products. *Tuijin Jishu/Journal of Propulsion Technology*, 44(5):2479–2488, 2023.

[47] Justin Zhang, Deborah Lee, Kylie Jungles, Diane Shaltis, Kayvan Najarian, Rajan Ravikumar, Georgiana Sanders, and Jonathan Gryak. Prediction of oral food challenge outcomes via ensemble learning. *Informatics in Medicine Unlocked*, 36:101142, 2023.

[48] Miroslava Nedyalkova, Mahdi Vasighi, Amirreza Azmoon, Ludmila Naneva, and Vasil Simeonov. Sequence-based prediction of plant allergenic proteins: Machine learning classification approach. *ACS Omega*, 2023.

[49] Pınar Cihan and Zeynep Banu Ozger. A new approach for determining sars-cov-2 epitopes using machine learning-based in silico methods. *Computational Biology*

and Chemistry, 98:107688, 2022.

[50] Ramisetty Kavya, Jabez Christopher, Subhrakanta Panda, and Y Bakthasingh Lazarus. Machine learning and xai approaches for allergy diagnosis. *Biomedical Signal Processing and Control*, 69:102681, 2021.

[51] Xiaodi Yang, Shiping Yang, Qinmengge Li, Stefan Wuchty, and Ziding Zhang. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Computational and structural biotechnology journal*, 18:153–161, 2020.

[52] Sevinç İlhan Omurca, Ekin Ekinci, Bengisu Çakmak, and Selin Gizem Özkan. Using machine learning approaches for prediction of the types of asthmatic allergy across the turkey. *Data Science and Applications*, 2(2):8–12, 2019.

[53] Joost Westerhout, Tanja Krone, Almar Snippe, Lilia Babe, Scott McClain, Gregory S Ladics, Geert F Houben, and Kitty CM Verhoeckx. Allergenicity prediction of novel and modified proteins: Not a mission impossible! development of a random forest allergenicity prediction model. *Regulatory Toxicology and Pharmacology*, 107:104422, 2019.

[54] Shahid Akbar, Maqsood Hayat, Muhammad Iqbal, and Mian Ahmad Jan. iacp-gaensc: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artificial intelligence in medicine*, 79:62–70, 2017.

[55] Shahid Akbar, Ashfaq Ahmad, Maqsood Hayat, Ateeq Ur Rehman, Salman Khan, and Farman Ali. iatbp-hyb-enc: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. *Computers in Biology and Medicine*, 137:104778, 2021.

[56] Shahid Akbar, Heba G Mohamed, Hashim Ali, Aamir Saeed, Aftab Ahmed, Sarah Gul, Ashfaq Ahmad, Farman Ali, Yazeed Yasin Ghadi, and Muhammad Assam. Identifying neuropeptides via evolutionary and sequential based multi-perspective descriptors by incorporation with ensemble classification strategy. *IEEE Access*, 2023.

[57] Zubaida Said Ameen, Hala Mostafa, Dilber Uzun Ozsahin, and Auwalu Saleh Mubarak. Accelerating sars-cov-2 vaccine development: Leveraging novel hybrid deep learning models and bioinformatics analysis for epitope selection and classification. *Processes*, 11(6):1829, 2023.

[58] Yiyun Zhang, Zijian Dai, Xinjie Zhao, Changyu Chen, Siqi Li, Yantong Meng, Zhuoma Suonan, Yuge Sun, Qun Shen, Liyang Wang, et al. Deep learning drives efficient discovery of novel antihypertensive peptides from soybean protein isolate. *Food Chemistry*, 404:134690, 2023.

[59] Mandana Behbahani, Parisa Rabiei, and Hassan Mohabatkar. A comparative analysis of allergen proteins between plants and animals using several computational tools and chou's pseaac concept. *International Archives of Allergy and Immunology*, 181(11):813–821, 2020.

[60] Ahmed A Metwally, Philip S Yu, Derek Reiman, Yang Dai, Patricia W Finn, and David L Perkins. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via long short-term memory networks. *PLoS computational biology*, 15(2):e1006693, 2019.

[61] Zhenjiao Du, Yixiang Xu, Changqi Liu, and Yonghui Li. plm4alg: Protein language model-based predictors for allergenic proteins and peptides. *Journal of Agricultural and Food Chemistry*, 2023.

[62] Muhammad Attique, Tamim Alkhalifah, Fahad Alturise, and Yaser Daanial Khan. Deepbce: Evaluation of deep learning models for identification of immunogenic b-cell epitopes. *Computational Biology and Chemistry*, 104:107874, 2023.

[63] Chun He, Xinhai Ye, Yi Yang, Liya Hu, Yuxuan Si, Xianxin Zhao, Longfei Chen, Qi Fang, Ying Wei, Fei Wu, et al. Deepalgpro: an interpretable deep neural network model for predicting allergenic proteins. *Briefings in Bioinformatics*, 24(4):bbad246, 2023.

[64] Wajdi Alghamdi, Muhammad Attique, Ebraheem Alzahrani, Malik Zaka Ullah, and Yaser Daanial Khan. Lbcepred: a machine learning model to predict linear b-cell epitopes. *Briefings in Bioinformatics*, 23(3):bbac035, 2022.

[65] Minh N Nguyen, Nora L Krutz, Vachiranee Limviphuvadh, Andreas L Lopata, G Frank Gerberick, and Sebastian Maurer-Stroh. Allercatpro 2.0: a web server for predicting protein allergenicity potential. *Nucleic Acids Research*, 50(W1):W36–W43, 2022.

[66] Francisco M Garcia-Moreno and Miguel A Gutiérrez-Naranjo. Allerdet: A novel web app for prediction of protein allergenicity. *Journal of Biomedical Informatics*, 135:104217, 2022.

[67] Robin Pearce, Yang Li, Gilbert S Omenn, and Yang Zhang. Fast and accurate ab initio protein structure prediction using deep learning potentials. *PLoS Computational Biology*, 18(9):e1010539, 2022.

[68] Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Akshara Pande, Chakit Arora, and Gajendra PS Raghava. Algpred 2.0: an improved method for predicting allergenic proteins and mapping of ige epitopes. *Briefings in Bioinformatics*, 22(4):bbaa294, 2021.

[69] Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Naorem Leimarembi Devi, and Gajendra PS Raghava. Chalpred: A web server for prediction of allergenicity of

chemical compounds. *Computers in Biology and Medicine*, 136:104746, 2021.

[70] Sebastian Maurer-Stroh, Nora L Krutz, Petra S Kern, Vithiagaran Gunalan, Minh N Nguyen, Vachiranee Limviphuvadh, Frank Eisenhaber, and G Frank Gerberick. Allercatpro—prediction of protein allergenicity potential from the protein sequence. *Bioinformatics*, 35(17):3020–3027, 2019.

[71] Fabio Dall'Antonia and Walter Keller. Spade web service for prediction of allergen ige epitopes. *Nucleic Acids Research*, 47(W1):W496–W501, 2019.

[72] Amadeo Sena-Torralba, Yeray Pallás-Tamarit, Sergi Morais, and Ángel Maquieira. Recent advances and challenges in food-borne allergen detection. *TrAC Trends in Analytical Chemistry*, 132:116050, 2020.

[73] Jae Hwan Shin, Y Veera Manohara Reddy, Tae Jung Park, and Jong Pil Park. Recent advances in analytical strategies and microsystems for food allergen detection. *Food chemistry*, 371:131120, 2022.

[74] Jaison Jeevanandam, Dominic Agyei, Michael K Danquah, and Chibuike Udenigwe. Food quality monitoring through bioinformatics and big data. In *Future Foods*, pages 733–744. Elsevier, 2022.

[75] Ivan Dimitrov, Ivan Bangov, Darren R Flower, and Irini Doytchinova. Allertop v. 2—a server for in silico prediction of allergens. *Journal of molecular modeling*, 20:1–6, 2014.

[76] Lei Zhou, Chu Zhang, Fei Liu, Zhengjun Qiu, and Yong He. Application of deep learning in food: a review. *Comprehensive reviews in food science and food safety*, 18(6):1793–1811, 2019.

[77] Piyush Bhardwaj, Ashish Tyagi, Shashank Tyagi, Joana Antão, and Qichen Deng. Machine learning model for classification of predominantly allergic and non-allergic asthma among preschool children with asthma hospitalization. *Journal of Asthma*, 60(3):487–495, 2023.

[78] Richard E Goodman, Motohiro Ebisawa, Fatima Ferreira, Hugh A Sampson, Ronald van Ree, Stefan Vieths, Joseph L Baumert, Barbara Bohle, Sreedevi Lalithambika, John Wise, et al. Allergenonline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Molecular nutrition & food research*, 60(5):1183–1198, 2016.

[79] David L Wheeler, Deanna M Church, Scott Federhen, Alex E Lash, Thomas L Madden, Joan U Pontius, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Tatiana A Tatusova, et al. Database resources of the national center for biotechnology. *Nucleic acids research*, 31(1):28–33, 2003.

[80] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele

Magrane, et al. The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl_1):D154–D159, 2005.

[81] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

[82] Atsushi Ikai. Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry*, 88(6):1895–1898, 1980.

[83] R Graphical Manual and HR Documentation. 38 hydrophobicity values for amino acids from various sources.

[84] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4:23–55, 1985.

[85] Li Yang, Mao Shu, Kaiwang Ma, Hu Mei, Yongjun Jiang, and Zhiliang Li. St-scale as a novel amino acid descriptor and its application in qsam of peptides and analogues. *Amino acids*, 38:805–816, 2010.

[86] HU Mei, Zhi H Liao, Yuan Zhou, and Shengshi Z Li. A new set of amino acid descriptors and its application in peptide qsars. *Peptide Science: Original Research on Biomolecules*, 80(6):775–786, 2005.

[87] Maria Sandberg, Lennart Eriksson, Jörgen Jonsson, Michael Sjöström, and Svante Wold. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *Journal of medicinal chemistry*, 41(14):2481–2491, 1998.

[88] Gabriele Cruciani, Massimo Baroni, Emanuele Carosati, Monica Clementi, Roberta Valigi, and Sergio Clementi. Peptide studies by means of principal properties of amino acids derived from mif descriptors. *Journal of Chemometrics*, 18(3-4):146–155, 2004.

[89] Andrea Zaliani and Emanuela Gancia. Ms-whim scores for amino acids: a new 3d-description for peptide qsar and qspr studies. *Journal of chemical information and computer sciences*, 39(3):525–533, 1999.

[90] Alexander G Georgiev. Interpretable numerical descriptors of amino acid space. *Journal of Computational Biology*, 16(5):703–723, 2009.

[91] Hans G Boman. Antibacterial peptides: basic facts and emerging concepts. *Journal of internal medicine*, 254(3):197–215, 2003.

[92] Enrique Audain, Yassel Ramos, Henning Hermjakob, Darren R Flower, and Yasset Perez-Riverol. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. *Bioinformatics*, 32(6):821–827, 2016.

[93] Bengt Bjellqvist, Graham J Hughes, Christian Pasquali, Nicole Paquet, Florence Ravier, Jean-Charles Sanchez, Séverine Frutiger, and Denis Hochstrasser. The

focusing positions of polypeptides in immobilized ph gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14(1):1023–1031, 1993.

[94] Rex Malcolm Chaplin Dawson, Daphne C Elliott, William H Elliott, and Kenneth M Jones. *Data for biochemical research*. Clarendon press, 2002.

[95] Gisela Gabernet, Alex T Müller, Jan A Hiss, and Gisbert Schneider. Membranolytic anticancer peptides. *MedChemComm*, 7(12):2232–2245, 2016.

[96] Albert L Lehninger, David L Nelson, Michael M Cox, et al. *Lehninger principles of biochemistry*. Macmillan, 2005.

[97] Robert K Murray, Daryl K Granner, Peter A Mayes, and Victor W Rodwell. Harper's illustrated biochemistry. 2003.

[98] John D Rodwell. Heterogeneity of component bands in isoelectric focusing patterns. *Analytical biochemistry*, 119(2):440–449, 1982.

[99] Antonio Sillero and Andres Maldonado. Isoelectric point determination of proteins and other macromolecules: oscillating method. *Computers in biology and medicine*, 36(2):157–166, 2006.

[100] TW Graham Solomons and Craig B Fryhle. *Organic chemistry*. John Wiley & Sons, 2008.

[101] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.

[102] Gerard JP van Westen, Remco F Swier, Jörg K Wegner, Adriaan P IJzerman, Herman WT van Vlijmen, and Andreas Bender. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *Journal of cheminformatics*, 5(1):1–11, 2013.

[103] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[104] Noel Lopes, Bernardete Ribeiro, Noel Lopes, and Bernardete Ribeiro. Deep belief networks (dbns). *Machine Learning for Adaptive Many-Core Machines-A Practical Approach*, pages 155–186, 2015.

[105] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[106] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[107] Robi Polikar. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34, 2012.

[108] Ivan Dimitrov, Darren R Flower, and Irini Doytchinova. Allertop-a server for in silico prediction of allergens. In *BMC bioinformatics*, volume 14, pages 1–9. BioMed

Central, 2013.

[109] Sudipto Saha and Gajendra Pal Singh Raghava. Algpred: prediction of allergenic proteins and mapping of ige epitopes. *Nucleic acids research*, 34(suppl_2):W202–W209, 2006.

[110] Ivan Dimitrov, Lyudmila Naneva, Irini Doytchinova, and Ivan Bangov. Allergenfp: allergenicity prediction by descriptor fingerprints. *Bioinformatics*, 30(6):846–851, 2014.

[111] Pablo Caballero, Guadalupe Ortiz, Alfonso Garcia-de Prado, and Juan Boubeta-Puig. Paving the way to collaborative context-aware mobile applications: a case study on preventing worsening of allergy symptoms. *Multimedia Tools and Applications*, 80(14):21101–21133, 2021.

[112] Robert A Wood. Food allergen immunotherapy: current status and prospects for the future. *Journal of Allergy and Clinical Immunology*, 137(4):973–982, 2016.

[113] Rod A Herman, Ping Song, Henry P Mirsky, and Jason M Roper. Evidence-based regulations for bioinformatic prediction of allergen cross-reactivity are needed. *Regulatory Toxicology and Pharmacology*, 120:104841, 2021.

[114] Mark WEJ Fiers, Gijs A Kleter, Herman Nijland, Ad ACM Peijnenburg, Jan Peter Nap, and Roeland CHJ Van Ham. Allermatch™, a webtool for the prediction of potential allergenicity according to current fao/who codex alimentarius guidelines. *BMC bioinformatics*, 5:1–6, 2004.

[115] Srinidhi Sudharson, Tanja Kalic, Christine Hafner, and Heimo Breiteneder. Newly defined allergens in the who/iuis allergen nomenclature database during 01/2019-03/2021. *Allergy*, 76(11):3359–3373, 2021.

[116] Somya Goyal. Heterogeneous stacked ensemble classifier for software defect prediction. In *2020 sixth international conference on parallel, distributed and grid computing (PDGC)*, pages 126–130. IEEE, 2020.

[117] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[118] Jun-Mo Jo. Effectiveness of normalization pre-processing of big data to the machine learning performance. *The Journal of the Korea institute of electronic communication sciences*, 14(3):547–552, 2019.

[119] Azlyna Senawi, Hua-Liang Wei, and Stephen A Billings. A new maximum relevance-minimum multicollinearity (mrmmc) method for feature selection and ranking. *Pattern Recognition*, 67:47–61, 2017.

[120] Saqib E Awan, Mohammed Bennamoun, Ferdous Sohel, Frank M Sanfilippo, Benjamin J Chow, and Girish Dwivedi. Feature selection and transformation by ma-

chine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PloS one*, 14(6):e0218760, 2019.

[121] A Yongquan Yang, B Haijun Lv, C Ning Chen, D Yang Wu, and E Zhongxi Zheng. Ftbme: feature transferring based multi-model ensemble. *Multimedia Tools and Applications*, 79:18767–18799, 2020.

[122] Alfred V Aho and John E Hopcroft. *The design and analysis of computer algorithms*. Pearson Education India, 1974.

[123] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[124] Jalal A Nasiri, Nasrollah Moghadam Charkari, and Saeed Jalili. Least squares twin multi-class classification support vector machine. *Pattern Recognition*, 48(3):984–992, 2015.

[125] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[126] Junliang Fan, Xin Ma, Lifeng Wu, Fucang Zhang, Xiang Yu, and Wenzhi Zeng. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management*, 225:105758, 2019.

[127] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[128] Elopy Sibanda. Increasing trend of sensitisation to food and inhalant allergen sources in zimbabwe. *Current Allergy & Clinical Immunology*, 26(4):214–219, 2013.

[129] Ruby Pawankar, Giorgio Walter Canonica, Stephen T Holgate, Richard F Lockey, and Michael S Blaiss. Wao white book on allergy. *Milwaukee, WI: World Allergy Organization*, 3:156–157, 2011.

[130] Claudia L Gray. Food allergy in south africa. *Current allergy and asthma reports*, 17:1–9, 2017.

[131] O Oluwole, OG Arinola, GA Falade, MA Ige, GA Falusi, T Aderemi, D Huo, IO Olopade, and CO Olopade. Allergy sensitization and asthma among 13-14 year old school children in nigeria. *African health sciences*, 13(1):144–153, 2013.

[132] Michael E Levin, Claudia L Gray, and Javier Marrugo. Food allergy: international and developing world perspectives. *Current Pediatrics Reports*, 4:129–137, 2016.

[133] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[134] Caroline Mushayi, Farai Nyabadza, Esther Chigidi, Hope Mataramvura, Lorraine Pfavayi, Simbarashe Rusakaniko, and Elopy Nimele Sibanda. A mathematical model for the prediction of the prevalence of allergies in zimbabwe. *World Allergy*

Organization Journal, 14(7):100555, 2021.

[135] Cristina Lanzas, Kale Davies, Samantha Erwin, and Daniel Dawson. On modelling environmentally transmitted pathogens. *Interface focus*, 10(1):20190056, 2020.

[136] RAM NARESHA and Agraj Tripathi. A nonlinear mathematical model for asthma: effect of environmental pollution. 2009.

[137] Sunil Kumar, Ali Ahmadian, Ranbir Kumar, Devendra Kumar, Jagdev Singh, Dumitru Baleanu, and Mehdi Salimi. An efficient numerical method for fractional sir epidemic model of infectious disease by using bernstein wavelets. *Mathematics*, 8(4):558, 2020.

[138] Robin N Thompson. Epidemiological models are important tools for guiding covid-19 interventions. *BMC medicine*, 18:1–4, 2020.

[139] Zulqurnain Sabir. Stochastic numerical investigations for nonlinear three-species food chain system. *International Journal of Biomathematics*, 15(04):2250005, 2022.

[140] Zulqurnain Sabir, Mohamed R Ali, and R Sadat. Gudermannian neural networks using the optimization procedures of genetic algorithm and active set approach for the three-species food chain nonlinear model. *Journal of Ambient Intelligence and Humanized Computing*, 14(7):8913–8922, 2023.

[141] Tareq Saeed, Zulqurnain Sabir, Mohammed Sh Alhodaly, Hamed H Alsulami, and Yolanda Guerrero Sánchez. An advanced heuristic approach for a nonlinear mathematical based medical smoking model. *Results in Physics*, 32:105137, 2022.

[142] Muhammad Umar, Fazli Amin, Hafiz Abdul Wahab, and Dumitru Baleanu. Unsupervised constrained neural network modeling of boundary value corneal model for eye surgery. *Applied Soft Computing*, 85:105826, 2019.

[143] BO Wang, JF Gomez-Aguilar, Zulqurnain Sabir, Muhammad Asif Zahoor Raja, Wei-Feng Xia, HADI Jahanshahi, Madini O Alassafi, and Fawaz E Alsaadi. Numerical computing to solve the nonlinear corneal system of eye surgery using the capability of morlet wavelet artificial neural networks. *Fractals*, 30(05):2240147, 2022.

[144] Zulqurnain Sabir, Hafiz Abdul Wahab, Mohamed R Ali, and R Sadat. Neuron analysis of the two-point singular boundary value problems arising in the thermal explosion's theory. *Neural Processing Letters*, 54(5):4297–4324, 2022.

[145] Zulqurnain Sabir. Neuron analysis through the swarming procedures for the singular two-point boundary value problems arising in the theory of thermal explosion. *The European Physical Journal Plus*, 137(5):638, 2022.

[146] Kanit Mukdasai, Zulqurnain Sabir, Muhammad Asif Zahoor Raja, R Sadat, Mohamed R Ali, and Peerapongpat Singkibud. A numerical simulation of the fractional order leptospirosis model using the supervise neural network. *Alexandria Engineer-*

ing Journal, 61(12):12431–12441, 2022.

[147] Thongchai Botmart, Zulqurnain Sabir, Muhammad Asif Zahoor Raja, Mohamed R Ali, R Sadat, Ayman A Aly, Ali Saad, et al. A hybrid swarming computing approach to solve the biological nonlinear leptospirosis system. *Biomedical Signal Processing and Control*, 77:103789, 2022.

[148] Utpal Barman and Ridip Dev Choudhury. Smartphone image based digital chlorophyll meter to estimate the value of citrus leaves chlorophyll using linear regression, lmbp-ann and scgbp-ann. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2938–2950, 2022.

[149] Tusar Kanti Dash, Sandeep Singh Solanki, and Ganapati Panda. Multi-objective approach to speech enhancement using tunable q-factor-based wavelet transform and ann techniques. *Circuits, Systems, and Signal Processing*, 40(12):6067–6097, 2021.

[150] TK Dash and SS Solanki. Investigation on the effect of the input features in the noise level classification of noisy speech. 2019.

[151] Samuel J White, Meriel Moore-Colyer, E Marti, D Hannant, Vinzenz Gerber, L Coüetil, EA Richard, and M Alcocer. Antigen array for serological diagnosis and novel allergen identification in severe equine asthma. *Scientific reports*, 9(1):15170, 2019.

[152] Elisa Lazzarini, Andrea Pace, Ilaria Trozzi, Martina Zangheri, Massimo Guardigli, Donato Calabria, and Mara Mirasoli. An origami paper-based biosensor for allergen detection by chemiluminescence immunoassay on magnetic microbeads. *Biosensors*, 12(10):825, 2022.

[153] Tusar Kanti Dash, Sandeep Singh Solanki, and Ganapati Panda. Improved phase aware speech enhancement using bio-inspired and ann techniques. *Analog Integrated Circuits and Signal Processing*, 102:465–477, 2020.

[154] Ayse Nur Akkilic, Zulqurnain Sabir, Shahid Ahmad Bhat, and Hasan Bulut. A radial basis deep neural network process using the bayesian regularization optimization for the monkeypox transmission model. *Expert Systems with Applications*, 235:121257, 2024.

[155] Zulqurnain Sabir, Shahid Ahmad Bhat, Muhammad Asif Zahoor Raja, and Sharifah E Alhazmi. A swarming neural network computing approach to solve the zika virus model. *Engineering Applications of Artificial Intelligence*, 126:106924, 2023.