Development of a Smart Clinical Decision Support System for Screening of Celiac Disease using AI

By

Rimsha Mallhi

Masters of Science in Bioinformatics (2021)

MS-BI-00000360812

Supervisor

Dr. Zamir Hussain

SCHOOL OF INTERDISCIPLINARY ENGINEERING & SCIENCES

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY(NUST)

ISLAMABAD, PAKISTAN

AUGUST,2023

Development of smart
Clinical decision support system for screening of celiac disease using AI

By

Rimsha Mallhi

Masters of Science in Bioinformatics (2021)

MS-BI-00000360812

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Bioinformatics

Thesis Supervisor:

Dr. Zamir Hussain

SCHOOL OF INTERDISCIPLINARY ENGINEERING & SCIENCES

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY(NUST)

ISLAMABAD, PAKISTAN

Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms **Rimsha Mallhi** Registration No. **00000360812** of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: ____________

Name of Supervisor: Dr. ZAMIR HUSSAIN

Date: Aug 23, 2023

Signature of HoD with stamp: ____________

Dr. Fouzia Malik
HoD Sciences
Associate Professor
SINES - NUST, Sector H-12
Islamabad

Date: 24-8-2023

Countersign by

Signature (Dean/Principal): ____________

Associate Professor
SINES - NUST, Sector H-12
Islamabad

Date: 25/8/23

Declaration

I Rimsha Mallhi, certify that this research work titled *"Development of smart clinical decision support system for screening of celiac disease using AI"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Rimsha Mallhi
Masters of Science in Bioinformatics (2021)
MS-BI-00000360812

SCHOOL OF INTERDISCIPLINARY ENGINEERING & SCIENCES
NATIONAL UNIVERSITY OF SCIENCES AND
TECHNOLOGY(NUST)
ISLAMABAD, PAKISTAN

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Rimsha Mallhi
Masters of Science in Bioinformatics (2021)
MS-BI-00000360812

Copyright Statement

Acknowledgements

I am deeply grateful for the guidance and blessings bestowed upon me by Allah Subhana-Watala throughout the course of this thesis. Every new thought and improvement in this work is a testament to His divine guidance. I humbly acknowledge that without His priceless help and direction, I would have been unable to achieve anything.

I extend my heartfelt appreciation to my beloved parents (my father Zulfiqar Ali Mallhi and my mother Mussarat un Nisa), whose firm support and nurturing have been instrumental in shaping me into the person I am today. Their love, encouragement, and sacrifices have been a constant source of strength and inspiration in every aspect of my life. Words cannot express the depth of my gratitude for their unwavering belief in me and their endless sacrifices to provide me with the best opportunities. Their presence and belief in me have been the pillars of my success, and I am forever grateful for their unwavering love and support.

I would like to express my sincere gratitude to my supervisor, Dr. Zamir Hussain, for his invaluable assistance and guidance throughout the thesis. His expertise, patience, and dedication have been instrumental in shaping the direction of my research. I am also grateful for his exceptional teaching in the fields of machine learning and data analysis, which have greatly enriched my understanding and knowledge in the context of my research.

I would like to express my heartfelt thanks to Dr. Rehan Zafar Paracha and Dr. Tariq Saeed for their participation as members of my thesis guidance and evaluation committee. I am also grateful and would like to express sincere thanks to Principal SINES, Dr. Hammad Cheema and HOD SINES Dr. Fouzia Parveen Malik. Additionally, I extend my gratitude to NUST SCHOOL OF INTERDISCIPLINARY ENGINEERING & SCIENCE (SINES) for their collaboration and cooperation.

I am indebted to my sister Rabia Mallhi for her unwavering support and cooperation. Whenever I encountered challenges or obstacles, she always provided timely solutions and guidance. And without my best friend, Zubaida Khan's unwavering support, I would not have been able to complete this thesis. Her presence and encouragement have been invaluable, and I am fortunate to have had such a dedicated and supportive friend by my side.

Finally, I would like to extend my appreciation to all individuals, my friends who have provided valuable assistance and support throughout my study. Their contributions, whether big or small, have made a significant impact on the successful completion of this thesis. I am forever grateful to all those who have contributed to my journey and supported me in various ways. May Allah bless each one of them abundantly for their kindness and generosity.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment.

Table of Contents

List of Figures

List of Tables

Abstract

In the era of Artificial Intelligence (AI) and Intelligent Computing, the revolution in the healthcare sector is underway. Studies have focused on the development of decision-support processes for healthcare professionals to enhance disease screening and diagnostics. This study proposes a decision support system for diagnosing celiac disease (CD) using primary data, a complex autoimmune disorder affecting millions worldwide. CD diagnosis is complicated by socioeconomic factors, healthcare disparities, and limited access to advanced facilities and diagnostic technologies. Conventional methods are cost-prohibitive and lack of awareness contributes to underdiagnoses or misdiagnoses in developing countries.

The study focused on improving detection rates of CD by utilizing AI-based approaches. The study aimed to ensure that no cases of CD go undetected and minimize the risk of misdiagnosing celiac cases as non-celiac. The experimentation phase employed 5 automated classifiers available in Google Colab Notebook: decision trees, Bayesian classifier, XGBoost algorithm, support vector machine, and logistic regression. The assessment parameters considered encompassed accuracy, sensitivity, specificity, and the area under the ROC curve (AUC). These models were selected for their proven ability to handle both continuous and categorical data, including categorical dependent variables, within a classification task. Additionally, considering the limitations of previous applications of AI-based diagnostic methods, comprehensive data preprocessing and feature engineering techniques have been introduced including the application of Recursive Feature Elimination (RFE). Among the array of AI models examined the XGBoost classifier showed the highest accuracy of 97.0%, a sensitivity of 0.98, a false-negative ratio of 1 and an AUC of 0.91. The outcomes of the study are helpful in terms of a step towards the development of a smart clinical decision support system (CDSS). Future directions include market validation of the proposed process and transformation into a smart application for ease of adoption for the end users. The study's methodology, which encompasses primary data collection, robust preprocessing, and meticulous feature engineering, not only enhances predictive accuracy but also establishes a pioneering CD data repository. This repository, brimming with comprehensive patient information, is poised to reshape CD research. In essence, the study's detailed results underscore the transformative potential of AI-driven diagnostic approaches in tackling the complexities of celiac disease.

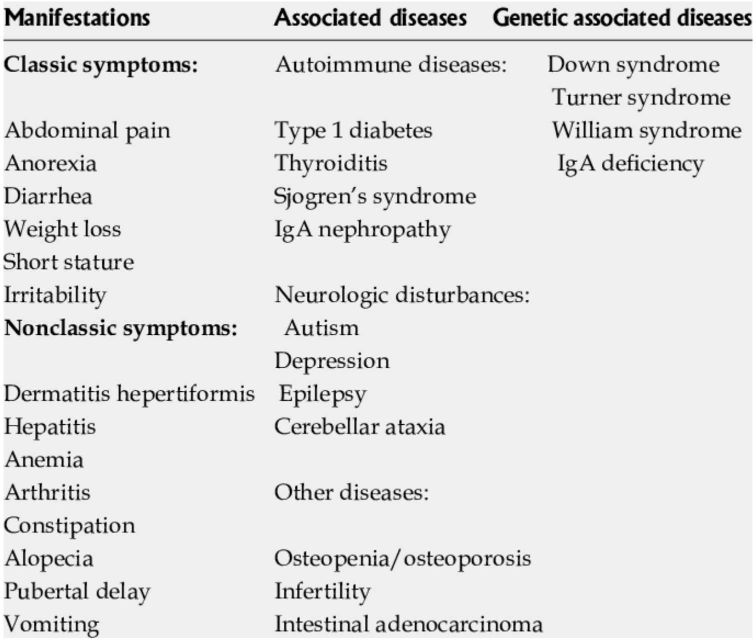
Key Words: *AI in diagnosis, Decision Support System, Intelligent Computing, Medical Informatics, Health Informatics, Celiac Disease, Autoimmune Disorder*

Chapter 1

Introduction

1.1 Celiac Disease:

CD is a permanent intolerance to gluten, the immune system mistakenly thinks that gluten — a protein in wheat, barley, rye, and oats — is a foreign invader [1]. It is an autoimmune disorder that can affect both males and females, regardless of their sex chromosomes [2]. CD is primarily associated with certain genetic variants in the human leukocyte antigen (HLA) genes, particularly HLA-DQ2 and HLA-DQ8 [3]. These genes are found on chromosome 6 and are involved in the regulation of the immune system[4]. Inheritance of these genetic variants can increase the risk of developing CD, but it does not follow a typical X-linked or Y-linked pattern of inheritance [5]. The likelihood of getting another autoimmune ailment increases with age for those with CD [6]. Celiac patients are also at high risk of developing stomach and intestinal cancer, esophageal squamous cell carcinoma, and colon cancer [7]. Many individuals have skin allergies like Dermatitis herpetiformis which is a long-lasting (chronic) skin disorder that causes itchy bumps and blisters[8]. Due to the extensive range of clinical symptoms and the involvement of numerous human systems, CD may be regarded as a syndrome. Compared to other autoimmune illnesses, CD exhibits peculiar characteristics [9]. Figure 1 showcases a diverse array of clinical manifestations associated with CD (CD) and associated diseases, highlighting the wide-ranging symptoms and presentations of this condition.



Manifestations	**Associated diseases**	**Genetic associated diseases**
Classic symptoms:	Autoimmune diseases:	Down syndrome
		Turner syndrome
Abdominal pain	Type 1 diabetes	William syndrome
Anorexia	Thyroiditis	IgA deficiency
Diarrhea	Sjogren's syndrome	
Weight loss	IgA nephropathy	
Short stature		
Irritability	Neurologic disturbances:	
Nonclassic symptoms:	Autism	
	Depression	
Dermatitis hepertiformis	Epilepsy	
Hepatitis	Cerebellar ataxia	
Anemia		
Arthritis	Other diseases:	
Constipation		
Alopecia	Osteopenia/osteoporosis	
Pubertal delay	Infertility	
Vomiting	Intestinal adenocarcinoma	

Figure 1: Clinical Manifestation of CD

Malnutrition may result from CD if left untreated or undiagnosed and when a person with CD consumes gluten, their body overreacts to the protein and harms their villi, which are tiny projections that resemble fingers and are found along the small intestine's wall. As small intestine doesn’t absorb nutrients from food when the villi are damaged and thus causes malnutrition which leads to a deficiency of certain vitamins and minerals that can cause conditions such as iron deficiency anemia, vitamin B12, and folate deficiency anemia, osteoporosis – a condition where your bones become brittle and weak [10]. Figure 2 provides a comprehensive depiction of the structural differences between normal and damaged villi by CD, illustrating how these structures facilitate the absorption and passage of nutrients. Misdiagnosed CD may, over time, increase the chance of significant side effects like small bowel cancer, lymphoma, and female infertility or miscarriages and also increase the chance of acquiring other autoimmune diseases [11].

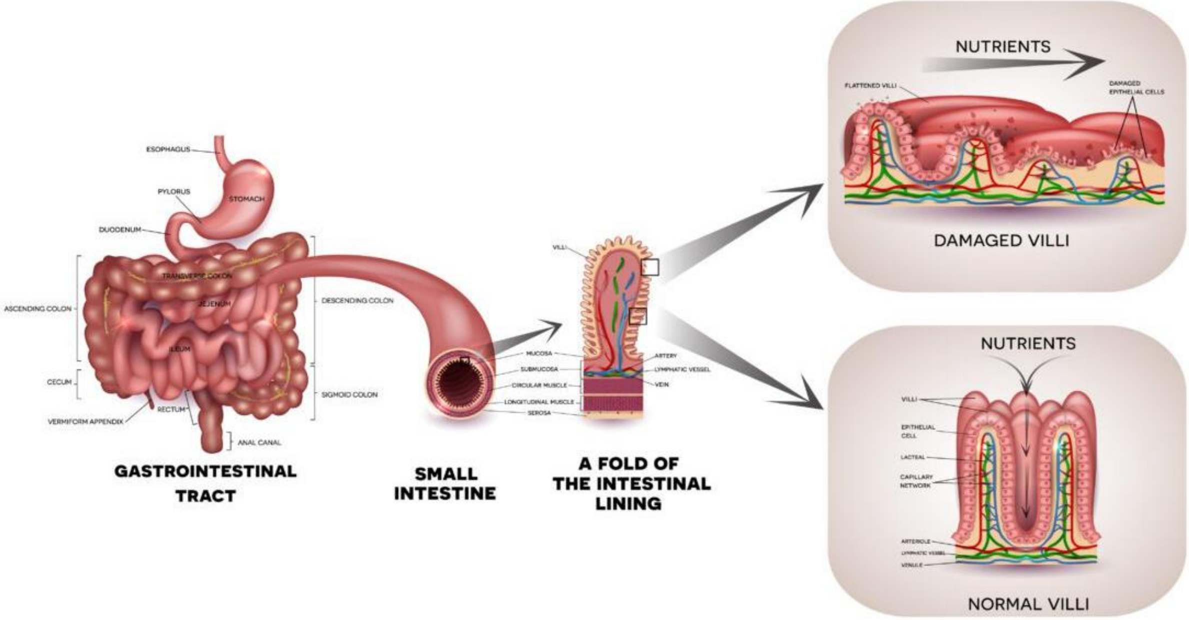


Figure 2: Comprehensive Depiction of the Structural Differences between Normal and Damaged Villi by CD

1.2 CD Statistics:

Individuals with CD have an overall increased mortality rate estimated at 57% [12]. Though the prevalence rate of this disease is about 1% of the population worldwide, approximately 90% of its cases remain undiagnosed [13]. The statistics of developing countries paint a concerning picture, particularly in Pakistan. According to a study, it is estimated that 1-3% of the Pakistani population has CD, with countless presumably undiagnosed [14]. In the Punjab region, a study found a prevalence of 1.2% [15]. Similarly, the Hazara Division of the Khyber Pakhtunkhwa province reported a prevalence of 9.0% [16]. However, there is a lack of comprehensive data regarding the prevalence of CD in other regions of Pakistan.

Figure 2 illustrates the increasing prevalence of CD. The research, conducted by BMC Health researchers provides authentic evidence of the rising occurrence of CD, highlighting its significance as a public health concern. The findings of this study emphasize the need for improved diagnostic coding and effective management of CD within primary care settings.

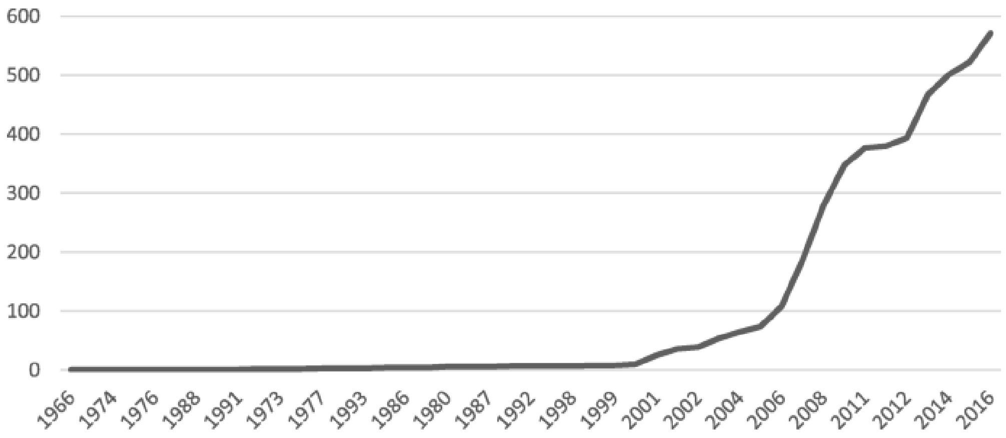


Figure 3: Prevalence of CD

1.3 Diagnostic Procedures and Problems Associated with Them

The currently recommended tests are the serum IgA-tissue transglutaminase antibody (TTG) and the IgA-endomysial antibody (EMA). These tests have a sensitivity and specificity of greater than 90% [17]. The TTG is currently the test of choice and is widely used as the main diagnostic test for CD in Pakistan [18], the patient must be consuming a normal, gluten-containing diet at the time of testing. TTG testing helps doctors identify that the person is at risk of having the disease and they further go for a biopsy for a complete diagnosis. During the test, blood is examined for antibodies. Increased antibody protein levels signify an immunological response to gluten[19]. The normal considered range of TTG <15 U/mL [20], a positive (greater) number indicates the presence of CD.

The majority of celiac specific serological test kits are currently imported from Europe and North America and are expensive which in result increase the diagnostic tests price for general public[21]. The cost of an anti-TTG test in Pakistan varies based on the lab and city where the test is performed, average price for celiac screening tests in Pakistan is from PKR 2000 to 5000 [22]. Another problem is that the diagnostic accuracy of these tests is studied for Caucasian people, and so the antibody level cutoffs are defined for these populations. Because of differences in genetic makeup and gluten consumption, these cutoffs for a positive test that have been set for the Caucasian population may not have the same diagnostic accuracy for Asian patients [23].

Video capsule endoscopy is also effective in the examination of CD patients [24], but takes time, requires endoscopic assistant knowledge, and is rarely worthwhile and is expensive so the general public of Pakistan cannot afford it[14]. Another crucial factor in CD diagnosis is that not all endoscopic biopsy specimens are reviewed and analyzed by gastrointestinal pathologists. It is critical to evaluate who is providing an opinion on the biopsy specimen. Not all general pathologists are aware of the full range of pathologic alterations seen in CD as it was initially defined and used worldwide [25]. Endoscopy cost in Pakistan ranges from PKR 30,000 to 150,000. It varies depending on the doctor, materials used, and hospital expenditure/equipment [26].

1.4 Awareness among Health Professionals

The CD is regarded as clinically problematic because it might appear to be many other diseases and ailments, confusing both patients and medical professionals [27]. The most recent clinical guidelines for the diagnosis and treatment of CD were released by the American College of Gastroenterology (ACG) in 2023, yet there are questions about practitioner comprehension and adherence [28]. Indeed, a study conducted among physicians in Pakistan to investigate their knowledge of the differences between IBS and CD that CD is often misdiagnosed as IBS [29]. According to another research finding, only 42% of people thought CD was a small intestine disorder, while the remainder thought it was a large intestine disorder [30].

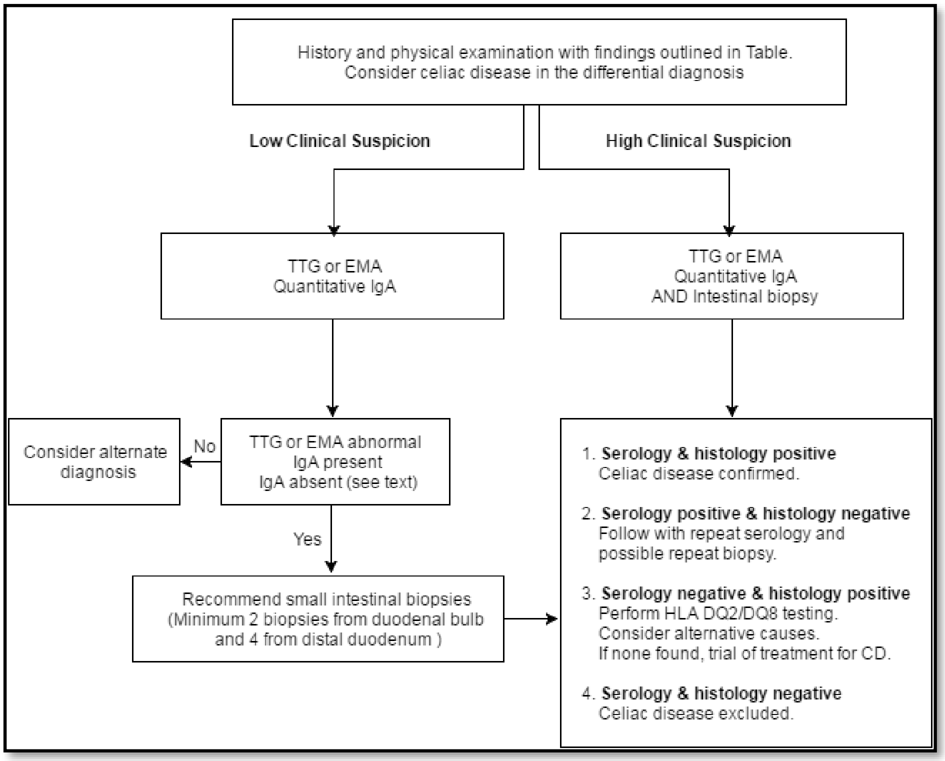

History and physical examination with findings outlined in Table.
Consider celiac disease in the differential diagnosis
Low Clinical Suspicion
High Clinical Suspicion
TTG or EMA
Quantitative IgA
TTG or EMA
Quantitative IgA
AND Intestinal biopsy
Consider alternate diagnosis
No
TTG or EMA abnormal
IgA present
IgA absent (see text)
Yes
Recommend small intestinal biopsies
(Minimum 2 biopsies from duodenal bulb and 4 from distal duodenum)
1. Serology & histology positive
Celiac disease confirmed.
2. Serology positive & histology negative
Follow with repeat serology and possible repeat biopsy.
3. Serology negative & histology positive
Perform HLA DQ2/DQ8 testing.
Consider alternative causes.
If none found, trial of treatment for CD.
4. Serology & histology negative
Celiac disease excluded.

Figure 4: Typical algorithm used for the evaluation of CD

1.5 Treatment:

The cornerstone of CD management is the adherence to a strict gluten-free diet. This approach requires individuals diagnosed with CD to eliminate all sources of gluten, including wheat, barley, rye, and their derivatives, from their diet [31]. The exclusion of gluten-containing foods and the adoption of gluten-free alternatives, such as rice, corn, millet, quinoa, and gluten-free flours, form the basis of the treatment [32] . Education and support play vital roles in the successful implementation of the gluten-free diet, as individuals need to develop a deep understanding of which foods are safe to consume, navigate social situations, and locate local resources for gluten-free products[27]. Regular follow-up with healthcare professionals, including gastroenterologists and registered dietitians, is crucial to monitor the individual's progress, address any concerns or complications, and conduct nutritional assessments to ensure the adequacy of the gluten-free diet. Through diligent adherence to a gluten-free lifestyle and ongoing healthcare support, individuals with CD can effectively manage their condition and improve their quality of life.

1.6 Justification for Selection of CD as Research Topic:

The selection of CD as a research topic in the context of Pakistan is of significant importance due to several reasons. Firstly, CD is a global health concern, affecting individuals of various ethnicities and geographical regions, including Pakistan. The prevalence of CD in Pakistan is reported to be relatively high, yet awareness and understanding of the condition among the general population and healthcare professionals remain limited. Secondly, investigating CD in Pakistan provides an opportunity to shed light on the unique challenges faced by individuals living in this region, including cultural dietary practices, availability and affordability of gluten-free products, and potential variations in genetic and environmental factors that may influence disease presentation and management. Lastly, as a celiac patient living in Pakistan, my personal experience serves as a motivation and a driving force behind this research endeavor. By examining CD within the Pakistani context, this study aims to contribute to the existing body of knowledge, raise awareness, and ultimately improve the diagnosis, treatment, and overall management of CD in Pakistan, benefiting individuals like myself and the broader population.

1.7 Problem Statement

Pakistan, like other developing countries, having a prevalence rate of 1-3 % and 90% presumably undiagnosed or misdiagnosed CD cases require an effective CDSS that will enhance the accuracy of timely diagnosis and significantly improve various aspects of patients' care.

1.8 Proposed Solution

The central focus of this study is to offer a solution that involves the development of a smart CDSS for timely and accurate CD diagnosis. By incorporating AI in diagnosis, the system can analyze a wide range of patient data and symptoms, considering the intricate patterns and correlations that might be missed by human clinicians alone. This human-centered AI approach will empower healthcare professionals with an intelligent computing tool that complements their expertise and aids in making informed decisions during the diagnostic process. Implementing a smart CDSS based on AI algorithms and human-centered AI principles has the potential to revolutionize CD diagnosis and management, leading to timely and accurate diagnoses, improved patient care, and enhanced quality of life. By optimizing effectiveness and efficiency, the system reduces the burden on patients and healthcare resources, ultimately benefiting individuals affected by CD.

1.9 Goal of the Study

Our main objective is to assist the concerned medical professionals through the development of a CDSS that will improve the techniques used for this complex disease screening.

1.10 General Objective

Patients suffering from CD suffer in silence in Pakistan as they face many challenges due to a lack of awareness about the disease. The general objective of this study is to provide awareness among medical professionals in the health sector, the general public, and also to the industrial sector which will result in more investments in the production of gluten-free products in Pakistan.

1.11 Specific Objectives

- To analyze the effectiveness of the ongoing screening procedure of CD cases, their false negative ratio, and associated details considering primary data.
- To develop a fully/partially automated CDSS for medical professionals for optimal decision-making.

1.12 Benefits

- Accurate diagnosis can be achieved with the creation of AI solutions to raise knowledge and awareness of the CD. The benefit of this study is to offer the ability to direct a knowledgeable user toward a diagnosis.
- Pakistan's digital health market is continuously expanding, and this study will be an opportunity to collaborate with various healthcare providers.

Chapter 2

Literature Review

2.1. Traditional Techniques for Diagnosis of CD

CD diagnosis traditionally relies on a combination of serological tests, genetic testing, and intestinal biopsy, as discussed in Chapter 1. The limitations of these traditional techniques have motivated researchers to explore alternative approaches, particularly those integrating AI and machine learning algorithms, to enhance the accuracy and efficiency of CD diagnosis. These emerging techniques leverage advanced computational methods to analyze complex patterns in patient data, including clinical information, genetic profiles, and histopathological images. By extracting relevant features and training predictive models on large datasets, these AI-powered systems aim to improve diagnostic accuracy, reduce invasive procedures, and optimize resource utilization.

The traditional methods have certain limitations, including false-positive results, inability to definitively confirm the disease, and invasiveness. The integration of AI and machine learning algorithms presents an opportunity to overcome these limitations by leveraging patient data for more accurate and efficient diagnosis. The examples from the literature demonstrate the potential of AI-powered systems in enhancing CD diagnosis.

2.2. CDSSs and Artificial Intelligence

In general, the CDSSs based on AI are very efficient if applied to training with large data sets; but large data sets are not always available in the medical domain [33]. The introduction of CDSSs for diagnosing CD could improve diagnostic work-up, allowing cost, time, and labor savings and improving the procedure’s safety, avoiding biopsy sampling and prolonged sedation associated with the multiple biopsy protocol. In particular, CDSS based on AI is enjoying growing research interest in solving classification problems in a wide range of application fields [34], especially in medicine, where the possibility of presenting classification results together with a measurement of the association is very tempting [35]. The interest of the scientific community in the development of CDSS systems, also thanks to new performing machine learning techniques, is certainly growing [36], [37].

2.3. Application of AI in Diagnosis

Artificial intelligence (AI) has revolutionized the field of disease diagnosis by offering innovative approaches that enhance accuracy, efficiency, and personalized care. Numerous studies have explored the application of AI in disease diagnosis, harnessing its capabilities to analyze complex data and generate valuable insights[38]. The figure provides a visual representation of the anticipated growth and market potential of AI in the healthcare sector. It showcases the expected size of the AI in healthcare market by the year 2030, indicating the significant role AI technologies are predicted to play in transforming healthcare delivery and decision-making processes. The figure serves as a valuable reference for understanding the projected expansion of AI applications in the healthcare industry, highlighting the growing importance and investment in this emerging field.

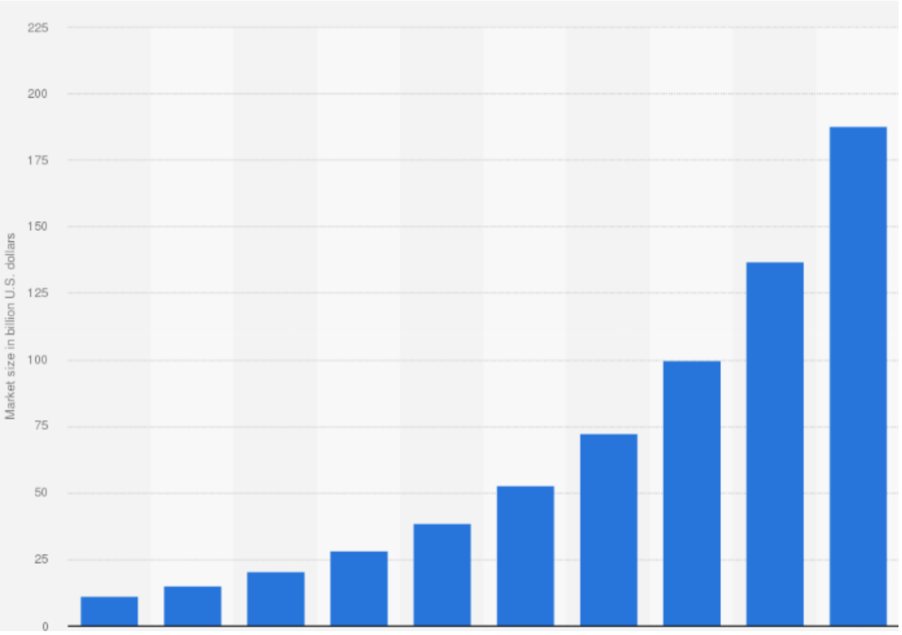


Figure 5: Market Potential of AI in the Healthcare Sector

In a study conducted by Pace et al. [39], the authors explored the use of artificial neural networks (ANN) and linear discriminant analysis (LDA) approaches in combination with gastroesophageal reflux disease (GERD) questionnaire to develop a novel model for distinguishing between healthy individuals and those with GERD. The study aimed to improve the diagnostic accuracy and efficiency of GERD detection by leveraging machine learning techniques. The researchers collected data from a sample of individuals who underwent upper gastrointestinal endoscopy and completed the GERD questionnaire. The questionnaire contained a series of questions related to the symptoms experienced by the participants. By combining the questionnaire responses with clinical data, such as age and sex, the researchers trained the ANN and LDA models to classify individuals as either healthy or having GERD. The results of the study demonstrated the potential of the ANN and LDA models in effectively distinguishing between healthy individuals and those with GERD. The combined model achieved a high accuracy rate, showcasing the ability of machine learning techniques to extract meaningful patterns from the collected data and provide valuable insights for diagnosis. This study by Pace et al. [31] highlights the potential of employing machine learning techniques in the diagnosis of GERD. By utilizing ANN and LDA algorithms, along with a GERD questionnaire, the researchers were able to develop a model that enhances the accuracy of GERD detection and aids in distinguishing between healthy individuals and those with the condition. The findings of this study contribute to the growing body of research exploring the application of machine learning in improving diagnostic processes for GERD and highlight the potential of such approaches in enhancing patient care and outcomes.

In the field of dermatology, Esteva et al. (2017) developed a deep-learning algorithm capable of diagnosing skin cancer by analyzing images of skin lesions. Their study demonstrated the algorithm's accuracy, which was comparable to that of dermatologists, showing the potential of AI in assisting with skin cancer diagnosis [40].

For breast cancer detection, McKinney et al. (2020) developed an AI system that analyzed mammograms. Their deep learning model outperformed radiologists in accurately identifying breast cancer, showcasing the potential for AI to augment the diagnostic capabilities of healthcare professionals in breast cancer screening [41].

In the context of lung diseases, Ardila et al. (2019) utilized AI to analyze chest radiographs for detecting multiple diseases, including pneumonia, tuberculosis, and lung cancer. Their deep learning model achieved high accuracy, demonstrating the potential of AI in aiding radiologists in diagnosing lung diseases more efficiently [3].

In the field of ophthalmology, Gulshan et al. (2016) developed an AI algorithm capable of diagnosing diabetic retinopathy by analyzing retinal images. The deep learning system demonstrated high sensitivity and specificity, suggesting its potential as a valuable tool for diabetic retinopathy screening and diagnosis [42].

Lastly, in the domain of neurological disorders, Diogo VS et al. (2019) explored the use of AI in diagnosing Alzheimer's disease by analyzing brain magnetic resonance imaging (MRI) scans. Their study demonstrated the potential of AI algorithms to accurately distinguish between Alzheimer's disease and other forms of dementia, aiding clinicians in making more precise diagnoses [43].

These studies collectively emphasize the potential of AI in disease diagnosis across various medical specialties. The application of AI techniques, such as deep learning algorithms, image analysis, and pattern recognition, allows for more accurate and efficient diagnoses, enabling timely interventions and improved patient outcomes.

2.4. Application of CDSS in Disease Diagnosis:

The application of CDSSs (CDSS) combined with artificial intelligence (AI) techniques has significantly transformed disease diagnosis, leading to improved accuracy, efficiency, and patient outcomes [44]. Several recent studies have explored the use of CDSS and AI in disease diagnosis, highlighting their advantages and revolutionary impact.

In the field of cardiovascular diseases, Attia et al. (2021) developed a CDSS using AI algorithms to analyze electrocardiogram (ECG) data for the detection of arrhythmias. Their study demonstrated that the AI-powered CDSS achieved high accuracy in identifying various arrhythmias, enabling timely interventions and improved patient management [45]. This advancement aids healthcare professionals in accurately diagnosing and treating cardiac conditions, potentially reducing complications and mortality rates.

For the diagnosis of neurological disorders, Havaei et al. (2019) developed a CDSS powered by deep learning algorithms to analyze brain magnetic resonance imaging (MRI) scans for the detection of multiple sclerosis (MS). Their model demonstrated superior performance in identifying MS lesions, assisting radiologists in making accurate diagnoses and enabling early interventions [46]. This advancement improves the accuracy and efficiency of MS diagnosis, facilitating prompt treatment initiation.

In the context of infectious diseases, Rajkomar et al. (2018) developed a CDSS utilizing AI techniques to predict sepsis onset in hospitalized patients. Their study demonstrated that the AI-powered CDSS accurately identified patients at risk of sepsis, allowing for early interventions and potentially reducing sepsis-related morbidity and mortality [47]. This advancement aids clinicians in making timely and informed decisions, leading to improved patient outcomes.

In the field of oncology, Liu et al. (2020) developed a CDSS utilizing machine learning algorithms to analyze medical images for the detection of lung cancer. Their model achieved high accuracy in identifying lung cancer nodules, assisting radiologists in accurate diagnosis and potentially improving survival rates through early detection [48]. This advancement aids in the early detection and personalized treatment planning of lung cancer, enhancing patient care and outcomes.

Moreover, in the domain of gastroenterology, Naz et al. (2020) utilized a CDSS powered by AI to analyze endoscopic images for the detection of gastrointestinal lesions. Their model demonstrated high accuracy in identifying lesions, assisting endoscopists in making accurate diagnoses and potentially reducing unnecessary procedures [49]. This advancement improves the efficiency and effectiveness of gastrointestinal disease diagnosis, leading to better patient care.

In 2017, Zia and Syed Saood created a clinical decision assistance system using a hybrid reasoning method. They employed a database management system (DBMS) and a knowledge-based CDSS (KBCDSS) to build their model. As an inference mechanism, the case base reasoning (CBR) technique was used as the suggested system's primary methodology, and a support vector machine (SVM) was included to classify cases and anticipate answers and it had an accuracy of 96.4% [50].

Based on feature extraction and optimal classification, Ashir Javeed and associates created the CDSS (CDSS) in 2022 for impartial cesarean section prediction. They applied the ROSE method or random oversampling example. A random forest (RF) model was used for classification while principal component analysis was used to extract features from the dataset. On training data, their proposed approach had an accuracy of 96.29%, while on test data, it had an accuracy of 97.12% [51].

These recent studies collectively demonstrate the advantages of CDSS and AI in disease diagnosis, including enhanced accuracy, efficiency, and personalized care. The integration of AI techniques into CDSS enables comprehensive analysis of patient data, leading to more accurate diagnoses, improved decision support, and optimized patient management. By revolutionizing the diagnostic process, CDSS and AI have the potential to transform healthcare delivery, improve patient outcomes, and contribute to the advancement of medical practice.

2.5. Existing CDSS for CD Screening

In recent years, several studies have explored the use of AI-powered CDSS for CD screening. This section reviews the existing literature on CDSS development for CD and summarizes the key findings and methodologies employed.

Several studies have utilized machine learning techniques to analyze data related to CD, aiming to improve its diagnosis and treatment. Tenorio et al. (2011) developed a

CDSS (CDSS) for CD diagnosis by integrating machine learning models and artificial intelligence techniques. They utilized a dataset of 178 clinical cases and incorporated 35 symptoms, including those associated with high-risk populations. The CDSS achieved promising results, with the AODE algorithm demonstrating the highest accuracy (80.0%), sensitivity (0.78), specificity (0.80), and area under the curve (AUC) value of 0.84 during testing [52]. However, it is important to note that the proposed system was not implemented, and further progress in the study was not reported.

Robert L. and associates reviewed the expert system for CD risk assessment and decision-making in 2019 that was developed by Tenorio and colleagues in 2011 [53], they examined and interrogated objectives and goals to build for the creation of a new CDSS based on evidence-based knowledge that serves as a robust system for clinical environments and a teaching tool. According to them, one of the key objectives of this CDSS was to consider the requirement for an instructional model that combines the precise language of symptoms, symptomatology, and other linked diseases and would be included in an electronic health record (EHR). They concluded and all agreed that the CD-CDSS is medically accurate and can direct healthcare providers through the diagnosis procedure.

In 2013, Shirts and colleagues employed the nearest neighbor algorithm to predict CD based on tissue transglutaminase antibody (TTG) levels and positive endoscopy results. Their approach focused on identifying similar patients to make clinically relevant predictions [54]. Another study by Ludvigsson in 2016 utilized natural language processing (NLP) techniques to screen potential celiac patients for further testing. By searching electronic medical records (EMRs) using NLP, they identified 216 CD patients and 280 controls, determining the optimal number of hits required for CD cases. The developed algorithms demonstrated a sensitivity of 17.1%, a specificity of 88.5%, and a positive predictive value of 0.9% [55].

In their study, Nasiriyan-Rad et al. [56] focused on the grading of CD using a fuzzy cognitive map (FCM) approach and the particle swarm optimization (PSO) algorithm. The objective of their research was to develop a computational model that could effectively grade the severity of CD based on clinical data. The authors utilized the

PSO algorithm to optimize the FCM parameters and improve the accuracy of the grading system. By incorporating various clinical features and their relationships, the FCM model was able to capture the complex interactions among different factors involved in CD progression. The PSO algorithm, known for its optimization capabilities, helped refine the FCM structure and weights, leading to improved accuracy in disease grading. The study achieved promising results, demonstrating the effectiveness of the proposed approach in accurately grading the severity of CD. The developed FCM-PSO model has the potential to assist healthcare professionals in making informed decisions regarding treatment strategies and monitoring disease progression, ultimately leading to improved management and personalized care for CD patients.

In his thesis, Reis [57] focused on CD diagnosis using an expert system approach. The objective of the research was to develop a computational model that could assist in the accurate diagnosis of CD based on patient symptoms and clinical data. The author utilized an expert system, which is a knowledge-based system designed to mimic the decision-making process of human experts, to create a diagnostic tool for CD. The expert system incorporated a rule-based approach, where a set of predefined rules derived from expert knowledge in the field of CD was used to evaluate patient data and provide a diagnosis. Through the development and implementation of the expert system, Reis aimed to improve the efficiency and accuracy of CD diagnosis. Although the specific achievements and outcomes of the study were not explicitly mentioned in the provided reference, the use of an expert system in CD diagnosis has the potential to streamline the diagnostic process, reduce errors, and facilitate prompt interventions for patients suspected of having CD.

In their study, Thukral and Bal [58] aimed to diagnose CD in North-Indian patients using a fuzzy logic probabilistic system. The researchers developed a computational model that incorporated fuzzy logic, a mathematical framework for handling uncertainty, to assess the likelihood of CD based on clinical parameters. The fuzzy logic system utilized linguistic variables and membership functions to represent the imprecision and uncertainty associated with CD diagnosis. By considering various clinical features, such as serological markers and symptoms, the model calculated the probability of CD diagnosis for each patient. The study achieved promising results, demonstrating the efficacy of the fuzzy logic probabilistic system in diagnosing CD in North-Indian patients. The model provided a quantitative measure of the likelihood of CD, aiding healthcare professionals in making accurate diagnoses and facilitating timely interventions. The use of fuzzy logic in CD diagnosis allows for the incorporation of uncertainty and imprecision, improving the diagnostic process and ultimately leading to better management and care for individuals with CD in the North-Indian population.

A study that summarized recent trends in computer-aided coeliac disease diagnosis using upper endoscopy. The authors discussed the advancements in technology and proposed pipelines for fully automated patient-wise diagnosis, as well as integrating expert knowledge into the automated decision-making process. The study by Gadermayr et al. [59] highlights the importance of utilizing computational methods to improve the accuracy and efficiency of CD diagnosis during upper endoscopy. By leveraging automated classification algorithms, it becomes possible to analyze endoscopic images and extract relevant features for identifying signs of CD. The authors discuss the current status of computer-aided diagnosis in this context and provide insights into future directions and potential advancements in the field. This research contributes to the growing body of literature on computer-aided diagnosis for CD and underscores the potential of integrating advanced computational techniques into the diagnostic process. The proposed pipelines and integration of expert knowledge can enhance the performance of automated systems and improve their clinical applicability. By leveraging such approaches, clinicians can benefit from reliable and efficient decision support during endoscopic procedures, leading to more accurate diagnoses and improved patient outcomes [59].

2.6. Challenges and Opportunities

Despite these advancements in machine learning applications for CD, there is currently a gap in the literature regarding specific machine learning models focused on CD in Pakistan. However, to our knowledge, to date, there are no works in which clinical support systems for CD screening have been developed, i.e. designed to work from non-invasive diagnostic tests (and to limit the use of biopsy) within the context of Pakistan. This highlights the need for further research in this area, considering the potential benefits of machine learning techniques in improving the accuracy and efficiency of CD diagnosis in the local population. By developing and implementing tailored machine learning models, healthcare professionals in Pakistan can potentially enhance the screening and management of CD for better patient outcomes.

While some studies have explored the use of AI techniques in CD diagnosis, existing CDSS models for CD have not been widely implemented or reached the development phase. For instance, studies utilizing machine learning algorithms, expert systems, or fuzzy logic approaches have demonstrated promising results in diagnosing CD. However, these models have not been translated into practical clinical tools or implemented in routine practice. This lack of practical implementation and limited progress in the development of CDSS for CD highlights a research gap that needs to be addressed. Therefore, there is a need for a new CDSS specifically designed for CD diagnosis, one that overcomes the barriers faced by previous models and advances beyond the research phase to practical implementation. Such a CDSS would fill the existing research gap, enabling healthcare professionals to benefit from the advancements in AI and CDSS technology and providing accurate, efficient, and personalized diagnoses for patients with CD. Furthermore, existing CDSS models often focus on specific populations or geographical regions, leading to a lack of generalizability and limited applicability in diverse patient populations. Therefore, a new CDSS for CD should be built on the basis of a comprehensive and multi-modal approach, incorporating AI techniques, machine learning algorithms, and expert knowledge to improve the accuracy and efficiency of CD diagnosis across different patient populations. Such a CDSS would address the existing research gap, enabling timely and accurate diagnoses, facilitating personalized treatment strategies, and ultimately improving patient outcomes in the field of CD diagnosis.

Chapter 3

Methodology

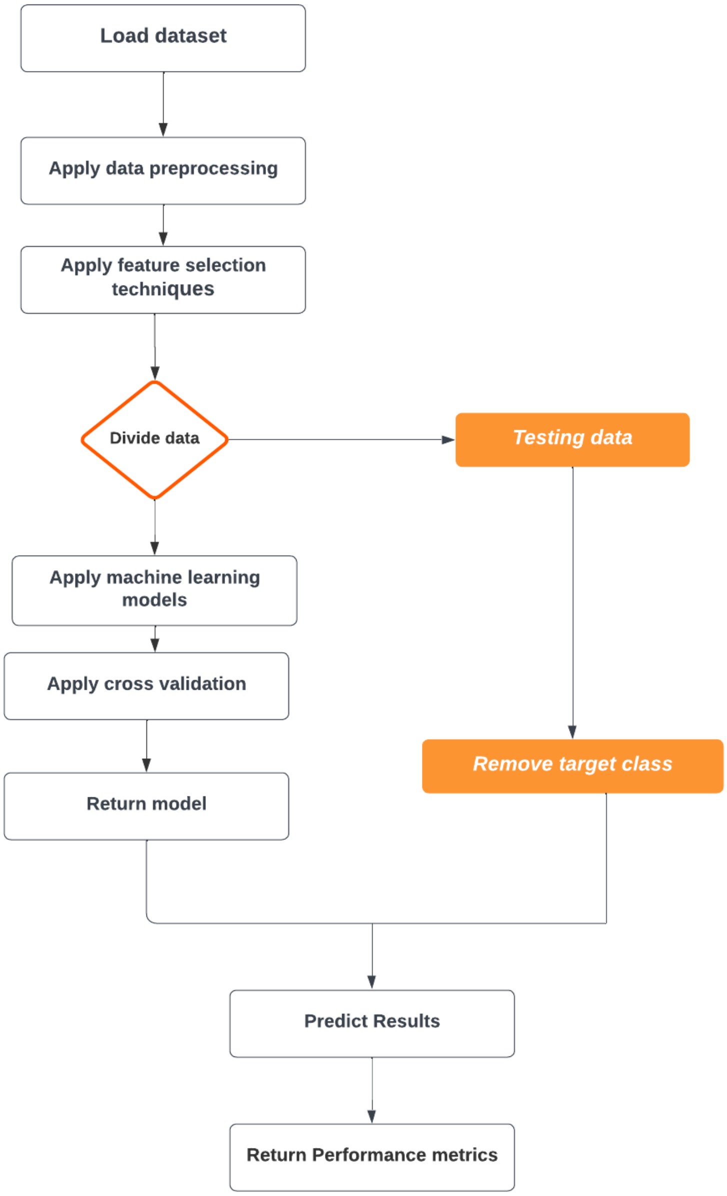


Figure 6: Flow Chart of Proposed Methodology

3.1. Data Description

In this thesis research, data is collected from three different resources: online surveys and various hospitals. The online surveys were conducted to gather primary data directly from participants, while the labeled data from different hospitals served as additional sources of information.

The online surveys were designed and distributed using a structured questionnaire, targeting a specific population relevant to the research objectives. Participants were asked to provide responses to a series of questions related to CD. The sampling method used for this is convenience sampling, which selects individuals who are easily accessible or willing to participate, and snowball sampling, also known as chain referral sampling, that identifies and recruit's participants through referrals from initial participants. Email and social media were used to contact respective participants. Closed-ended and open-ended questions about the demographics, signs, and severity of CD were included in the survey.

In addition to the online surveys, data was obtained from different hospitals. This data was obtained through collaboration and shared by the respective labs, ensuring its relevance and reliability to the research topic.

The data contains demographic information (e.g., age, gender, region), clinical characteristics (e.g., symptoms & complications), and laboratory test results (e.g., serum levels of anti-tissue transglutaminase antibodies-TTG). The data has a total of 2481 instances. The dataset includes both qualitative and quantitative attributes. There is a total of 16 attributes (including the target variable). The quantitative variables which are continuous in nature are Case no., Age (in years), and TTG count (AU/ml). The remaining 12 attributes are all categorical in nature, they all have yes or no values. The binary attributes include Gender, Anemia, Diarrhea, Skin rash, Thyroid, Psychological, Stunt growth, Nausea, Abdominal Pain, Fatigue, a Family member having CD, and a class variable (celiac or normal). Table 1:

Table 1: Description of attributes

	Name of attribute	**Description**
1.	Case no.	Represents the case count.
2.	Age	Represents the age of the patients (in years).
3.	Gender	Represents the gender of the patients. (male=0, female=1)
4.	TTG count	The TTG(tissue-transglutaminase) count describes the number of TTG and it was confirmed through the respective serological report. The unit considered is AU/ml.
5.	Anemia	Describes whether the patient had anemia as a symptom or not at the time of diagnosis. It has a value of yes or no.
6.	Diarrhea	Describes whether the patient had diarrhea as a symptom or not at the time of diagnosis. It has a value of yes or no.
7.	Skin rash	Describes whether the patient had skin rash as a symptom or not at the time of diagnosis. It has a value of yes or no
8.	Thyroid	Describes whether the patient had thyroid as a symptom or not at the time of diagnosis. It has a value of yes or no
9.	Psychological	Describes whether the patient had any psychological issue as dizziness, brain fog, migraine, etc. as a symptom or not at the time of diagnosis. It has a value of yes or no
10.	Short Stature	Describes whether the patient had stunt growth as a symptom or not at the time of diagnosis. It has a value of yes or no

11.	Nausea	Describes whether the patient had nausea as a symptom or not at the time of diagnosis. It has a value of yes or no
12.	Abdominal pain	Describes whether the patient had abdominal pain as a symptom or not at the time of diagnosis. It has a value of yes or no
13.	Fatigue	Describes whether the patient had fatigue as a symptom or not at the time of diagnosis. It has a value of yes or no
14.	Family member with CD	Describes whether the patient has any other family member having CD. It has a value of yes or no
15.	Other symptoms	Describes what other symptoms other than those mentioned above, the patient had as a symptom at the time of diagnosis. It has a value of yes or no
16.	Disease Status	The target variable (celiac or normal).

3.2. Data Preprocessing

3.2.1. Encoding Categorical Variables

The data is cleaned and preprocessed as an initial step to remove any incomplete or inconsistent records. In addition, some variables are recoded or transformed to better fit the statistical analysis. For example, 12 features with two classes of yes or no (Gender, Anemia, Diarrhea, Skin rash, Diabetes, Thyroid, Psychological, Stunt growth, Nausea, Abdominal Pain, Fatigue, a Family member having CD, and a class variable) are encoded (no=0, yes=1).

3.2.2. Handling Missing Values

The dataset has some missing values for certain variables, such as age and serum levels of anti-tissue transglutaminase (TTG) antibodies. To handle these missing values, we implement the linear trend imputation method, which is a simple and widely used approach for dealing with missing data in epidemiological research [60].

Linear trend imputation involves imputing missing values by assuming that the missing values follow the same linear trend as the observed values. This approach assumes that the values are missing at random (MAR), which means that the probability of absence depends only on the observed values and not on the missing values themselves. In this study, we used linear trend imputation to impute missing values for age and serum levels of TTG antibodies. We first examined the distribution of the observed values and found that they followed a linear trend over time. We then used the observed values to estimate the slope of the linear trend and used this slope to impute the missing values. After imputing the missing values, we also performed a sensitivity analysis to assess the impact of the missing value imputation on the results of the analysis. We compared the results obtained from the imputed dataset to those obtained from the complete dataset to ensure that the imputation did not introduce any bias or change the conclusions of the analysis.

Overall, the linear trend imputation method effectively handled missing values in our dataset and allowed us to use all available data in our analysis while minimizing bias due to missing data.

3.2.1. Data Partitioning

To evaluate the performance of the developed models and ensure their generalizability, the dataset was split into training and testing sets. A suitable strategy, such as stratified sampling, was employed to ensure a representative distribution of the target variable in both sets. This partitioning allowed for model development and training on the training set, followed by evaluation and validation on the independent testing set. It provided an unbiased assessment of the models' performance on unseen data.

3.2.1. Data Imabalance

One challenge encountered in this study was the imbalance in the dataset, where one class (e.g., CD) significantly outweighed the other (e.g., non-CD). Despite efforts to balance the data using techniques such as oversampling or under sampling, it was not possible to achieve a balanced distribution due to the limited availability of samples in the minority class. The justification for not balancing the data lies in the limitations and constraints of the dataset. In real-world scenarios, imbalanced data is a common occurrence, particularly in rare disease cases or situations where the occurrence of certain events is infrequent. In such cases, artificially balancing the data may introduce bias or compromise the representation of the real-world distribution. It is important to acknowledge and address the challenges associated with imbalanced data rather than artificially manipulating it. To mitigate the impact of class imbalance, appropriate evaluation metrics and model selection strategies were employed. Techniques such as stratified sampling, cross-validation, and performance metrics such as precision, recall, and F1-score were used to assess the models' performance on both classes, prioritizing the accurate identification of the minority class. This approach allowed us to focus on correctly predicting the presence of CD, even in the imbalanced dataset, and provide insights into its associated factors. Despite the inherent challenges posed by imbalanced data, the study recognizes the importance of addressing this issue and acknowledges its potential impact on model performance. By reporting and interpreting the results in consideration of the imbalanced nature of the dataset, the study ensures a balanced and comprehensive analysis while providing insights into the predictive capabilities and limitations of the models in real-world scenarios.

3.3. Feature Selection and Extraction

3.3.1. Chi-square Statistic

The chi-square test of independence is used to examine the associations between categorical variables and the label class (presence or absence of CD). This statistical test allowed us to assess the significance of the relationships between these variables and the CD status. By analyzing the chi-square statistic and the corresponding p-values, we evaluated the strength and significance of the associations. The chi-square test specifically examines the independence of variables and helps determine whether there is a meaningful relationship between them. Our analysis focused on assessing the associations between the categorical variables and the CD label class, providing valuable insights into the variables that exhibit a significant influence on CD and contributing to our understanding of the disease. The formula for calculating the chi-square statistic in a chi-square test of independence is:

$$\boldsymbol{\chi\text{^}2 = \Sigma\,[(O\_ij - E\_ij)\text{^}2\ /\ E\_ij]}$$

where:

χ^2 is the chi-square statistic, O_ij is the observed frequency in each cell of the contingency table, and E_ij is the expected frequency in each cell of the contingency table. The chi-square statistic is computed by summing the squared differences between the observed and expected frequencies, divided by the expected frequencies for each cell in the contingency table. The resulting value is a measure of the discrepancy between the observed and expected frequencies, indicating the degree of association or independence between the variables being tested.

The approach focuses on selecting features based on their individual association with the disease status, independent of any machine learning models. This comprehensive feature selection process enhances the accuracy and reliability of our predictive model, leading to more informed decision-making in the diagnosis and treatment of patients with CD.

3.3.2. Recursive Feature Elimination

In our study, we employed the recursive feature elimination (RFE) method with XGBoost, a state-of-the-art gradient boosting algorithm, to select a subset of high-importance features for the screening of CD. RFE is a robust feature selection technique that systematically evaluates the importance of variables and iteratively eliminates less significant features based on their impact on the model's performance. By leveraging the inherent capability of XGBoost to assess feature importance, we were able to identify a set of key variables that play a crucial role in the accurate prediction of CD. Table 4 shows chosen variables by RFE. By incorporating these carefully selected high-importance features identified through the rigorous RFE process, our screening model not only enhances the accuracy and efficiency of CD detection but also provides valuable insights into the underlying factors contributing to the condition. This comprehensive approach to feature selection, combining the power of XGBoost and the RFE method, ensures that our screening model is robust, reliable, and poised to make a significant impact in the early identification and management of CD.

3.3.3. Literature Review

Existing literature has identified these features as relevant indicators of CD. For instance, age has been recognized as a risk factor, with the disease commonly diagnosed during childhood or adulthood. Gender has also been investigated, showing a higher prevalence among females. Other symptoms such as anemia, skin rash, thyroid disorders, gastrointestinal issues (nausea, diarrhea, abdominal pain), fatigue, and weight loss have been reported as common manifestations of CD. Additionally, family history of CD has been found to increase the likelihood of developing the condition. These associations have been documented in various studies and clinical guidelines, providing a solid foundation for the inclusion of these features in the screening process.

By these multiple factors, we ensure the inclusion of relevant features that are informative for the screening of CD. Table 5, presents the selected features that were utilized for the development of the machine learning models.

3.4. Machine Learning Algorithms for CD Screening

For the screening of CD, selecting an appropriate machine-learning model is crucial to ensure accurate and reliable predictions [61]. Considering the dataset's characteristics, which consist of a mix of continuous and categorical variables with a size of 2481, several models have been chosen for evaluation.

3.4.1. Logistic Regression

Firstly, *logistic regression model* was selected due to its wide applicability in binary classification tasks. Logistic regression can handle both continuous and categorical variables with appropriate encoding [62].

$$\boldsymbol{P}\ (\boldsymbol{y} = \mathbf{1} \mid \boldsymbol{X})\ =\ \mathbf{1} \,/\, (\mathbf{1}\ +\ \boldsymbol{e}\hat{}(-\boldsymbol{z})) \qquad \textbf{(1)}$$

The logistic regression equation (1) represents the probability (P) of the binary outcome variable (y) being equal to 1, given the input features (X). The equation utilizes the logistic function, which transforms the linear combination (z) of the input features into a probability value between 0 and 1. The linear combination (z) is obtained by taking the sum of the product between each input feature (X) and its corresponding coefficient (β), and adding the intercept term (β0). The logistic function, defined as 1 divided by 1 plus the exponential of the negative of (z), converts the linear combination into a valid probability. This equation allows the logistic regression model to estimate the coefficients (β0, β1, β2, ..., βn) during training and subsequently predict the probability of the outcome variable being 1 for new instances based on their feature values. A general representation of logistic model is shown in Figure 6.

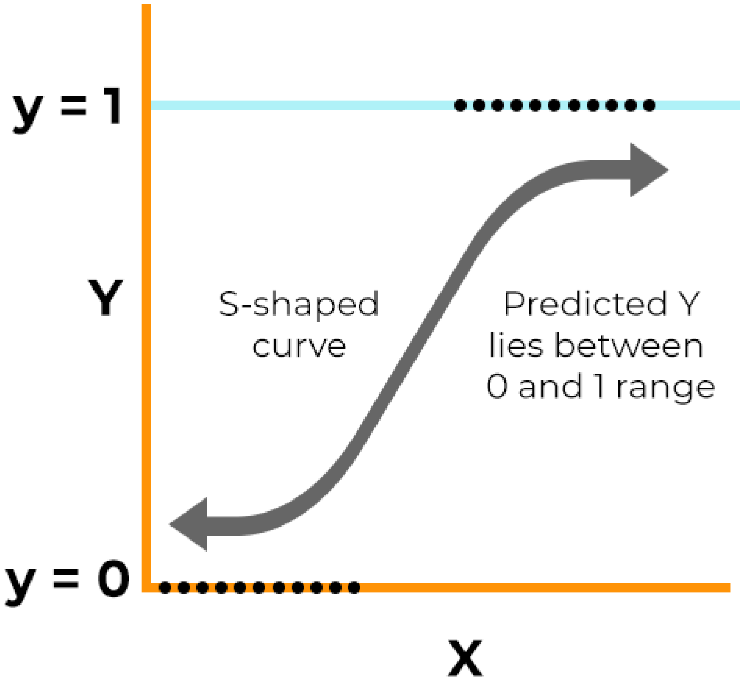


Figure 7: General working of Logistic model

3.4.2. Random Forest model

Secondly, *the random forest model* was chosen as an ensemble learning method for disease screening. Scikit-learn computes node importance in decision trees using Gini Importance [63], assuming two child nodes:

$$\boldsymbol{nij} \ = \boldsymbol{WjCj} — \boldsymbol{Wleft(j)Cleft(j)} — \boldsymbol{Wright(j)Cright}(j)$$

where:

nij is the importance of node j, Wj is the weighted number of samples reaching node j, Cj is the impurity value of node j, left(j) is the child node from left split on node j and right(j) is child node from right split on node j. The importance for each feature on a decision tree is then calculated as:

$$\boldsymbol{f\ i_i} \quad = \quad \frac{\boldsymbol{\Sigma j: node\ j\ splits\ on\ feature\ i\ nij}}{\boldsymbol{\Sigma k€all\ nodes\ nik}}$$

where:

$\boldsymbol{f}\,i_{\mathrm{i}}$ is the importance of feature i and nij is the importance of node j. These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values.

$$\boldsymbol{normfii} \quad = \quad \frac{\boldsymbol{f\ i\ i}}{\boldsymbol{\Sigma j€all\ features\ fij}}$$

The final feature importance in Random Forest is obtained by averaging the importance values across all trees.

$$\boldsymbol{RFfii} \quad = \quad \frac{\boldsymbol{\Sigma j€all\ trees\ normfiij}}{\boldsymbol{T}}$$

where:

RFfi$_{\mathrm{i}}$ is the importance of feature i calculated from all trees in the Random Forest model, normfi$_{\mathrm{ij}}$ is the normalized feature importance for i in tree j and T is total number of trees. Flowchart for random forest classifier is shown in Figure 7.

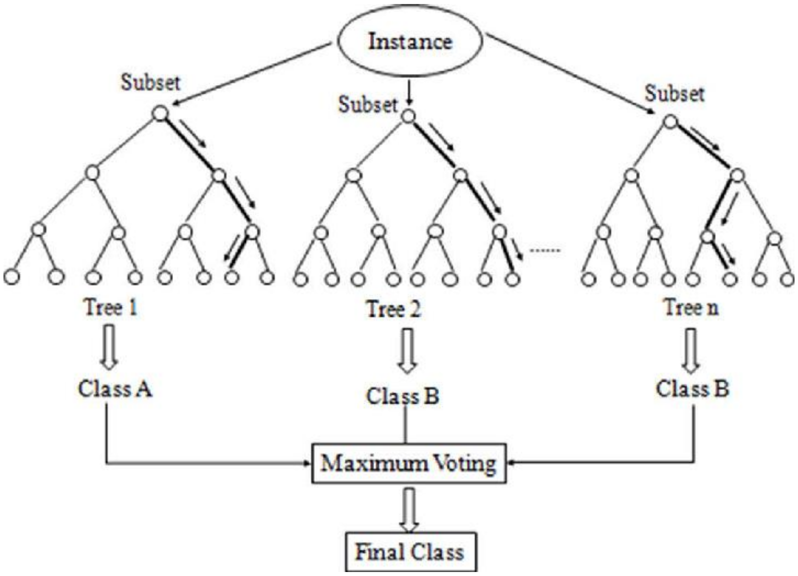


Figure 8: Flowchart of Random Forest Classifier

3.4.3. Naïve Bayes model

Naïve Bayes model was built. The Gaussian Naive Bayes algorithm assumes that the features are continuous and follow a Gaussian (normal) distribution. Given a set of input features X = {X_1, X_2, ..., X_n}, and a target variable y, the model estimates the probability of each class C_k (where k ranges from 1 to the number of classes) given the input features [64]. The posterior probability of a class C_k given the input features X can be calculated using Bayes' theorem:

$$\boldsymbol{P(C\_k|X)\ =\ (P(X|C\_k)\ *\ P(C\_k))\ /\ P(X)}$$

where:

P(C_k|X) is the posterior probability of class C_k given the input features X.P(X|C_k) is the likelihood of the input features X given class C_k, assuming a Gaussian distribution for each feature. P(C_k) is the prior probability of class C_k, which represents the probability of encountering class C_k in the training dataset. P(X) is the probability of the input features X, which serves as a normalization factor and can be calculated as the sum of P(X|C_k) * P(C_k) over all classes.

To calculate P(X|C_k), the Gaussian Naive Bayes model assumes that the features are conditionally independent given the class. Therefore, the likelihood of the input features X given class C_k can be calculated as the product of the individual feature probabilities:

$$\boldsymbol{P(X|C\_k)\ =\ P(X\_1|C\_k)\ *\ P(X\_2|C\_k)\ *\ldots*\ P(X\_n|C\_k)}$$

where:

each P(X_i|C_k) is the probability density function (PDF) of the i-th feature X_i given class C_k, assumed to be a Gaussian distribution. To train the Gaussian Naive Bayes model, the algorithm estimates the parameters (mean and variance) of the Gaussian distributions for each feature and class based on the training dataset. These parameters are then used to calculate the likelihood probabilities for the input features during the prediction phase. Illustration of how naïve Bayes works is shown in Figure 8.

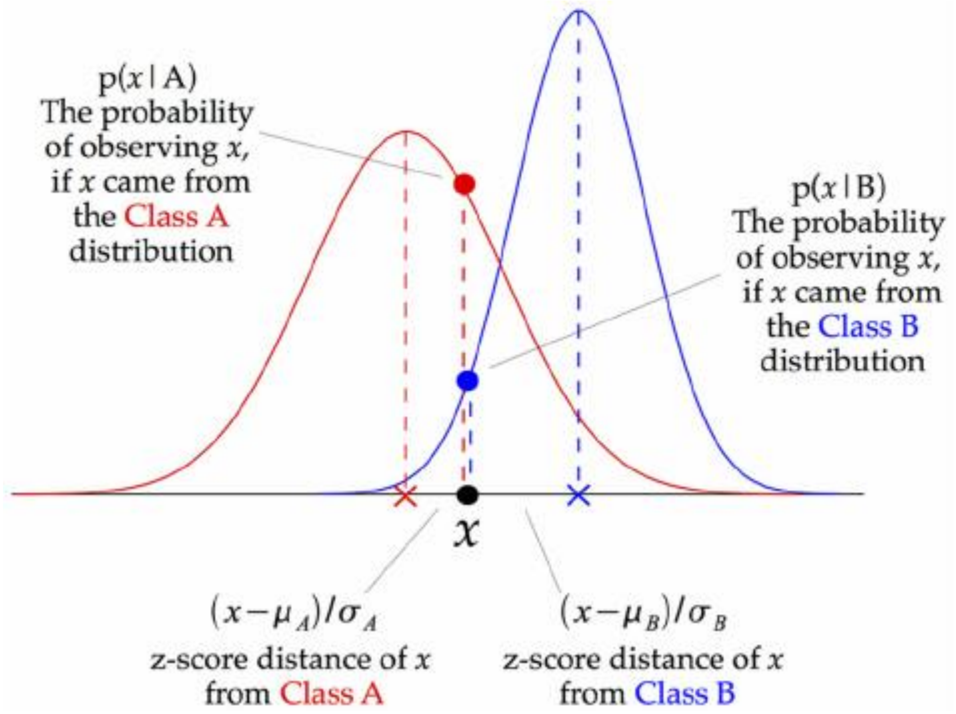

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Figure 9: Illustration of how naïve Bayes works

3.4.4. Extreme Gradient Boosting (XGBoost)

In addition, *Extreme Gradient Boosting (XGBoost)* was considered for their powerful and versatile nature in classification tasks. This algorithm is a high-performance implementation of gradient-boosted decision trees and can handle a mix of continuous and categorical variables [65]. A general and simplified working structure of XGBoost in presented in Figure 9. Given the relatively large size of the dataset, XGBoost model offer excellent performance and can handle imbalanced class distributions. The term "Gradient Boosting" was coined by Friedman in 2001 and is commonly applied to structured and tabular data [66]. XGBoost leverages the second-order Taylor expansion of the loss function and incorporates a regularization term to strike a balance between model complexity and minimizing the loss function. To make predictions with a dataset containing n examples and m features [67], Equation (2) is employed.

$$\boldsymbol{\hat{y}\ i = \phi(xi) = \sum k = 1Kfk(xi), fk\ \epsilon\ F} \tag{2}$$

where:

F= $\{f\ (x)=wq(x)\}$ ($q : Rm \rightarrow T$, $w \in RT$) is the spacing of the trees, q is the structure of each tree and T is the number of leaves of the tree? Therefore, each f is an independent tree structure. The regularized objective can be minimized as follows in (3).

$$\boldsymbol{L\ (\phi) = \sum il\ (\hat{y}\ i, yi)\ + \sum k\Omega\ (fk)} \tag{3}$$

where:

$\Omega(f)=\Upsilon T+12\lambda\|w\|$ 2. In this case, l is the convex of the loss function and Ω

is the penalization of the complexity of the model. With the aim of improving the objective i instance and t iteration are added and using a second-order approximation obtaining (6).

$$\boldsymbol{L(t) \simeq \sum i = 1n\ [l\ (yi, \hat{y}\ i(t-1)\ + gi\ ft(xi) + 12hif2t(xi))]\ + \Omega(ft)} \tag{4}$$

where:

$gi=\partial\hat{y}\ l\ (yi,\ \hat{y}\ i(t-1))\ and\ hi=\partial 2\hat{y}\ (t-1)\ l\ (yi,\ \hat{y}\ i(t-1))$ are first and second order gradient statistics on the loss function. Removing the constant terms and defining I_j = { i|q(x_i)= j} as the instance of j [41]. Also, defining $w*j=-\Sigma$ $i\in Ij\ gi\Sigma\ i\in Ij\ hi+\lambda$

finally, (7) is obtained. This can be used as a scoring function to measure the quality of a tree.

$$\boldsymbol{L(t)(q) = -12\sum Tj = 1(\Sigma\ i \in Ij\ gi)2\Sigma\ i \in Ij\ hi + \lambda + \gamma T} \tag{5}$$

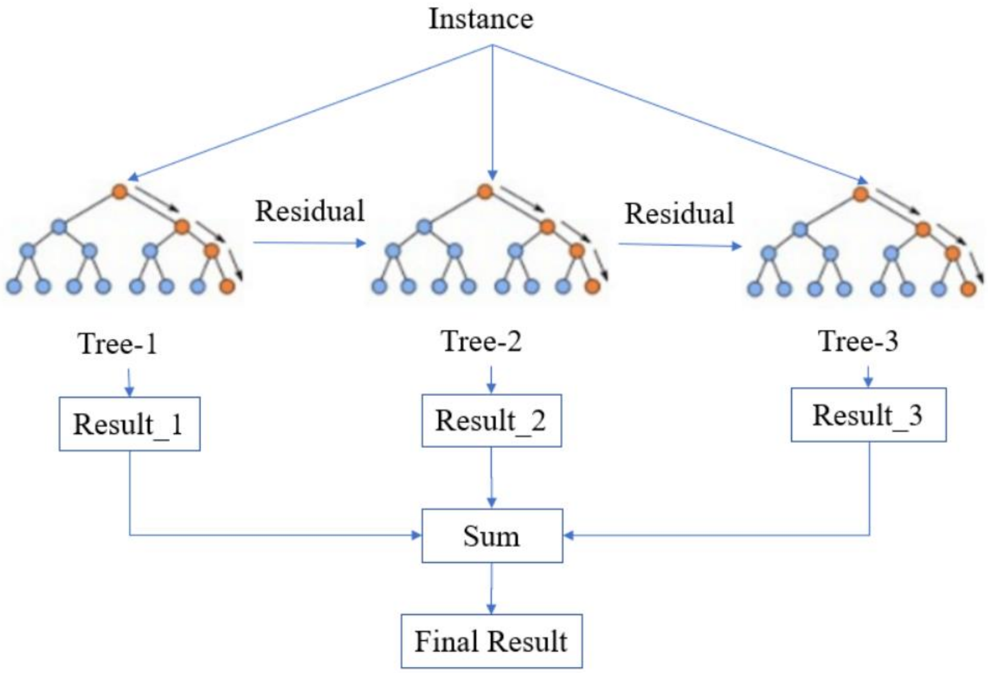


Figure 10: Simplified working structure of XGBoost

3.4.5. Support Vector Classifier model (SVC)

Lastly, *support vector classifier (SVC)* was chosen as a robust and effective classification algorithm for the given dataset. SVC, a supervised learning technique, is employed for classification, outlier detection, and feature selection tasks. Its primary goal is to establish an optimal hyperplane that maximizes the separation between classes [68]. In a two-dimensional space, the hyperplane corresponds to a line. For three-dimensional space, it becomes a two-dimensional plane. Similarly, SVM constructs n-dimensional hyperplanes in $Rn-1$, where n represents the dimensionality or feature count of the data. A kernel function is needed to map the data [69].

Given a set of n observations with x representing the training data and y the class of the label, as seen in (6).

$$\boldsymbol{S} = \{(\boldsymbol{x1}, \boldsymbol{y1}), \dots, (\boldsymbol{xi}, \boldsymbol{yi})\} \quad \textbf{(6)}$$

The decision function for nonlinear data of the algorithm is given by (7). Where m is the bias parameter and α determines the maximal margin classifier, a parameter related to the input vector.

$$\boldsymbol{f(x) = sgn\ (\textstyle\sum i = 1N\alpha iyKi\ (xi.x)\ + m)} \qquad \textbf{(7)}$$

where K is the kernel function. In this study, we used linear kernel.

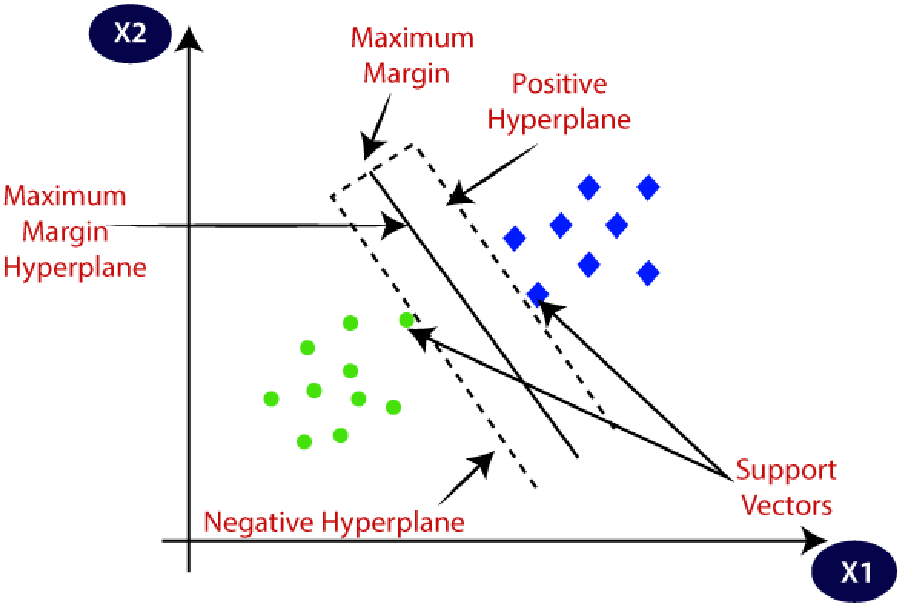


Figure 11: SVC algorithm

3.5. Justification for using these models in our Analysis

The selection of machine learning models, including Random Forest, XGBoost, SVM, Logistic Regression, and Naive Bayes, for CD screening is justified based on several factors pertaining to the dataset and the classification problem.

Firstly, the size of the dataset plays a crucial role in model selection [70]. With a sufficiently large dataset, models like Random Forest, XGBoost, and Logistic Regression can effectively handle the volume of data and provide reliable predictions. Additionally, Naive Bayes is known for its scalability and ability to handle large datasets efficiently [71]. Secondly, the complexity of the classification problem necessitates models that can capture intricate relationships between variables. Random Forest and XGBoost are ensemble methods that excel at capturing non-linear relationships and interactions, making them suitable for identifying complex patterns in the data [72]. Logistic Regression, on the other hand, provides interpretable results and allows for a better understanding of the relationship between the variables and the disease status [73]. Naive Bayes, although based on a simple probabilistic framework, can still capture certain dependencies between variables and perform well in classification tasks [74]. Furthermore, the presence of both continuous and categorical variables in the dataset requires models that can handle mixed data types. Logistic Regression and Naive Bayes can handle both continuous and categorical variables with appropriate encoding techniques [75]. Random Forest and XGBoost can naturally handle a mix of data types without extensive preprocessing, making them suitable for the heterogeneous nature of the CD dataset [76].

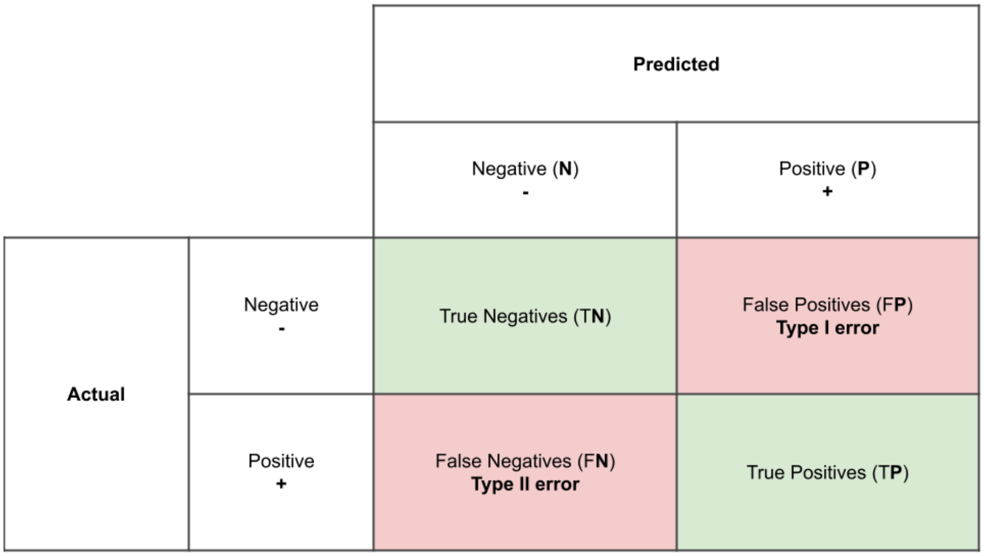
In summary, the selection of Random Forest, XGBoost, SVM, Logistic Regression, and Naive Bayes models for CD screening is justified based on the dataset's size, complexity of the problem, and the presence of mixed data types. These models provide a comprehensive set of options, combining accuracy, interpretability, scalability, and the ability to handle different types of variables, to effectively address the binary classification task of CD screening.

3.6. Evaluation Metrics

In this section, we discuss the evaluation metrics used to assess the performance of the machine learning models for CD screening. The following metrics were employed: Confusion Matrix, Accuracy, Precision, Recall, F1-Score, and ROC Curves.

3.6.1. Confusion Matrix

The Confusion Matrix is a table that provides a comprehensive summary of the model's performance by displaying the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. It is a valuable tool for assessing the model's ability to correctly classify instances and identify potential errors. An illustration of the confusion matrix is shown in Figure 12.



		Predicted	
		Negative (**N**) **-**	Positive (**P**) **+**
Actual	Negative **-**	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive **+**	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 12: Confusion Matrix

3.6.2. Accuracy

Accuracy measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances to the total number of instances. It is a widely used metric to evaluate the model's performance, especially in balanced datasets, as it provides a general overview of the model's predictive power. The formula for calculating the accuracy is given as:

$$\boldsymbol{Accuracy = (TP + TN) / (TP + TN + FP + FN)}$$

3.6.3. Precision

Precision quantifies the model's ability to accurately identify positive instances. It is calculated as the ratio of true positives to the sum of true positives and false positives. Precision focuses on minimizing false positive predictions and is particularly important when the cost of misclassifying positive instances is high. The formula for calculating the precision is given as:

$$\boldsymbol{Precision = TP / (TP + FP)}$$

3.6.4. Recall

Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify positive instances from the actual positive samples. It is calculated as the ratio of true positives to the sum of true positives and false negatives. Recall is important when the goal is to minimize false negatives, ensuring that all positive instances are captured. The formula for calculating the recall is given as:

$$\boldsymbol{Recall = TP / (TP + FN)}$$

3.6.5. F1-Score

The F1-Score is a harmonic mean of precision and recall, providing a balanced measure of the model's performance. It combines precision and recall into a single metric that considers both false positives and false negatives. F1-Score is useful when there is an uneven distribution of classes or when the cost of both false positives and false negatives needs to be considered. The formula for calculating the f1-score is given as:

$$\boldsymbol{F1 - Score = 2 * (Precision * Recall) / (Precision + Recall)}$$

3.6.6. ROC Curves

Receiver Operating Characteristic (ROC) curves are graphical representations that illustrate the model's performance across various classification thresholds. They plot the true positive rate (sensitivity) against the false positive rate (1 - specificity) at different threshold settings. ROC curves provide insights into the trade-off between sensitivity and specificity and help determine an optimal threshold for classification. ROC curves are obtained by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The formulas for TPR and FPR are as follows:

$$\boldsymbol{TPR = TP \,/\, (TP + FN)}$$

$$\boldsymbol{FPR = FP \,/\, (FP + TN)}$$

In conclusion, the evaluation metrics used in this study, including the Confusion Matrix, Accuracy, Precision, Recall, F1-Score, and ROC Curves, provide a comprehensive assessment of the machine learning models' performance for CD screening. These metrics enable a thorough analysis of the models' ability to correctly classify instances, quantify the balance between true positives and false positives, and assess the overall predictive power of the models.

Chapter 4

Results

4.1. Descriptive Statistics and Data Visualization

The dataset used in this study consisted of 2481 samples and 16 variables related to CD screening. The target variable, CD status, was distributed with 80% of the samples classified as positive. The descriptive statistics revealed that the age of the participants ranged from 03 to 65 years, with a mean age of 14 years. Other variables, such as gender, TTG count, anemia, nausea, fatigue, weight loss, short stature, abdominal pain, family member with CD, and diarrhea, exhibited varying distributions and levels of prevalence.

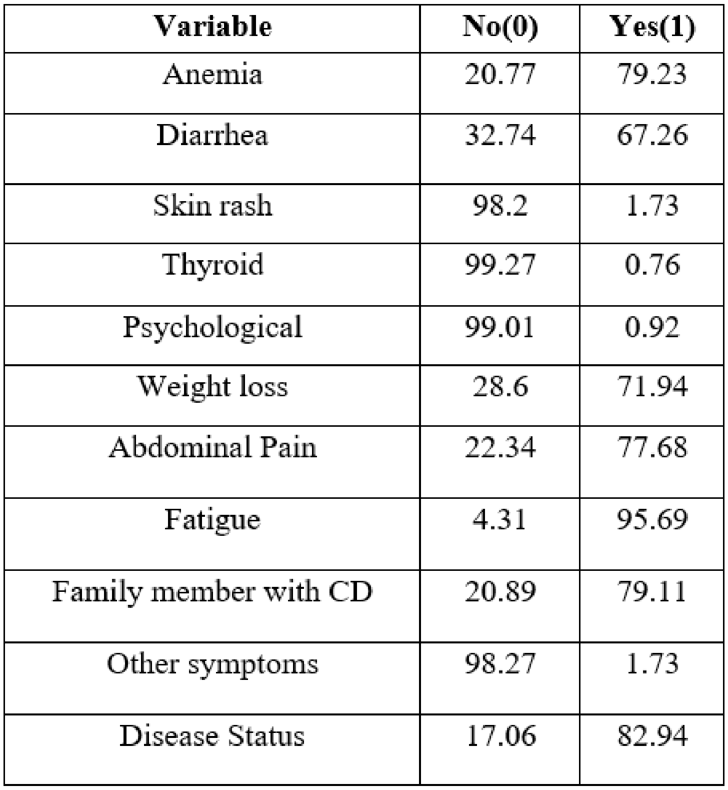
To get a better understanding of the data, a descriptive analysis has been done. It provides a comprehensive overview of the data, allows the identification of patterns, trends, and relationships, and helps in understanding the characteristics and nature of the data.

To communicate the findings: Table 2, shows the mean age, the standard deviation (SD), skewness, kurtosis, and the range (minimum and maximum) of quantitative variables. Table 2, shows the percentages of each symptom among two classes (yes, no)

Table 2: Descriptive Statistics of Quantitative Variables

Variable	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Age_in_years	3	65	14	8.60	1.16	2.5
TTG count	0	964.3	43.32	61.2	4.07	35.9

Table 3: Symptom Prevalence among Classes



Variable	No(0)	Yes(1)
Anemia	20.77	79.23
Diarrhea	32.74	67.26
Skin rash	98.2	1.73
Thyroid	99.27	0.76
Psychological	99.01	0.92
Weight loss	28.6	71.94
Abdominal Pain	22.34	77.68
Fatigue	4.31	95.69
Family member with CD	20.89	79.11
Other symptoms	98.27	1.73
Disease Status	17.06	82.94

To enhance the understanding of the data, data visualization techniques have been used. Figure 3, shows the distribution of people having CD or not. Figure 4, demonstrates the percentages of female and male participants in the dataset. Figure 5 shows the histograms of quantitative continuous variables.

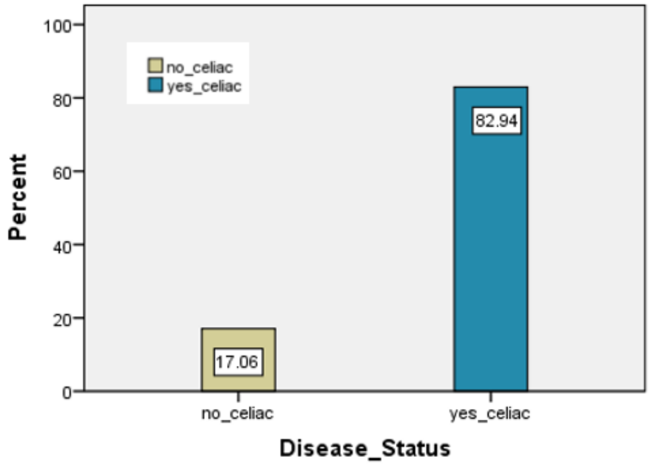


Figure 13: Distribution of Individuals with CD

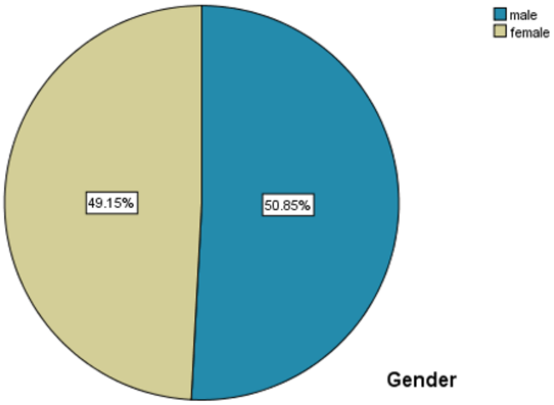


Figure 14: Gender Distribution of Participants

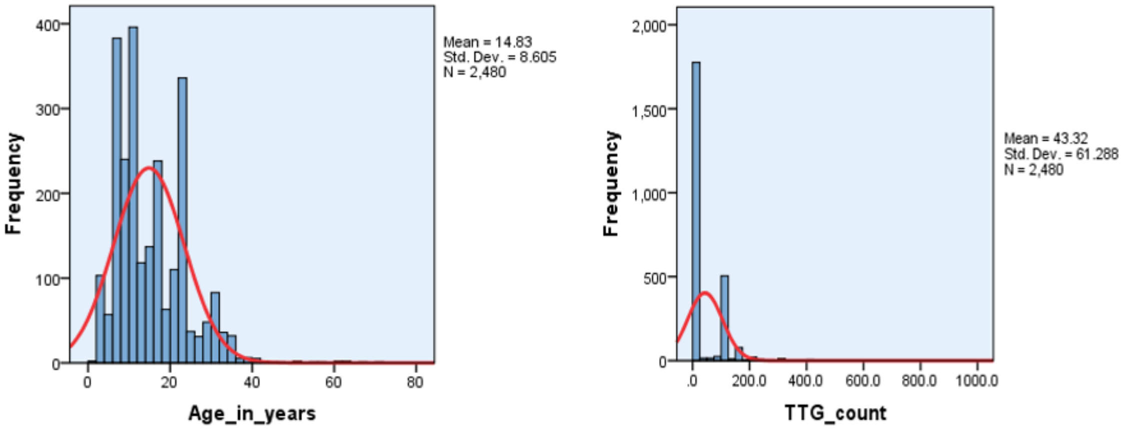


Figure 15: Histograms of Quantitative Continuous Variables

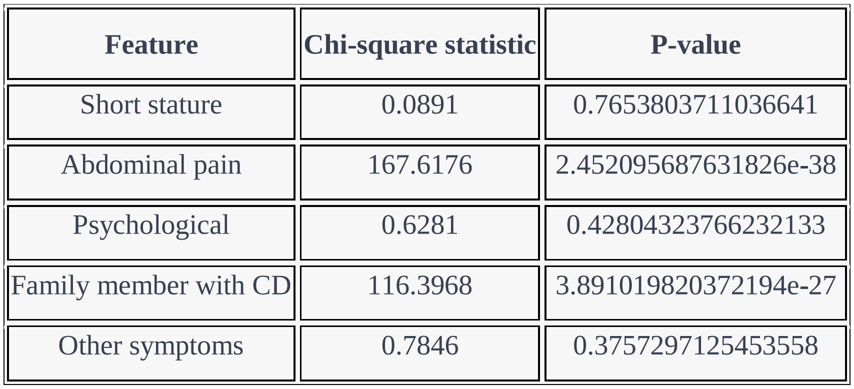
4.2. Feature Selection

4.2.1. Feature Selection by Chi-square statistics:

The selection of the chi-square test for association and feature engineering in our thesis was a strategic decision driven by the nature of our variables and the specific objectives of our research. Given that we were working with categorical variables related to CD, the chi-square test provided a suitable method to examine the associations between these variables and the disease status. By applying this test, we were able to identify the significant relationships and prioritize the most relevant features for our predictive models. This approach not only enhanced the accuracy and reliability of our screening system but also provided valuable insights into the underlying factors influencing CD. By leveraging the power of the chi-square test, our study contributes to the existing knowledge and understanding of this condition, paving the way for improved diagnosis and management strategies. By analyzing the chi-square statistic and the corresponding p-values, we evaluated the strength and significance of the associations. Table 4, shows significant associations of variables with the disease status, as evidenced by their chi-square statistics and p-values.

Table 4: Associations of Variables with Disease Status

Feature	Chi-square statistic	P-value
Anemia	1071.0751	6.391794601440761e-235
Gender	6.7997	0.009117544448928004
Skin rash	1.3438	0.24636195659456672
Thyroid	0.9752	0.32337855680560457
Nausea	0.0104	0.9188694131059258
Fatigue	3.1460	0.0761134736484347
Weight loss	164.0249	1.4938848191516352e-37
Diarrhea	4.6716	0.030666090797333564

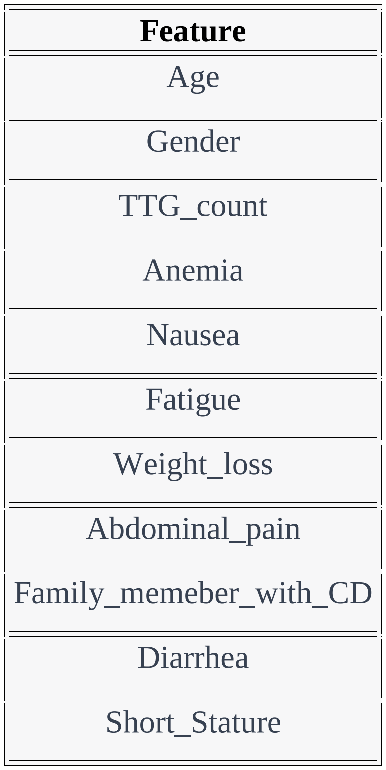


Feature	Chi-square statistic	P-value
Short stature	0.0891	0.7653803711036641
Abdominal pain	167.6176	2.452095687631826e-38
Psychological	0.6281	0.42804323766232133
Family member with CD	116.3968	3.891019820372194e-27
Other symptoms	0.7846	0.3757297125453558

4.2.2. Feature Selection by XGBOOST- Recursive Feature Elimination (RFE) Algorithm

In our study, we employed the recursive feature elimination (RFE) method with XGBoost, a state-of-the-art gradient boosting algorithm, to select a subset of high-importance features for the screening of CD. RFE is a robust feature selection technique that systematically evaluates the importance of variables and iteratively eliminates less significant features based on their impact on the model's performance. By leveraging the inherent capability of XGBoost to assess feature importance, we were able to identify a set of key variables that play a crucial role in the accurate prediction of CD. Table 5 shows chosen variables by RFE. By incorporating these carefully selected high-importance features identified through the rigorous RFE process, our screening model not only enhances the accuracy and efficiency of CD detection but also provides valuable insights into the underlying factors contributing to the condition. This comprehensive approach to feature selection, combining the power of XGBoost and the RFE method, ensures that our screening model is robust, reliable, and poised to make a significant impact in the early identification and management of CD.

Table 5: Selected Variables by Recursive Feature Elimination (RFE)

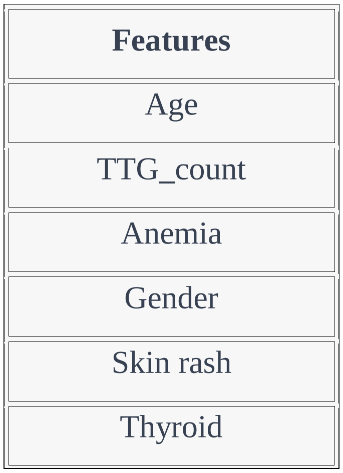


Feature
Age
Gender
TTG_count
Anemia
Nausea
Fatigue
Weight_loss
Abdominal_pain
Family_memeber_with_CD
Diarrhea
Short_Stature

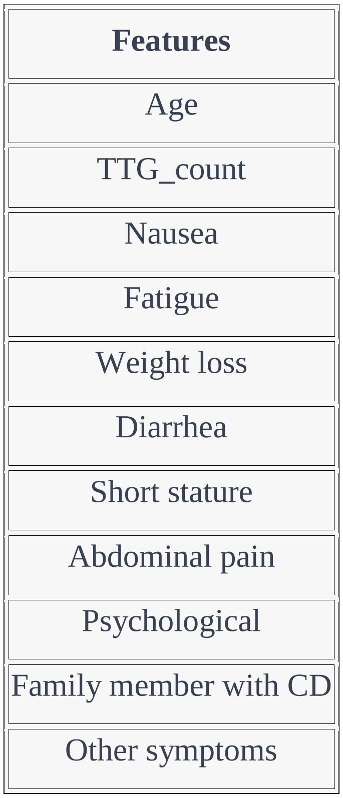
4.2.3. Final Pool:

In the process of feature engineering, an initial set of 15 features was considered as mentioned in Table 6.

Table 6: Key Features for CD Prediction



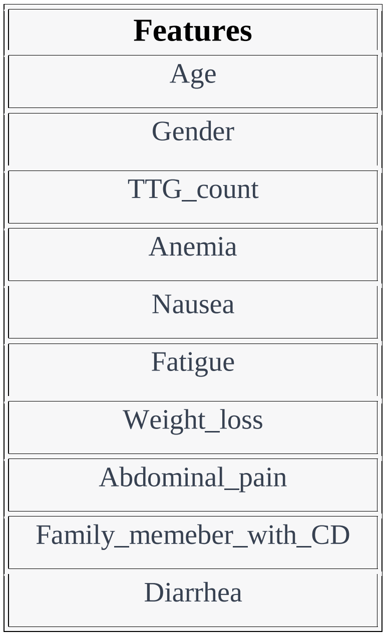
Features
Age
TTG_count
Anemia
Gender
Skin rash
Thyroid



Features
Age
TTG_count
Nausea
Fatigue
Weight loss
Diarrhea
Short stature
Abdominal pain
Psychological
Family member with CD
Other symptoms

However, meticulous analysis and evaluation of our machine learning models revealed an intriguing finding. The performance of the models experienced a significant boost when employing a refined feature set consisting of 10 carefully selected variables. The refined feature set comprising 10 variables as mentioned in Table 7, which exhibited both high chi-square values and were selected through the Recursive Feature Elimination (RFE) method, demonstrated a substantial improvement in the performance of the models. We ensured that the selected features were not only highly relevant but also had a substantial impact on the predictive performance of the models. Notably, even with the reduction in the number of features, the XGBoost model consistently maintained its high accuracy, reinforcing its robustness and superiority in CD prediction. This compelling result further highlights the effectiveness of XGBoost in capturing the essential patterns and relationships present in the data, regardless of the feature set size. By focusing on the most influential and informative features, the refined model achieved superior accuracy, demonstrating its capability for precise CD prediction. This meticulous feature selection process, along with the consistent performance of XGBoost, underscores the importance of thoughtful analysis and highlights the significance of selecting a parsimonious yet effective set of features for accurate prediction tasks.

Table 7: Final Pool of Selected Features for CD Screening



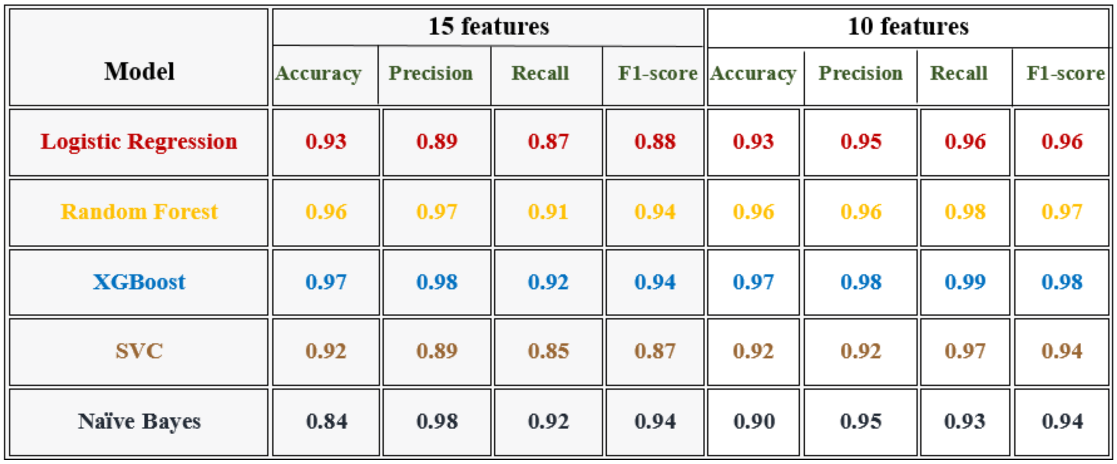
Features
Age
Gender
TTG_count
Anemia
Nausea
Fatigue
Weight_loss
Abdominal_pain
Family_memeber_with_CD
Diarrhea

Table 7 showcases the final pool of selected features that have been identified as crucial for CD screening. These features have undergone a rigorous selection process, considering their statistical significance and predictive importance. The table presents the specific variables that have demonstrated meaningful associations with the disease status, offering valuable insights for accurate and reliable screening. By incorporating these selected features into the screening model, we enhance the ability to identify individuals at risk of CD, facilitating early detection and appropriate interventions.

4.3. Model Evaluation

The performance of five machine learning models was evaluated using both the 15-feature and 10-feature sets. The evaluation metrics, including accuracy, precision, recall, and F1-score, were computed for each model. The results showed that all models achieved higher performance metrics when using the 10-feature set compared to the 15-feature set, except for the XGBoost model, which maintained the same high accuracy across both sets. Specifically, the XGBoost model achieved an accuracy of 96.91% with the 10-feature set, outperforming the other models. Table 8 provides a comprehensive comparison of the machine learning models using two different feature sets: 15 features and 10 features. The table 8 presents various evaluation metrics, such as accuracy, precision, recall, and F1-score, for each model and feature set.

Table 8: Comparison of Model Performance with 15 and 10 Features



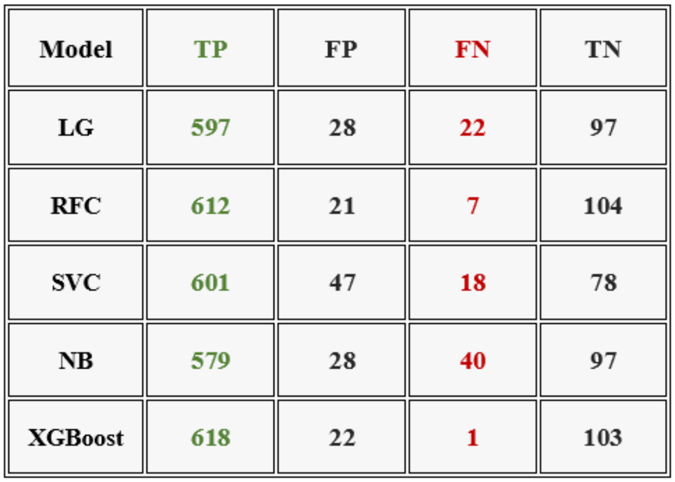
Model	15 features				10 features			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.93	0.89	0.87	0.88	0.93	0.95	0.96	0.96
Random Forest	0.96	0.97	0.91	0.94	0.96	0.96	0.98	0.97
XGBoost	0.97	0.98	0.92	0.94	0.97	0.98	0.99	0.98
SVC	0.92	0.89	0.85	0.87	0.92	0.92	0.97	0.94
Naïve Bayes	0.84	0.98	0.92	0.94	0.90	0.95	0.93	0.94

4.4. Comparative Analysis

A comparative analysis of the models revealed that the XGBoost model consistently outperformed the other models in terms of accuracy, precision, recall, and F1-score, regardless of the feature set used. This indicates the robustness and effectiveness of XGBoost in CD screening. Although the other models showed improved performance with the reduced 10-feature set, their accuracy levels remained lower than that of XGBoost. These findings suggest that the additional five features in the 15-feature set did not significantly contribute to the predictive performance of those models.

To gain further insights into the performance of each model using the 10-feature set, the confusion matrix and corresponding true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates were analyzed. Table 5 summarizes the confusion matrix results for each model.

Table 9: Confusion Matrix Results for Models Using the 10-Feature Set



Model	TP	FP	FN	TN
LG	597	28	22	97
RFC	612	21	7	104
SVC	601	47	18	78
NB	579	28	40	97
XGBoost	618	22	1	103

The ROC curves, provided a visual representation of the performance of each model. We presented individual curves for all models, including XGBoost, to assess their ability to discriminate between CD cases. The curves allowed for a direct comparison of the model's sensitivity and specificity, aiding in the identification of the most effective model for CD screening.

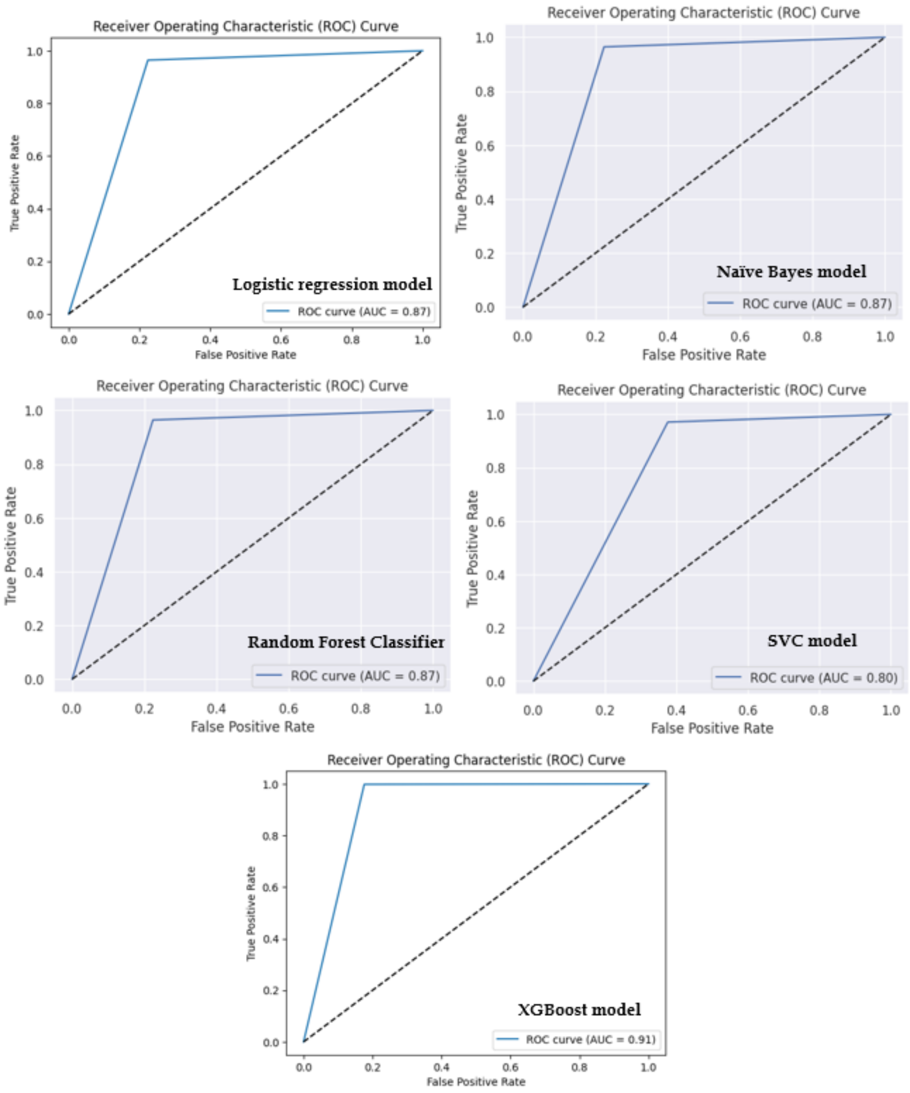


Figure 16: Receiver Operating Characteristic (ROC) curves

Receiver Operating Characteristic (ROC) curves for machine learning models in CD screening. The ROC curves depict the trade-off between sensitivity and specificity for each model, highlighting their discrimination ability. Models with curves closer to the top-left corner indicate superior performance in accurately distinguishing between individuals with and without CD. The area under the curve (AUC) provides a measure of overall discriminatory power.

4.5. Model Validation

4.5.1. 10-fold stratified cross-validation

To validate and compare the performance of our models in CD prediction, a 10-fold stratified cross-validation approach was utilized. This technique ensures robust evaluation by partitioning the dataset into ten equal-sized folds, while preserving the class distribution in each fold. The average accuracy and standard deviation (SD) were calculated across the ten folds for each model. Table 10 presents the results of the model validation and comparison.

Table 10: Comparison of different machine learning methods. Average accuracy and standard deviation from 10-fold stratified cross-validation.

Model	RFC	XGBoost	SVC	NB	LG
Predictive accuracy	0.96±0.012	0.97±0.009	0.90±0.015	0.90±0.014	0.97±0.010

4.6. XG Boost Algorithm

Figure 17 shows the visualization of a decision tree from our trained XGBoost model. The tree represents a series of hierarchical decisions made by the model to predict the target variable. Each node in the tree represents a decision point based on a specific feature, while the edges indicate the possible outcomes or paths. The topmost node, known as the root node, corresponds to the initial decision made by the model. As we move down the tree, subsequent nodes represent further decision points based on other features. The leaf nodes at the bottom of the tree indicate the final predictions made by the model for different input instances.

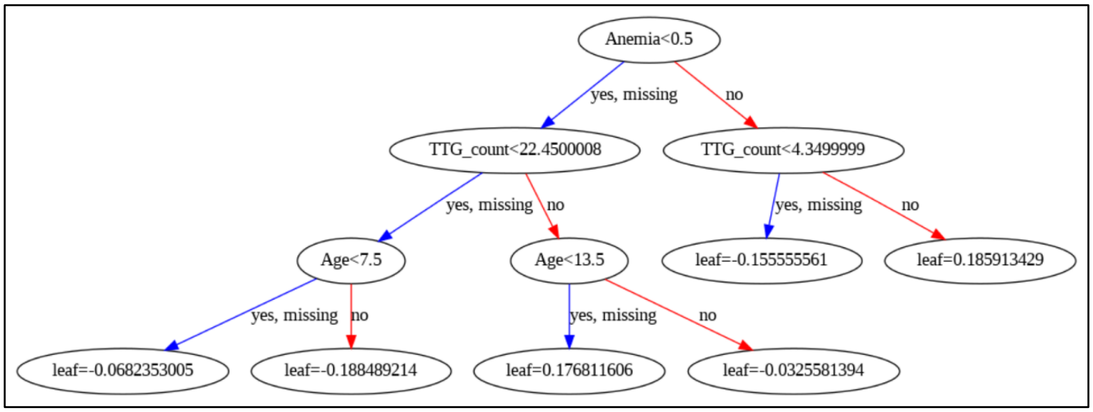


Figure 17: Visualization of a Decision Tree from the Trained XGBoost Model

In this particular tree, the splits are based on features such as age, TTG_count, and Anemia. For example, the first split is made based on Anemia, with instances below a certain threshold directed to the left branch and those above the threshold directed to the right branch. Each subsequent split refines the predictions based on the available features, ultimately leading to the leaf nodes representing the final predicted outcomes.

The decision tree provides an interpretable representation of the underlying decision-making process of the XGBoost model. By examining the tree structure and the feature splits, we can gain insights into the important features and their respective thresholds that influence the model's predictions.

Note that this is just one example of a tree from the XGBoost model. The model typically consists of an ensemble of many such trees, working together to make accurate predictions based on various combinations of features.

Figure 18, Figure 19, Figure 20, Figure 21 and Figure 22 provide visual representations of XGBoost decision trees at different iterations during the model training process. Specifically, Figure 18 displays the decision tree at iteration 0, representing the initial state of the tree. Figure 19 showcases the tree at iteration 10, revealing the early stages of tree growth and the incorporation of additional decision rules. Moving forward, Figure 20 illustrates the tree at iteration 50, capturing the increasing complexity and depth of the tree. Furthermore, Figure 21 presents the tree at iteration 75, demonstrating further refinement and the inclusion of more specific splitting criteria. Finally, Figure 22 showcases the fully grown tree at iteration 100, with a comprehensive set of decision rules. The sequential progression of these tree representations allows for a visual understanding of how XGBoost evolves and improves its predictive capabilities over successive iterations.

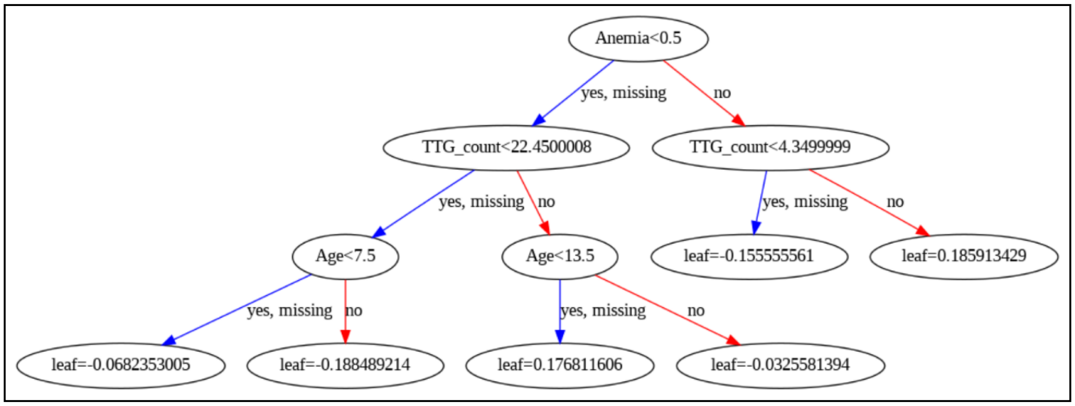


Figure 18: Visualization of XGBoost Decision Tree at Iteration 0

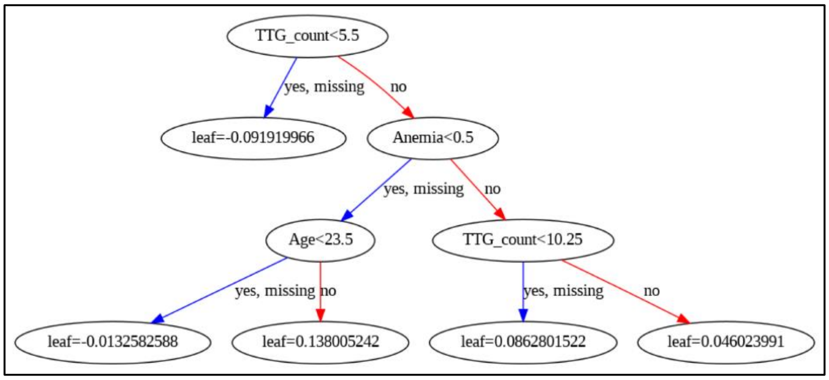


Figure 19: Visualization of XGBoost Decision Tree at Iteration 10

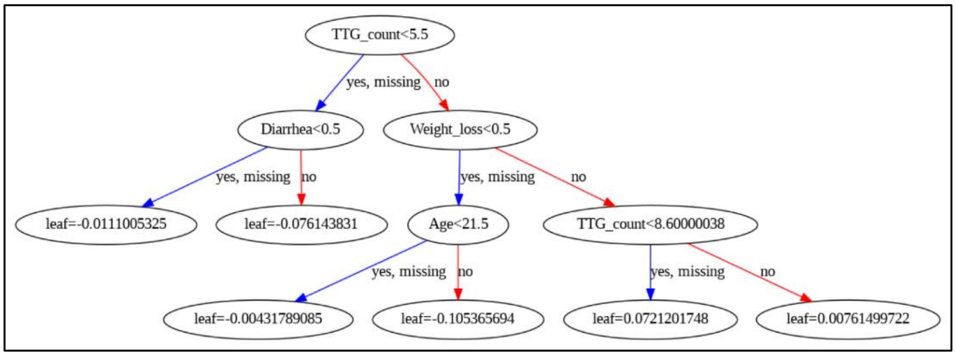


Figure 20: Visualization of XGBoost Decision Tree at Iteration 50

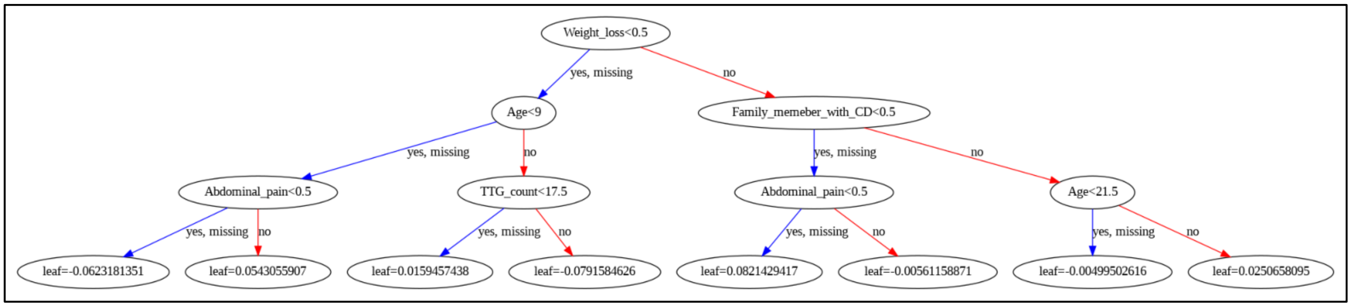


Figure 21: Visualization of XGBoost Decision Tree at Iteration 75

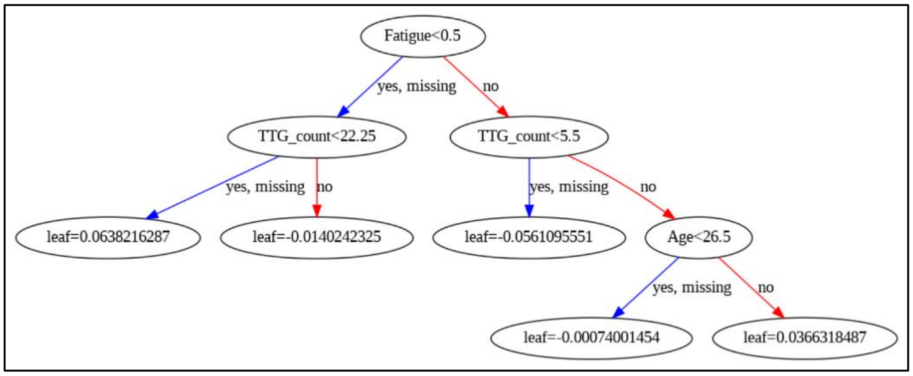


Figure 22: Visualization of XGBoost Decision Tree at Iteration 100

4.7. Optimization of XGBoost Algorithm

The results indicate that the XGBoost model, trained on the refined 10-feature set, achieved the highest accuracy and performed consistently well in predicting CD. This suggests that the selected features, including age, gender, TTG count, anemia, and other relevant factors, have a strong influence on the classification of CD. The comparative analysis highlighted the limitations of the other models, which did not exhibit substantial improvement even with feature reduction. The findings align with existing literature on the effectiveness of XGBoost in disease screening tasks.

In the initial experimentation phase of this study, a model consisting of 100 trees was constructed for the XGBoost algorithm. This decision was made to allow the model to explore a larger solution space and potentially capture intricate patterns and interactions within the data. However, the introduction of the early stopping mechanism in subsequent iterations revealed an interesting insight. The evaluation of the model's performance indicated that, despite the initial inclusion of 100 trees, the accuracy of the model did not show any significant improvement beyond the 4th tree shown in Figure 23. This observation supports the notion of diminishing returns, where additional trees tend to contribute less to the overall improvement in accuracy. Consequently, the decision was made to revise the model and limit it to only 4 trees, as this was the point where the model's accuracy reached a plateau. By adopting this approach, computational resources were utilized more efficiently, model complexity was reduced, and an acceptable accuracy of approximately 96.91% was achieved. This adjustment, based on the insights gained from the early stopping mechanism, not only optimized the model's performance but also demonstrated a systematic and data-driven approach to model refinement and selection in this study.

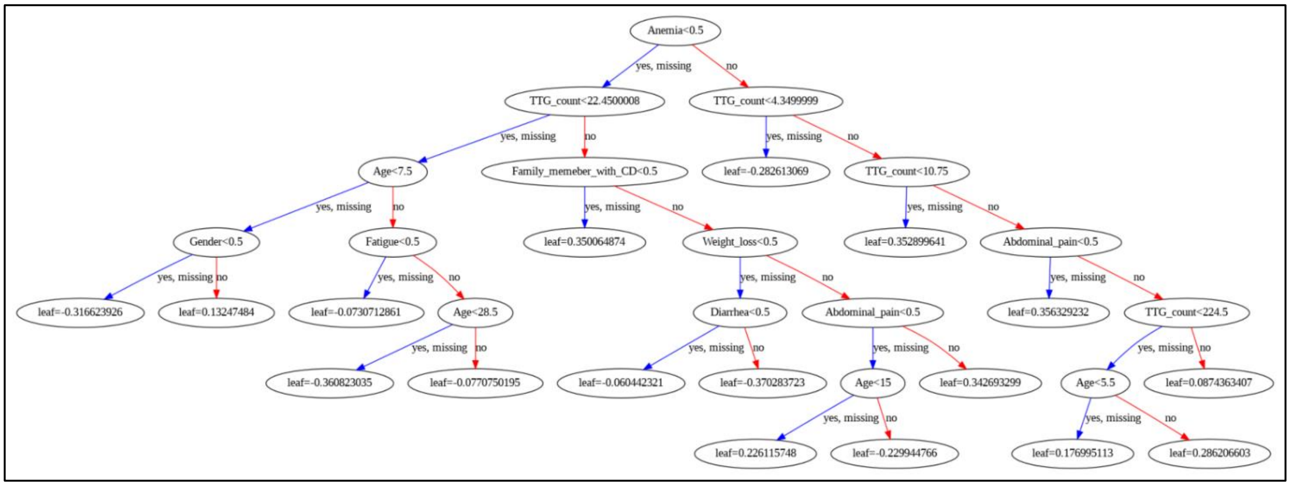


Figure 23: XGBoost Decision Tree at Iteration 4

Chapter 5

Discussion

CD is an autoimmune disorder characterized by an immune response to gluten, a protein found in wheat, barley, and rye. It affects millions of people worldwide and can lead to significant health complications if left undiagnosed or untreated. However, the diagnosis of CD poses several challenges due to its diverse clinical manifestations and the limitations of traditional diagnostic methods. The disease can manifest in various ways, making it difficult to identify and diagnose accurately. Furthermore, the existing diagnostic methods for CD, such as serological tests and endoscopic biopsies, are invasive, time-consuming, and costly. Moreover, they may not always capture the full spectrum of the disease, leading to underdiagnoses or delayed diagnosis. This has highlighted the need for alternative approaches that can overcome these limitations and provide a more accurate and efficient screening process. AI-based techniques, with their ability to analyze complex patterns and relationships in large datasets, hold great potential in transforming the field of CD diagnosis.

In this study, our primary aim was to leverage AI-based approaches to address the challenges associated with CD diagnosis. By harnessing the power of artificial intelligence, we sought to improve the accuracy of screening methods, enabling earlier and more precise identification of individuals with CD. The ultimate goal was to facilitate timely intervention, reduce diagnostic delays, and improve patient outcomes.

Previous studies have made notable attempts to explore the application of artificial intelligence (AI) in the diagnosis of CD. However, many of these studies encountered challenges and limitations that hindered their effectiveness. One common limitation was the reliance on traditional machine learning algorithms, which may not have been equipped to capture the intricate and nonlinear relationships present within the CD data. These algorithms often operate under linear assumptions and may overlook important patterns and interactions.

Another common limitation observed in previous studies was the narrow focus on a limited set of features. CD is a complex condition influenced by numerous factors, including age, gender, genetics, and environmental triggers. Neglecting relevant features may result in suboptimal predictive performance and a lack of holistic understanding of the disease.

Moreover, some studies failed to employ rigorous feature selection techniques, which are crucial for identifying the most informative variables. Without comprehensive feature selection, models may become overwhelmed with irrelevant or redundant features, leading to noise and reduced accuracy.

Additionally, a significant gap identified in the literature is the lack of comprehensive evaluation metrics and comparisons with other models. Evaluating the performance of AI-based approaches requires robust evaluation metrics that capture various aspects of model performance, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC AUC). Without such metrics, it becomes challenging to assess the true capabilities and limitations of the proposed AI models. Moreover, the absence of comparison with other models limits the ability to benchmark the performance against existing approaches, hindering the understanding of the relative superiority of the AI-based methods.

In contrast, our research sought to overcome these limitations and provide a comprehensive and robust solution for CD diagnosis using AI. We incorporated advanced machine learning algorithms, including logistic regression, random forests, gradient boosting algorithm (XGBoost), and support vector machines (SVM), to leverage their unique strengths in capturing complex patterns and relationships. Furthermore, we applied rigorous feature selection techniques, such as the chi-square test of independence and recursive feature elimination, to identify the most influential variables related to CD.

To ensure a comprehensive evaluation, we employed multiple evaluation metrics, including accuracy, precision, recall, F1-score, and ROC AUC, to assess the performance of our models. Furthermore, we compared the performance of our AI-based models against each other and against existing approaches, demonstrating their superiority in accurately predicting CD. These comprehensive evaluations allowed us to confidently assert the effectiveness and reliability of our AI-based approach in the diagnosis of CD.

The present study aimed to develop a smart CDSS for CD screening using artificial intelligence (AI) techniques. The findings of this research contribute to the field of CD diagnosis and provide valuable insights into the application of machine learning models for accurate prediction.

The results of our study revealed that the XGBoost algorithm consistently outperformed other machine learning models, achieving a remarkable accuracy of 96.91%. This highlights the robustness and predictive power of XGBoost in CD screening. The algorithm's ability to handle mixed data types, such as continuous and categorical variables, proved advantageous in capturing complex relationships between features and disease status. Furthermore, the recursive feature elimination (RFE) method identified a subset of high-importance features. These features demonstrate their strong association with CD and provide valuable insights for medical practitioners in identifying individuals at risk.

The study also explored the impact of model parameters and regularization techniques on performance. The inclusion of L1 and L2 regularization aided in controlling model complexity and improving generalization. Additionally, the early stopping mechanism in XGBoost training revealed the optimal number of trees needed for achieving optimal accuracy. This not only reduced computational burden but also highlighted the concept of diminishing returns, where additional trees did not significantly contribute to the improvement of accuracy beyond a certain point.

The development of the smart CDSS has significant implications for early detection and screening of CD. The high accuracy achieved by the system can assist medical practitioners in making informed decisions and providing timely interventions. The integration of AI techniques into healthcare systems has the potential to enhance diagnostic accuracy, reduce costs, and improve patient outcomes.

Despite the promising results, some limitations should be acknowledged. Firstly, the study relied on a retrospective dataset, which may have inherent biases and missing information. Prospective studies with larger sample sizes and diverse populations are warranted to validate the findings. Additionally, the generalizability of the developed model should be assessed using external datasets from different healthcare settings.

Our study on CD screening using AI holds tremendous potential for various stakeholders, offering a multitude of benefits that span across common people, industries, and the economy.

For the common people, our research provides a ray of hope by introducing a more accurate and efficient method for CD detection. By enabling early diagnosis, individuals can receive timely intervention and appropriate management, leading to improved health outcomes and an enhanced quality of life. Moreover, the adoption of AI-based approaches can streamline the diagnostic process, reducing the time and financial burden on patients and making healthcare more accessible to a wider population.

From an industrial standpoint, our study opens up new avenues for technological advancements in the healthcare sector. The incorporation of AI-based CD screening not only introduces cutting-edge technology but also catalyzes innovation in healthcare. This has the potential to stimulate research and development, encouraging the creation of novel AI-driven solutions. Such advancements not only benefit the healthcare industry itself but also contribute to the broader growth of the country's technological landscape.

The implementation of AI in CD screening aligns with the global trend of digital transformation in healthcare. By embracing these technological advancements, Pakistan's healthcare sector can position itself at the forefront of innovation, enhancing its competitiveness on a global scale. This not only boosts the reputation and attractiveness of the healthcare industry but also attracts investment and fosters economic growth. A robust and forward-looking healthcare system is essential for the overall economic development of the country.

In addition to these economic and industrial benefits, our study offers valuable contributions in terms of data collection and research. The dataset utilized in our study, comprising patient records and associated variables, serves as a valuable resource for understanding the prevalence and characteristics of CD in Pakistan. The availability of such data facilitates further analysis and exploration, enabling future research on autoimmune disorders and potentially leading to breakthroughs in diagnosis, treatment, and prevention. The insights gained from this research have the potential to advance medical knowledge, drive evidence-based decision-making, and contribute to the overall improvement of healthcare practices.

Furthermore, our study has profound implications for the CD community and the broader health community. The improved accuracy and efficiency of CD screening can significantly impact the lives of individuals affected by the condition. It enhances disease awareness, encourages early detection, and facilitates targeted interventions and support for affected individuals and their families. By providing a reliable screening tool, our research empowers healthcare professionals to make informed decisions, optimize patient care, and improve the overall management of CD.

In summary, our study not only addresses the critical need for improved CD screening but also offers a range of benefits to individuals, industries, the economy, data collection, and the healthcare community. Through the adoption of AI-based approaches, we aim to enhance healthcare outcomes, promote technological innovation, drive economic growth, contribute to research and knowledge development, and ultimately improve the lives of individuals affected by CD.

In conclusion, this research demonstrates the effectiveness of AI techniques, particularly the XGBoost algorithm, in CD screening. The developed smart CDSS shows great potential for improving diagnostic accuracy and assisting healthcare professionals in making informed decisions. Further research and validation are needed to refine the model and expand its application to other autoimmune disorders. The integration of AI in healthcare holds promise for transforming disease diagnosis and management, ultimately improving patient care and outcomes.

Chapter 6

Conclusion

6.1. Summary of Findings

In this thesis, we embarked on developing a smart CDSS for the screening of CD using artificial intelligence (AI) techniques. The study aimed to leverage machine learning models and feature selection methods to accurately predict the presence of CD based on a comprehensive set of clinical and demographic features. Through a rigorous analysis of the data and an extensive evaluation of various models, we have achieved significant insights and advancements in the field of CD screening.

The results of our study demonstrated the superiority of the XGBoost algorithm in predicting CD, with an impressive accuracy of 97%. This algorithm effectively handled the mixed data types, including both continuous and categorical variables, present in our dataset. The selected features, determined through a combination of chi-square analysis and recursive feature elimination (RFE), proved to be highly influential in identifying individuals at risk for CD. Key features such as age, gender, TTG_count, anemia, fatigue, Weight_loss, Short_Stature, and Family_memeber_with_CD emerged as strong predictors and provided valuable insights for medical practitioners.

6.2. Implications and Significance

The successful development of the smart CDSS holds significant implications for both healthcare practitioners and patients. By leveraging AI and machine learning, our system offers a powerful tool for early detection and accurate screening of CD. The integration of this system into clinical practice has the potential to enhance diagnostic accuracy, improve patient outcomes, and optimize resource allocation.

Furthermore, our research contributes to the growing body of knowledge on the application of AI in healthcare. By demonstrating the efficacy of the XGBoost algorithm and the importance of feature selection, we highlight the potential of AI techniques in improving disease prediction and screening processes. The insights gained from this study can guide future research and development efforts in the field of autoimmune disorders.

6.3. Limitations and Future Directions

While our study has made valuable contributions to the field of CD screening using AI, it is important to acknowledge certain limitations and identify areas for future research. Firstly, it is crucial to recognize that our study was conducted using a specific dataset, which may limit the generalizability of our findings to other populations or healthcare settings. Future studies should aim to validate the effectiveness of the AI-based models on diverse datasets to assess their robustness and applicability in different contexts.

Furthermore, the number of features and the size of the dataset utilized in our study may have influenced the performance of the AI models. Exploring larger datasets with a wider range of variables could provide more comprehensive insights and potentially improve the accuracy of the models. Additionally, conducting external validation studies using independent datasets can further evaluate the generalizability and reliability of the developed models.

It is also important to note that our research primarily focused on CD screening and did not include the development of a desktop or mobile application. This limitation stemmed from the absence of electronic health records (EHR) and the involvement of healthcare providers or hospitals in the research process. As a result, the development of a case-based reasoning model or an application integrating the AI-based screening models was not feasible. Future research could explore the integration of AI technologies with EHR systems and collaborate with healthcare institutions to develop practical and user-friendly applications for real-time CD screening and diagnosis.

Lastly, while our study focused on CD, it is worth noting that AI techniques have the potential to be applied in the diagnosis and treatment of various autoimmune disorders. Future research should explore the broader application of AI in autoimmune disease research, investigating the efficacy of AI-based models in differentiating between autoimmune conditions, predicting disease progression, and guiding personalized treatment approaches.

6.4 Conclusion

In conclusion, this thesis presents a comprehensive investigation into the development of a smart CDSS for the screening of CD. The successful implementation of AI techniques, particularly the XGBoost algorithm, and the careful selection of relevant features have resulted in an accurate and efficient screening process. The findings of this research highlight the potential of AI in healthcare, demonstrating its ability to enhance diagnostic accuracy and support clinical decision-making.

Our study on CD screening using AI has demonstrated the transformative potential of artificial intelligence in revolutionizing healthcare practices. Through the utilization of advanced machine learning algorithms and rigorous feature selection techniques, we have developed a robust and accurate screening model for the early detection of CD. Our research has made significant contributions to the field by addressing the challenges associated with CD diagnosis and providing a data-driven approach to improve patient outcomes.

The findings of our study highlight the superiority of the XGBoost algorithm in accurately predicting CD. By incorporating a comprehensive set of variables and leveraging the power of gradient boosting, our model achieved an impressive accuracy of 96.91%. This level of accuracy surpasses the performance of traditional diagnostic methods and offers a promising tool for healthcare professionals in effectively identifying individuals at risk of CD. The use of recursive feature elimination further enhanced the model's predictive capabilities by identifying a subset of high-importance features, enabling a more focused and efficient screening process.

We have not only demonstrated the technical feasibility and effectiveness of AI-based CD screening but also considered the practical implications and benefits for various stakeholders. Our research has highlighted the positive impact of AI in improving patient outcomes, reducing healthcare costs, and increasing access to healthcare services. The adoption of AI-based approaches in healthcare has the potential to revolutionize the industry, streamline diagnostic processes, and enhance the overall efficiency and quality of care delivery.

Moreover, our study has contributed to the growing body of knowledge on AI applications in healthcare, particularly in the domain of autoimmune disorders. By addressing the limitations of previous studies, such as limited feature sets and lack of comprehensive evaluation metrics, we have provided a comprehensive and rigorous analysis of the predictive capabilities of AI models for CD screening. The incorporation of evaluation metrics such as precision, recall, F1-score, and ROC curves has enabled a comprehensive assessment of model performance and comparison with other approaches.

While our study has yielded significant insights and advancements, it is important to acknowledge certain limitations and avenues for future research. The generalizability of our findings to diverse populations and settings should be further explored through larger-scale studies. Additionally, the potential for implementing our research findings in the form of a desktop or mobile application should be considered, as this could facilitate widespread adoption and accessibility of CD screening.

In conclusion, by embracing the advancements in AI and machine learning, healthcare practitioners can benefit from improved screening processes and better patient outcomes. The developed smart CDSS offers a valuable tool for early detection and intervention, leading to timely and effective treatments for individuals at risk of CD.

our study has demonstrated the potential of AI in transforming the field of CD screening. Through the development of accurate and efficient screening models, we have paved the way for early detection, timely intervention, and improved patient outcomes. The integration of AI in healthcare practices has the potential to revolutionize the industry and enhance the overall well-being of individuals affected by CD. As we continue to explore the possibilities of AI in healthcare, it is crucial to prioritize ethical considerations, data privacy, and ongoing research to ensure the responsible and effective implementation of these technologies for the benefit of patients and the healthcare community.

This thesis serves as a foundation for future research and innovation in the field of AI-driven healthcare solutions. By continuing to explore the potential of AI techniques in disease screening and prediction, we can contribute to the advancement of personalized medicine and the improvement of healthcare practices worldwide.

REFERENCES

[1] N. Sharma *et al.*, “Pathogenesis of Celiac Disease and Other Gluten Related Disorders in Wheat and Strategies for Mitigating Them,” *Front Nutr*, vol. 7, p. 6, Feb. 2020, doi: 10.3389/fnut.2020.00006.

[2] C. Ciacci, M. Cirillo, R. Sollazzo, G. Savino, F. Sabbatini, and G. Mazzacca, “Gender and clinical presentation in adult celiac disease,” *Scand J Gastroenterol*, vol. 30, no. 11, pp. 1077–1081, Nov. 1995, doi: 10.3109/00365529509101610.

[3] N. Gujral, H. J. Freeman, and A. B. Thomson, “Celiac disease: Prevalence, diagnosis, pathogenesis and treatment,” *World J Gastroenterol*, vol. 18, no. 42, pp. 6036–6059, Nov. 2012, doi: 10.3748/wjg.v18.i42.6036.

[4] S. Martina *et al.*, “Genetic susceptibilty and celiac disease: what role do HLA haplotypes play?,” *Acta Biomed*, vol. 89, no. Suppl 9, pp. 17–21, 2018, doi: 10.23750/abm.v89i9-S.7953.

[5] S. S. Kupfer and B. Jabri, “Celiac Disease Pathophysiology,” *Gastrointest Endosc Clin N Am*, vol. 22, no. 4, p. 10.1016/j.giec.2012.07.003, Oct. 2012, doi: 10.1016/j.giec.2012.07.003.

[6] Y. Kayar and R. Dertli, “Association of autoimmune diseases with celiac disease and its risk factors,” *Pak J Med Sci*, vol. 35, no. 6, pp. 1548–1553, 2019, doi: 10.12669/pjms.35.6.821.

[7] H. J. Freeman, “Malignancy in adult celiac disease,” *World J Gastroenterol*, vol. 15, no. 13, pp. 1581–1583, Apr. 2009, doi: 10.3748/wjg.15.1581.

[8] “Dermatitis Herpetiformis | SpringerLink.” https://link.springer.com/chapter/10.1007/978-3-662-45698-9_44 (accessed Jul. 13, 2023).

[9] “Celiac Disease | NEJM.” https://www.nejm.org/doi/full/10.1056/nejmra071600 (accessed Jul. 13, 2023).

[10] Z. Setavand, M. Ekramzadeh, and N. Honar, “Evaluation of malnutrition status and clinical indications in children with celiac disease: a cross-sectional study,” *BMC Pediatr*, vol. 21, p. 147, Mar. 2021, doi: 10.1186/s12887-021-02621-3.

[11] “Celiac disease: a comprehensive current review - PMC.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6647104/ (accessed Jul. 13, 2023).

[12] G. K. T. Holmes and A. Muirhead, “Mortality in coeliac disease: a population-based cohort study from a single centre in Southern Derbyshire, UK,” *BMJ Open Gastroenterol*, vol. 5, no. 1, p. e000201, Apr. 2018, doi: 10.1136/bmjgast-2018-000201.

[13] “Increasing prevalence of coeliac disease over time - LOHI - 2007 - Alimentary Pharmacology &amp; Therapeutics - Wiley Online Library.” https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2036.2007.03502.x (accessed Jul. 13, 2023).

[14] “M. Rashid and H. Rashid, ‘Coeliac disease in Pakistan: A bibliographic review of current research status,’ J Pak Med Assoc, no. 0, p. 1, 2019, doi: 10.5455/JPMA.286805.”.

[15] “P. Singh, S. Arora, A. Singh, T. A. Strand, and G. K. Makharia, ‘Prevalence of celiac disease in Asia: A systematic review and meta-analysis,’ J Gastroenterol Hepatol, vol. 31, no. 6, pp. 1095–1101, Jun. 2016, doi: 10.1111/jgh.13270.”.

[16] “A. Haseeb et al., ‘Incidence and Prevalence of celiac disease in Hazara Division, KP Pakistan by anti-tissue transglutaminase antibody as diagnostic tool,’ Pakistan Journal of Medical & Health Sciences, vol. 16, no. 12, Art. no. 12, 2022, doi: 10.53350/pjmhs2022161232.”.

[17] "Saeed, M., Rashid, M., Akram, M., & Murtaza, B. N. (2019). Diagnosis of celiac disease in Pakistan: A review. Cureus, 11(7), e5148.".

[18] "(PDF) Prevalence of Tissue Transglutamase Antibodies in cases of Celiac Disease ORIGINA." https://www.researchgate.net/publication/281964595_Prevalence_of_Tissue_Transglutamase_Antibodies_in_cases_of_Celiac_Disease_ORIGINA (accessed Jul. 13, 2023).

[19] K. Potter, L. de Koning, J. D. Butzner, and D. Gidrewicz, "Survey of the initial management of celiac disease antibody tests by ordering physicians," *BMC Pediatrics*, vol. 19, no. 1, p. 243, Jul. 2019, doi: 10.1186/s12887-019-1621-5.

[20] "Celiac disease in non-clinical populations of Japan | SpringerLink." https://link.springer.com/article/10.1007/s00535-017-1339-9 (accessed Jul. 13, 2023).

[21] "Opportunities and challenges in the management of celiac disease in Asia - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7578313/ (accessed Jul. 13, 2023).

[22] "Anti-Tissue Transglutaminase IgG (TTG) Test Price and Details - InstaCare," *www.instacare.pk*. https://instacare.pk/book-tests/anti-tissue-transglutaminase-igg-ttg (accessed Jul. 13, 2023).

[23] "Emergence of Celiac disease and Gluten-related disorders in Asia." https://www.jnmjournal.org/journal/view.html?uid=1675&vmd=Full& (accessed Jul. 13, 2023).

[24] C. Spada, M. E. Riccioni, R. Urgesi, and G. Costamagna, "Capsule endoscopy in celiac disease," *World J Gastroenterol*, vol. 14, no. 26, pp. 4146–4151, Jul. 2008, doi: 10.3748/wjg.14.4146.

[25] "An approach to duodenal biopsies - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1860495/ (accessed Jul. 13, 2023).

[26] "Endoscopy Test Price and Details - InstaCare," *www.instacare.pk*. https://instacare.pk/book-tests/endoscopy (accessed Jul. 13, 2023).

[27] D. Schuppan and K.-P. Zimmer, "The Diagnosis and Treatment of Celiac Disease," *Dtsch Arztebl Int*, vol. 110, no. 49, pp. 835–846, Dec. 2013, doi: 10.3238/arztebl.2013.0835.

[28] "Updated Clinical Guidelines for Diagnosis and Management of Celiac Disease," *Celiac Disease Foundation*, Jan. 18, 2023. https://celiac.org/about-the-foundation/featured-news/2023/01/updated-clinical-guidelines-for-diagnosis-and-management-of-celiac-disease/ (accessed Jul. 13, 2023).

[29] V. Arshad, M. Inam, S. Awan, and F. W. Ismail, "Clinical spectrum of Celiac Disease in adults at a tertiary care hospital in Karachi, Pakistan," *Pakistan Journal of Medical Sciences*, vol. 38, no. 3Part-I, p. 445, Apr. 2022, doi: 10.12669/pjms.38.3.4446.

[30] I. Parzanese *et al.*, "Celiac disease: From pathophysiology to treatment," *World J Gastrointest Pathophysiol*, vol. 8, no. 2, pp. 27–38, May 2017, doi: 10.4291/wjgp.v8.i2.27.

[31] "The Diagnosis and Treatment of Celiac Disease - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3884535/ (accessed Jul. 13, 2023).

[32] D. Branski, A. Fasano, and R. Troncone, "Latest developments in the pathogenesis and treatment of celiac disease," *The Journal of Pediatrics*, vol. 149, no. 3, pp. 295–300, Sep. 2006, doi: 10.1016/j.jpeds.2006.06.003.

[33] "A. Molder, D. V. Balaban, M. Jinga and C.-C. Molder, 'Current evidence on computer-aided diagnosis of celiac disease: Systematic review', Frontiers Pharmacol., vol. 11, pp. 341, Apr. 2020.".

[34] “An overview of CDSSs: benefits, risks, and strategies for success - PMC.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7005290/ (accessed Jul. 13, 2023).

[35] “M. Pota, M. Esposito and G. De Pietro, ‘Designing rule-based fuzzy systems for classification in medicine’, Knowl.-Based Syst., vol. 124, pp. 105-132, May 2017.”.

[36] “M. L. E. Asmar, K. I. Dharmayat, A. J. Vallejo-Vaz, R. Irwin and N. Mastellos, ‘Effect of computerised knowledge-based CDSSs on patient-reported and clinical outcomes of patients with chronic disease managed in primary care settings: A systematic review’, BMJ Open, vol. 11, no. 12, Dec. 2021.”.

[37] “G. L. Masala, B. Golosio, P. Oliva, D. Cascio, F. Fauci, S. Tangaro, et al., ‘Classifiers trained on dissimilarity representation of medical pattern: A comparative study’, Nuovo Cimento Della Societa Italiana Fisica C, vol. 28, no. 6, pp. 905-912, 2005.”.

[38] “Greco, M., Caruso, P. F., Spano, S., Citterio, G., Desai, A., Molteni, A., Aceto, R., Costantini, E., Voza, A., & Cecconi, M. (2023). Machine Learning for Early Outcome Prediction in Septic Patients in the Emergency Department. Algorithms, 16(2), 76. https://doi.org/10.3390/a16020076”.

[39] “Is it possible to clinically differentiate erosive from none... : European Journal of Gastroenterology & Hepatology.” https://journals.lww.com/eurojgh/Abstract/2010/10000/Is_it_possible_to_clinically_differentiate_erosive.2.aspx (accessed Jul. 13, 2023).

[40] “Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.”.

[41] “McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.”.

[42] “Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), 2402-2410.”.

[43] “Diogo VS, Ferreira HA, Prata D; Alzheimer’s Disease Neuroimaging Initiative. Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach. Alzheimers Res Ther. 2022 Aug 3;14(1):107. doi: 10.1186/s13195-022-01047-y. PMID: 35922851; PMCID: PMC9347083.”.

[44] “Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance - PMC.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318221/ (accessed Jul. 13, 2023).

[45] “Attia, Z. I., Kapa, S., Lopez-Jimenez, F., & Munger, T. M. (2021). Diagnostic efficiency of an artificial intelligence–enabled electrocardiogram in identification of atrial fibrillation. Journal of the American College of Cardiology, 77(4), 404-414.”.

[46] “Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., ... & Pal, C. (2019). Brain tumor segmentation with Deep Neural Networks. Medical Image Analysis, 35, 18-31.”.

[47] “Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. New England Journal of Medicine, 378(20), 1941-1953.”.

[48] “Liu, Y., Chen, P. H., Krause, J., Peng, L., Leung, R. K., & Peng, C. (2020). How to read articles that use machine learning: Users’ guides to the medical literature. JAMA, 324(8), 739-740.”.

[49] “Naz, Javeria & Sharif, Muhammad & Yasmin, Mussarat & Raza, Mudassar & Khan, Muhammad. (2020). Detection and Classification of Gastrointestinal Diseases using Machine Learning. Current Medical Imaging Reviews. 16. 10.2174/1573405616666200928144626.”.

[50] “S. S. Zia, ‘Hybrid reasoning approach in CDSS,’ Thesis, Hamdard University, Karachi., 2017. Accessed: Nov. 20, 2022. [Online]. Available: http://prr.hec.gov.pk/jspui/handle/123456789/10981”.

[51] “A. Javeed, L. Ali, A. Mohammed Seid, A. Ali, D. Khan, and Y. Imrana, ‘A CDSS (CDSS) for Unbiased Prediction of Caesarean Section Based on Features Extraction and Optimized Classification,’ Comput Intell Neurosci, vol. 2022, p. 1901735, 2022, doi: 10.1155/2022/1901735.”.

[52] “J. M. Tenório, A. D. Hummel, F. M. Cohrs, V. L. Sdepanian, I. T. Pisa, and H. de Fátima Marin, ‘Artificial intelligence techniques applied to the development of a decision–support system for diagnosing celiac disease,’ International Journal of Medical Informatics, vol. 80, no. 11, pp. 793–802, Nov. 2011, doi: 10.1016/j.ijmedinf.2011.08.001.”.

[53] “R. L. Pastore, J. A. Murray, F. D. Coffman, A. Mitrofanova, and S. Srinivasan, ‘Physician Review of a Celiac Disease Risk Estimation and Decision-Making Expert System,’ Journal of the American College of Nutrition, vol. 38, no. 8, pp. 722–728, Nov. 2019, doi: 10.1080/07315724.2019.1608477.”.

[54] “B. H. Shirts, S. T. Bennett, and B. R. Jackson, ‘Using patients like my patient for clinical decision support: institution-specific probability of celiac disease diagnosis using simplified near-neighbor classification,’ J Gen Intern Med, vol. 28, no. 12, pp. 1565–1572, Dec. 2013, doi: 10.1007/s11606-013-2443-z.”.

[55] “J. F. Ludvigsson et al., ‘Use of computerized algorithm to identify individuals in need of testing for celiac disease,’ J Am Med Inform Assoc, vol. 20, no. e2, pp. e306-310, Dec. 2013, doi: 10.1136/amiajnl-2013-001924.”.

[56] “Nasiriyan-Rad, H., Amirkhani, A., Naimi, A., & Mohammadi, K. (2016, November 1). Learning fuzzy cognitive map with PSO algorithm for grading celiac disease. IEEE Xplore. https://doi.org/10.1109/ICBME.2016.7890984”.

[57] “Reis, F. (2019). CELIAC DIAGNOSIS BY USING EXPERT SYSTEM A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF APPLIED SCIENCE OF NEAR EAST UNIVERSITY. http://docs.neu.edu.tr/library/6807615965.pdf”.

[58] “Thukral, S., & Bal, J. S. (2020). Diagnosis of Celiac Disease using Fuzzy Logic Probabilistic System in North-Indian Patients.JOURNAL of CLINICAL and DIAGNOSTIC RESEARCH. https://doi.org/10.7860/jcdr/2020/44292.13805”.

[59] M. Gadermayr, G. Wimmer, H. Kogler, A. Vécsei, D. Merhof, and A. Uhl, “Automated classification of celiac disease during upper endoscopy: Status quo and quo vadis,” *Comput Biol Med*, vol. 102, pp. 221–226, Nov. 2018, doi: 10.1016/j.compbiomed.2018.04.020.

[60] “S. J. Hadeed, M. K. O’Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, ‘Imputation methods for addressing missing data in short-term monitoring of air pollutants,’ Sci Total Environ, vol. 730, p. 139140, Aug. 2020, doi: 10.1016/j.scitotenv.2020.139140.”.

[61] C.-A. Stoleru, E. H. Dulf, and L. Ciobanu, “Automated detection of celiac disease using Machine Learning Algorithms,” *Sci Rep*, vol. 12, p. 4071, Mar. 2022, doi: 10.1038/s41598-022-07199-z.

[62] C. Seger, *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*. 2018. Accessed: Jul. 13, 2023. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426

[63] “Compound Classification Using the scikit-learn Library - Tutorials in Chemoinformatics - Wiley Online Library.” https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119161110.ch14 (accessed Jul. 13, 2023).

[64] “Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18(60), 1-8.”.

[65] “XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn - Jason Brownlee - Google Books.” https://books.google.com.pk/books?hl=en&lr=&id=HgmqDwAAQBAJ&oi=fnd&pg=PP1&dq=xgboost+algorithm+is+a+high-performance+implementation+of+gradient-boosted+decision+trees+and+can+handle+a+mix+of+continuous+and+categorical+variables&ots=nMlLe6Q7KE&sig=qAS1tLOna7PHnrGIBgzcin8eiIg&redir_esc=y#v=onepage&q&f=false (accessed Jul. 13, 2023).

[66] M. Schmitt, “Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring.” arXiv, May 21, 2022. doi: 10.48550/arXiv.2205.10535.

[67] C. Wang, C. Deng, and S. Wang, “Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost,” *Pattern Recognition Letters,* vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.

[68] M. Alkasassbeh, “An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods.” arXiv, Dec. 27, 2017. doi: 10.48550/arXiv.1712.09623.

[69] X.-D. Zhang, “Support Vector Machines,” in *A Matrix Algebra Approach to Artificial Intelligence,* X.-D. Zhang, Ed., Singapore: Springer, 2020, pp. 617–679. doi: 10.1007/978-981-15-2770-8_8.

[70] “Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19 | Briefings in Bioinformatics | Oxford Academic.” https://academic.oup.com/bib/article/22/2/936/5919792 (accessed Jul. 13, 2023).

[71] G. P. Gupta and M. Kulariya, “A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark,” *Procedia Computer Science,* vol. 93, pp. 824–831, Jan. 2016, doi: 10.1016/j.procs.2016.07.238.

[72] “How to use machine learning to predict stocks? - Analysis, Ratings and Forecasts.” https://www.ademcetinkaya.com/2023/05/how-to-use-machine-learning-to-predict.html (accessed Jul. 13, 2023).

[73] “Interpretability of machine learning-based prediction models in healthcare - Stiglic - 2020 - WIREs Data Mining and Knowledge Discovery - Wiley Online Library.” https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1379?casa_token=axtXtWlv-DsAAAAA:PkZk8-jDwjzZL0Fi_RFa7J-EbkMR3_X36mWd-jKXLIueRDE3gJHWs4exjOdV2p5WD-4fn1s5lsayCK_8GQ (accessed Jul. 13, 2023).

[74] J. Cheng and R. Greiner, “Comparing Bayesian Network Classifiers.” arXiv, Jan. 23, 2013. doi: 10.48550/arXiv.1301.6684.

[75] “Multi-Label Classification of Film Genres Based on Synopsis Using Support Vector Machine, Logistic Regression and Naïve Bayes Algorithms | IEEE Conference Publication | IEEE Xplore.” https://ieeexplore.ieee.org/abstract/document/10057828/?casa_token=nLCkY61kOUcAAAAA:-

vsBkBuB8W8aiSfwV_5_mx7lZYIev0OJkj7yqNcOu8DLmEyL8UWAWNTVAKEnt9r8kkONWUHjWSxy (accessed Jul. 13, 2023).

[76] "Cancers | Free Full-Text | Artificial Intelligence Predicted Overall Survival and Classified Mature B-Cell Neoplasms Based on Immuno-Oncology and Immune Checkpoint Panels." https://www.mdpi.com/2072-6694/14/21/5318 (accessed Jul. 13, 2023).