# A Mobile Application for Diagnosing Celiac Disease using Decision Tree with Gradient Boosting Algorithm

A Thesis Proposal  
Presented to the Faculty of the  
College of Computer and Information Sciences  
Polytechnic University of the Philippines

In Partial Fulfilment  
of the Requirements for the Degree

Bachelor of Science in Computer Science

Baleta, Gabriel Mori  
Pajanustan, Joven

September, 2017

# TABLE OF CONTENTS

[A Mobile Application for Diagnosing Celiac Disease using Decision Tree with Gradient Boosting Algorithm i](#_Toc491108738)

[TABLE OF CONTENTS ii](#_Toc491108739)

[CHAPTER 1 The Problem and Its Background 1](#_Toc491108740)

[1.1 INTRODUCTION 1](#_Toc491108741)

[1.2 BACKGROUND OF THE STUDY 1](#_Toc491108742)

[1.3 STATEMENT OF THE PROBLEM 2](#_Toc491108743)

[1.4 THEORETICAL/CONCEPTUAL FRAMEWORK 3](#_Toc491108744)

[1.4.1 THEORETICAL FRAMEWORK 3](#_Toc491108745)

[1.4.2 CONCEPTUAL FRAMEWORK 3](#_Toc491108746)

[1.5 SIGNIFICANCE OF THE STUDY 3](#_Toc491108747)

[1.6 SCOPE AND LIMITATION OF THE STUDY 4](#_Toc491108748)

[1.7 OPERATIONAL TERMS 4](#_Toc491108749)

[CHAPTER 2 Review of Related Literature 5](#_Toc491108750)

[2.1. RELATED LITERATURE 5](#_Toc491108753)

[2.2. RELATED STUDIES 8](#_Toc491108754)

[2.3. SYNTHESIS OF THE STUDY 10](#_Toc491108755)

[CHAPTER 3 Research Methodology 11](#_Toc491108756)

[3.1. RESEARCH DESIGN 11](#_Toc491108760)

[3.2. SOURCES OF DATA 11](#_Toc491108761)

[3.3. INSTRUMENTATION 12](#_Toc491108762)

[3.3.1 SOFTWARE/HARDWARE TOOLS 12](#_Toc491108763)

[3.3.1.1 SYSTEM ARCHITECTURE 12](#_Toc491108764)

[3.3.1.2 DEVELOPMENT DETAILS 13](#_Toc491108765)

[3.3.2 RESEARCH INSTRUMENT 15](#_Toc491108766)

[3.4. DATA GENERATION/GATHERING PROCEDURE 16](#_Toc491108767)

[3.5. STATISTICAL TREATMENT OF DATA 16](#_Toc491108768)

[REFERENCES 17](#_Toc491108769)

[APPENDIX 18](#_Toc491108770)

[APPENDIX A: SAMPLE RESEARCH INSTRUMENT 18](#_Toc491108771)

[APPENDIX B: COMMUNICATIONS 18](#_Toc491108772)

[APPENDIX C: SCREENSHOTS 18](#_Toc491108773)

# CHAPTER 1 The Problem and Its Background

## INTRODUCTION

Celiac disease (CD) is an autoimmune disorder that’s triggered by eating gluten-containing food such as bread products. Gluten-containing food damages the intestinal villi of people with CD, which results in impaired absorption of nutrients and its consequences, i.e., malnutrition, osteoporosis, and iron deficiency. Chronic injury to the villi also increases the risk of intestinal lymphomas (a type of cancer) in people with CD [Gonzales 2017].

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [Wikipedia 2008].

A Mobile Application for Diagnosing Celiac Disease using Decision Tree with Gradient Boosting Algorithm is an android application that evaluates a user’s risk on having Celiac Disease based on their symptoms and set of risk factors. The application outputs the percentage of risk of the user and interprets whether the user is at classical celiac disease, non-classic celiac disease, or silent celiac disease.

## BACKGROUND OF THE STUDY

This Celiac disease can be difficult to diagnose because it affects people differently. There are more than 200 known celiac disease symptoms which may occur in the digestive system or other parts of the body. Some people develop celiac disease as a child, others as an adult.

When people with celiac disease eat gluten (a protein found in wheat, rye and barley), their body mounts an immune response that attacks the small intestine. These attacks lead to damage on the villi, small fingerlike projections that line the small intestine, that promote nutrient absorption. When the villi get damaged, nutrients cannot be absorbed properly into the body.

Celiac disease is hereditary, meaning that it runs in families. People with a first-degree relative with celiac disease (parent, child, sibling) have a 1 in 10 risks of developing celiac disease.

Celiac disease is very dangerous so that if there is a way to determine if person have it better do it. In this modern era smartphones are available in the market with its affordable price, the researchers realized that this will be a helpful application for the community, because through mobile application a person can determine if he/she have a high chance of having a celiac disease, because early diagnosis of a patient’s illness will have a high chance of survivability [Celiac Disease Foundation 1998].

## STATEMENT OF THE PROBLEM

The study aims to develop and implement a system that will diagnose if a person has a celiac disease. In addition to this, the researchers intend to answer the following questions:

1. What is the accuracy of the system in diagnosing if a person has a celiac disease in terms of the following symptoms:
   * + - 1. Abdominal pain, gas, bloating
         2. Weight loss
         3. Poor growth
         4. Diarrhea
         5. Extreme Weakness
         6. Nausea, vomiting
         7. Anemia
         8. Mood swings/depression
         9. Constipation
         10. Eczema
         11. Bone/Joint Pain
         12. Mouth Ulcers
         13. Muscle Cramps
         14. Easy Bruising

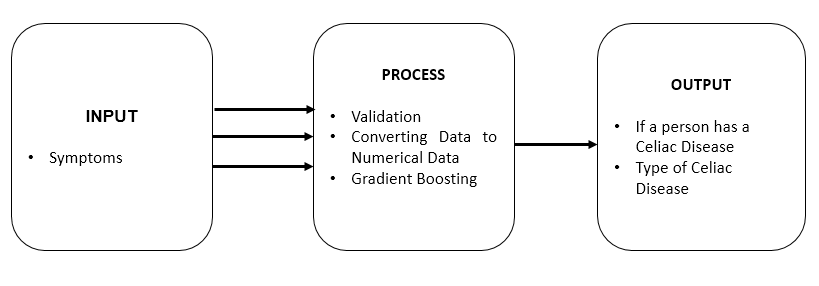
## THEORETICAL/CONCEPTUAL FRAMEWORK

## THEORETICAL FRAMEWORK

Figure 1.1 theoretical framework of the system

## CONCEPTUAL FRAMEWORK

The figure 1.2 below illustrates the conceptual framework of the system. The needed tools for the system are shown below. First, in the input phase, the user needs to answer the questions for symptoms. In the process phase includes the validation phase wherein only ages 18 and above can use this system then Gradient Boosting phase, where the data given will converted into numerical value first to process that will give the output of the system which will be the possible initial Celiac Disease diagnosis or the type of it.

****Figure 1.2 Conceptual Framework of the system

## SIGNIFICANCE OF THE STUDY

**Celiac Disease Patients**

This can be beneficial to the patients with Celiac disease for the early diagnosis. This can also help them decide to have an immediate consultation with the physician.

**Non-Celiac Disease Patients**

This can be beneficial to the patients without Celiac disease to determine if they have it and immediately consult a physician.

**Celiac Disease Experts**

The developed application will be a useful tool to give assistance for the experts in Celiac Disease. The tool can be used as an initial test / initial screening for their first-time patients to assess them if they are already at risk of having Celiac Disease.

**Future Researchers**

This will help the future researchers with the same topic of interest. This will also serve as guide or reference for them especially on the topics related to diagnosing of Celiac Disease.

## SCOPE AND LIMITATION OF THE STUDY

This system has the capacity to diagnose if the user has a Celiac Disease and the type of celiac disease only. The age input has a range from 16 years old and above only. The system has no capability of recommending an initial treatment. Although the research has reached its aims, there are still some unavoidable limitations. First, because of the limited time given, the research was conducted only on a small size of population.

## OPERATIONAL TERMS

*Celiac Disease* - is a serious autoimmune disorder that can occur in genetically predisposed people where the ingestion of gluten leads to damage in the small intestine.

*Gradient Boosting* - is a machine learning technique for [regression](https://en.m.wikipedia.org/wiki/Regression_(machine_learning)) and [classification](https://en.m.wikipedia.org/wiki/Classification_(machine_learning)) problems, which produces prediction model in the form of an [ensemble](https://en.m.wikipedia.org/wiki/Ensemble_learning) of weak prediction models, typically [decision trees](https://en.m.wikipedia.org/wiki/Decision_tree_learning).

*Android* – is a mobile operating system developed by Google. It is used by several smartphones and tablets. Examples include the Sony Xperia, the Samsung Galaxy, and the Google Nexus One. The Android operating system (OS) is based on the Linux kernel.

*Mobile Application* - most commonly referred to as an app, is a type of application software designed to run on a mobile device, such as a smartphone or tablet computer. Mobile applications frequently serve to provide users with similar services to those accessed on PCs.

*Ataxia* - is a [neurological sign](https://en.wikipedia.org/wiki/Neurological_sign) consisting of lack of voluntary [coordination of muscle movements](https://en.wikipedia.org/wiki/Motor_coordination) that includes [gait abnormality](https://en.wikipedia.org/wiki/Gait_abnormality). Ataxia is a non-specific clinical manifestation implying dysfunction of the parts of the [nervous system](https://en.wikipedia.org/wiki/Nervous_system) that coordinate movement, such as the [cerebellum](https://en.wikipedia.org/wiki/Cerebellum).

*Anemia* - is a condition that develops when your blood lacks enough healthy red blood cells or hemoglobin.

*ADHD* – or attention deficit hyperactivity disorder, is a chronic condition marked by persistent inattention, hyperactivity, and sometimes impulsivity. ADHD begins in childhood and often lasts into adulthood.

*Acid Reflux* - a condition in which acidic gastric fluid is regurgitated into the esophagus, causing heartburn.

*Lactose Intolerance* - means the body cannot easily digest lactose, a type of natural sugar found in milk and dairy products.

*Osteoporosis* - a medical condition in which the bones become brittle and fragile from loss of tissue, typically because of hormonal changes, or deficiency of calcium or vitamin D.

*Eczema* - a medical condition in which patches of skin become rough and inflamed, with blisters that cause itching and bleeding, sometimes resulting from a reaction to irritation (eczematous dermatitis) but more typically having no obvious external cause.

# CHAPTER 2 Review of Related Literature



## RELATED LITERATURE

**Decision trees for medical diagnosis**

In medical decision making (classification, diagnosing, etc.) there are many situations where decision must be made effectively and reliably. Conceptual simple decision making models with the possibility of automatic learning are the most appropriate for performing such tasks. Decision trees are a reliable and effective decision making technique that provide high classification accuracy with a simple representation of gathered knowledge and they have been used in different areas of medical decision making. In the paper we present the basic characteristics of decision trees and the successful alternatives to the traditional induction approach with the emphasis on existing and possible future applications in medicine [Podgorelec, V. et al 2002].

**Evolutionary design of decision trees for medical application**

Decision trees (DT) are a type of data classifiers. A typical classifier works in two phases. In the first, the learning phase, the classifier is built according to a preexisting data (training) set. Because decision trees are being induced from a known training set, and the labels on each example are known the first step can also be referred to as supervised learning. The second step is when the induced classifier is used for classification. Usually, prior to the first step several steps should be performed to improve the accuracy and efficiency of the classification: data cleaning, redundancy elimination, and data normalization. Classifiers are evaluated for accuracy, speed, robustness, scalability, and interpretability. DTs are widely used for exploratory knowledge discovery where comprehensible knowledge representation is preferred. The main attraction of DTs lies in the intuitive representation that is easy to understand and comprehend. Accuracy, however, is dependent on the learning data. One of the methods to improve the induction and other phases in the creation of a classifier is the use of evolutionary algorithms. They are used because the classic deterministic approach is not necessarily optimal with regard to the quality, accuracy, and complexity of the obtained classifier. In addition to the description of different evolutionary DT induction approaches, this paper also presents multiple examples of evolutionary DT applications in the medical domain [Kokol, P. et al 2012].

**Reason to test for celiac disease before you start a gluten-free diet**

You have non-celiac gluten sensitivity. Non-celiac gluten sensitivity (NCGS) is real. Some people get very sick eating gluten and they don’t have celiac disease. But, contrary to Internet lore, at this time there is no scientifically proven way to diagnose NCGS with a lab test. NCGS can only be diagnosed through a process of exclusion, under the supervision of a doctor to make sure other serious illnesses with similar symptoms are not missed. First, you rule out celiac disease (“ruling out” or “excluding” means testing for it to confirm you do not have it). If the results are negative, you can eliminate all gluten from your diet, including sources of cross-contamination. If a gluten-free diet resolves your symptoms, you and your doctor may conclude you have non-celiac gluten sensitivity. Having a celiac test is the first step in appropriately diagnosing NCGS. Celiac disease is genetic. If you have it, there is a chance your children have it or will develop it later. If you have celiac disease, but do not get it diagnosed and just assume it is non-celiac gluten sensitivity, your whole family can be impacted by this choice. Silent celiac disease. Even if your kids do not appear to have symptoms, they could have silent (asymptomatic) celiac disease. People with silent celiac disease are not aware of any symptoms, but damage is being done internally. The key to long-term health with celiac disease is early diagnosis and treatment. If you don’t test yourself for your own sake, do it for your children’s sake. [Buckhart, A 2013].

**How important is the timing of gluten introduction for children with celiac disease?**

The publication by Norris et al. highlights two major aspects of celiac disease. First, the complex nature of its etiology and pathogenesis, and second, the practical issue of what parents can do to minimize the risk of their children developing the disease. Contradictory advice concerning breast-feeding and the timing of gluten introduction into the infant diet has been offered to mothers for many years despite the lack of solid evidence. Celiac disease occurs as a result of the interplay between genetic and environmental factors. For its development an individual must carry the alleles that encode for the HLA CLASS II MOLECULES DQ2 or DQ8 (related to HLA-DR3 and HLA-DR4) and ingest gluten.1 The HLA genes, however, only account for 40% of the genetic influence for this disorder and furthermore these genes occur in up to 40% of the population. In addition, a vast number of people ingest wheat products. Why then does celiac disease only occur in 1% of the population when so many are at risk? The answer lies in the other genetic and environmental factors that contribute to the disease. Evidence shows that environmental factors contribute to the development of celiac disease in childhood. Some studies have shown that breast-feeding might modulate the presentation of the disease. One study revealed that children with celiac disease who were breast-fed had delayed onset and less severe, more atypical presentations of celiac disease compared with children who were not breast-fed.2 In addition, information has been provided by an analysis of the Swedish epidemic of celiac disease in infants in the early 1980s.3 Factors associated with this epidemic were lack of breast-feeding, large amounts of gluten in infant formula and the occurrence of infections in early life. The greatest protection was afforded to those infants who received small amounts of gluten while still being breast-fed.4 National guidelines were introduced following the publication of these studies, which resulted in a reduction in the number of children diagnosed with celiac disease [Norris J et al 2005].

**The Many Faces of Celiac Disease: Clinical Presentation of Celiac Disease in the Adult Population**

The major modes of presentation of patients with celiac disease are the classic diarrhea-predominant form and silent celiac disease. Those with silent celiac disease lack diarrhea, although they may present with manifestations of celiac disease that include an irritable bowel syndrome, anemia, osteoporosis, neurologic diseases, or malignancy. A significant proportion of patients are diagnosed through screening at-risk groups including relatives of patients and insulin-dependant diabetics. Nondiarrheal presentations now are seen more commonly than those with diarrhea. Patients with celiac disease have a greater burden of disease than the general population because of autoimmune diseases and malignancies. There is a need for screening studies of patients with conditions associated with celiac disease to determine whether the large numbers of people with undiagnosed celiac disease currently are seeking health care [Green, P H. R. 2005].

Cites literature to present topics related to the study.

## RELATED STUDIES

**The application of a decision tree to establish the parameters associated with hypertension**

Hypertension is an important risk factor for cardiovascular disease (CVD). The goal of this study was to establish the factors associated with hypertension by using a decision-tree algorithm as a supervised classification method of data mining. Data from a cross-sectional study were used in this study. A total of 9078 subjects who met the inclusion criteria were recruited. 70% of these subjects (6358 cases) were randomly allocated to the training dataset for the constructing of the decision-tree. The remaining 30% (2720 cases) were used as the testing dataset to evaluate the performance of decision-tree. Two models were evaluated in this study. In model I, age, gender, body mass index, marital status, level of education, occupation status, depression and anxiety status, physical activity level, smoking status, LDL, TG, TC, FBG, uric acid and hs-CRP were considered as input variables and in model II, age, gender, WBC, RBC, HGB, HCT MCV, MCH, PLT, RDW and PDW were considered as input variables. The validation of the model was assessed by constructing a receiver operating characteristic (ROC) curve.The prevalence rates of hypertension were 32% in our population. For the decision-tree model I, the accuracy, sensitivity, specificity and area under the ROC curve (AUC) value for identifying the related risk factors of hypertension were 73%, 63%, 77% and 0.72, respectively. The corresponding values for model II were 70%, 61%, 74% and 0.68, respectively.We have developed a decision tree model to identify the risk factors associated with hypertension that maybe used to develop programs for hypertension management [Tayefi, M. et al 2016].

**Men with celiac disease are shorter than their peers in the general population**

Objective Late diagnosis of celiac disease (CD) is increasingly common, the implications of which are largely unknown. Although short stature is a common sign of childhood CD, the data on the height of adult CD patients is conflicting. This study investigates the final height of men and women diagnosed with CD in adulthood and attempts to identify influencing factors. Patients and methods, we performed a cross-sectional study of 585 adults at the Celiac Disease Center at Columbia University, comparing their height with the control population (NHANES). Patients were included if they were older than 18 years of age at diagnosis and if baseline height and weight were available. In addition, we examined for differences in demographic and physical features, mode of presentation, and concomitant illnesses in shorter versus taller celiac patients. Results Men (n= 162) with CD diagnosed in adulthood were shorter than men in the general population (CD: 169.3±10.5 vs. 177.3±7.0 cm, P < 0.01) whereas women (n= 423) were not (CD: 166.3±9.4 vs. 163.2±6.7 cm). There were no statistically significant differences in age at diagnosis, BMI, concomitant autoimmune illnesses (hypothyroidism, type I diabetes, dermatitis herpetiformis) [Sonti, R. et al 2013].

**Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data**

Recent reports from our laboratory and others support the SELDI Protein Chip technology as a potential clinical diagnostic tool when combined with n-dimensional analyses algorithms. The objective of this study was to determine if the commercially available classification algorithm biomarker patterns software (BPS), which is based on a classification and regression tree (CART), would be effective in discriminating ovarian cancer from benign diseases and healthy controls. Serum protein mass spectrum profiles from 139 patients with either ovarian cancer, benign pelvic diseases, or healthy women were analyzed using the BPS software. A decision tree, using five protein peaks, resulted in an accuracy of 81.5% in the cross-validation analysis and 80% in a blinded set of samples in differentiating the ovarian cancer from the control groups. The potential, advantages, and drawbacks of the BPS system as a bioinformatic tool for the analysis of the SELDI high-dimensional proteomic data are discussed [Vlahou, A. et al 2002].

**Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness**

Dengue is re-emerging throughout the tropical world, causing frequent recurrent epidemics. The initial clinical manifestation of dengue often is confused with other febrile states confounding both clinical management and disease surveillance. Evidence-based triage strategies that identify individuals likely to be in the early stages of dengue illness can direct patient stratification for clinical investigations, management, and virological surveillance. Here we report the identification of algorithms that differentiate dengue from other febrile illnesses in the primary care setting and predict severe disease in adults. A total of 1,200 patients presenting in the first 72 hours of acute febrile illness were recruited and followed up for up to a 4-week period prospectively; 1,012 of these were recruited from Singapore and 188 from Vietnam. Of these, 364 were dengue RT-PCR positive; 173 had dengue fever, 171 had dengue hemorrhagic fever, and 20 had dengue shock syndrome as final diagnosis. Using a C4.5 decision tree classifier for analysis of all clinical, haematological, and virological data, we obtained a diagnostic algorithm that differentiates dengue from non-dengue febrile illness with an accuracy of 84.7%. The algorithm can be used differently in different disease prevalence to yield clinically useful positive and negative predictive values. Furthermore, an algorithm using platelet count, crossover threshold value of a real-time RT-PCR for dengue viral RNA, and presence of pre-existing anti-dengue IgG antibodies in sequential order identified cases with sensitivity and specificity of 78.2% and 80.2%, respectively, that eventually developed thrombocytopenia of 50,000 platelet/mm3 or less, a level previously shown to be associated with hemorrhage and shock in adults with dengue fever. This study shows a proof-of-concept that decision algorithms using simple clinical and hematological parameters can predict diagnosis and prognosis of dengue disease, a finding that could prove useful in disease management and surveillance [Tanner, L. et al 2008].

**The clinical decision analysis using decision tree**

Clinical practice, aimed at solving clinical problems of indi­vidual patients, is an act of continual decision-making. The best clinical decision refers to making a choice that maximizes effec­tiveness and minimizes harm. Nevertheless, when the sup­porting evidences were scant, decision-making depends on the subjective intuition of the physician and then becomes unpre­dictable and non-reproducible. Around the 1990s, Evidence-based Medicine (EBM), which was suggested as a methodology for making clinical decisions based on the best evidence, expanded across the entire field of healthcare, and the terminology “evidence-based deci­sion-making” was introduced. By overcoming the com­plexity of medical environment and the uncertainty of clinical decisions, the EBM aims to pursue qualitative improvements in healthcare. Because clinical decisions are also directly related to the development and expansion of clinical treatment guidelines, approval of new drugs, drug pre­scriptions, the applicability of medical insurance for procedures, and healthcare policies [Jong, M. 2014].

## SYNTHESIS OF THE STUDY

Base on the researchers, Celiac Disease is a genetic, so if any of your ancestors have Celiac Disease you will have a chance to have it too, and there are type of this disease this is called Silent Celiac Disease, the patients with this type of celiac disease is not aware that they already have Celiac Disease unless they will take test to determine if they have it.

The silent type of this disease attracted the researchers to develop such expert system that would be able to diagnose whether a user have Celiac Disease and type of it or not. There is no mobile application in the industry that can diagnose if the person has a chance of celiac disease, and gradient boosted decision tree algorithm is an excellent choice for diagnosing celiac disease so the researchers decided to use it, furthermore mobile application like this will be a helpful to the community because of its availability to Playstore for android or App Store for IOS smart phones.

# CHAPTER 3 Research Methodology



## RESEARCH DESIGN

The experimental method of research is used this study, Mobile Application for Diagnosing Celiac Disease using Decision Tree with Gradient Boosting Algorithm. Experimental research describes the process that a researcher undergoes of controlling certain variables and manipulating others to observe if the results of the experiment reflect that the manipulations directly caused the particular outcome. This is an experiment where one variable can be manipulated and randomized or control the other variables.

The researchers define this study as experimental because factors of heart attack are controlled and can be manipulated to see if there is effect on the risk of having celiac disease. From this, the researchers were able to test the accuracy of the system by comparing the results of the system based on the gastroenterologist’s diagnosis which is the experts to evaluate the system.

## SOURCES OF DATA

The symptoms presented in this study came from Celiac Disease: Evaluation of the Diagnosis and Dietary Compliance in Canadian Children (Rashin, M. et al, 2005), the research contains the data of Canadian children from Canadian Celiac Association, 5240 were surveyed with a questionnaire. Of the 2849 respondents with biopsy-confirmed celiac disease. Simple random sampling is the sampling technique to be used in this study.

## INSTRUMENTATION

## SOFTWARE/HARDWARE TOOLS

### SYSTEM ARCHITECTURE

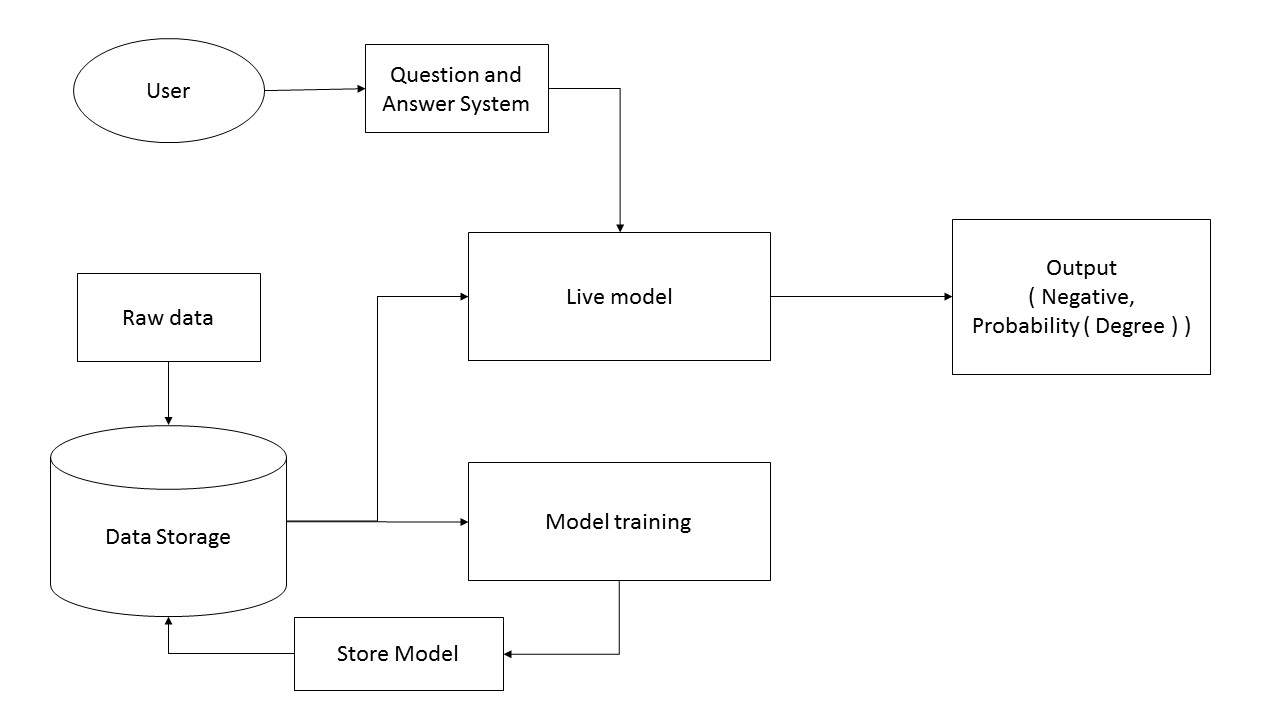


Figure 3.1 System Architecture

Figure 3.1 explains the flow of the system architecture of Celiac Disease Diagnosis Using Decision Tree with Generated Boosting. It begins with user answering the question of the system then the data will pass to the live model then the model will output the result if it is negative or if the result is positive then the system will output the level of the disease. Raw data is consisting of training data and test data and it is stored in the data storage. The raw data is used for training and improving the model which. The trained model will be stored in the data storage which will be used as the live model.

TABLE 1. Clinical Symptoms or Signs Prior to Diagnosis of Celiac Disease

|  |
| --- |
| Abdominal pain, gas, bloating |
| Weight loss |
| Poor growth |
| Diarrhea |
| Extreme weakness |
| Nausea, vomiting |
| Anemia |
| Mood swings/depression |
| Constipation |
| Eczema |
| Bone/joint pain |
| Mouth ulcers |
| Muscle cramps |
| Easy bruising |

These symptoms that were answered by the user will be saved to the working storage and will be processed by the model to generate the result.

Generated Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables

Decision tree algorithm is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

### DEVELOPMENT DETAILS

Figure 3.2 illustrates the series of steps done during and for the development of a software using Fountain Model.

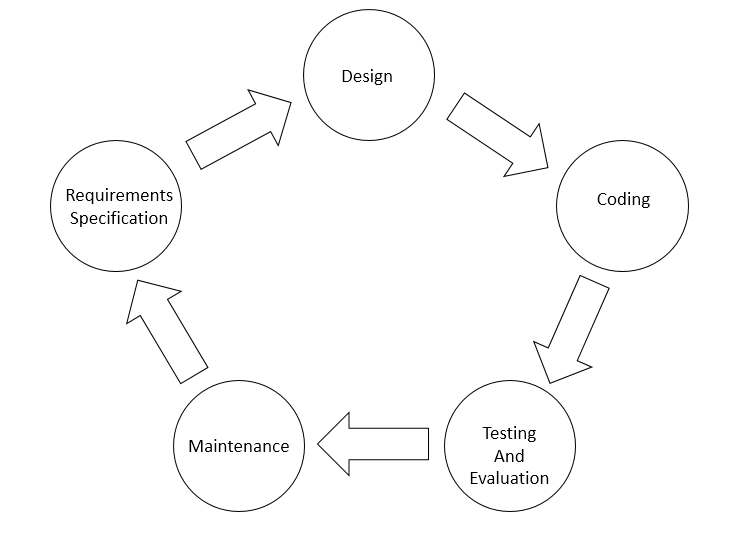


Figure 3.2 Fountain Model

**Requirement Specification**

The requirements that will be needed in the development of the system are analyzed in this phase.

- From the User – Risk factors and symptoms

- From the Gastroenterologist – How each risk factor and symptom affects the risk of having Celiac Disease.

**Design**

This phase includes the development of design of the system. The user interface as well as the set of rules that will be needed to calculate the risk of having celiac disease.

**Coding**

This is where the specifications and design are translated into code in order to develop the system. The researchers used Android Studio and Java as the programming language in developing the system.

**Testing and Evaluation**

During the testing phase, the formulated medical cases were used as the input to the system using Firefly Aurii. Later on, the system’s diagnosis report was evaluated by the gastroenterologist. And to get the accuracy of the system in diagnosing the risk of celiac disease, the diagnosis of the system was compared was compared to the diagnosis of the endocrinologist.

## 3.3.2 RESEARCH INSTRUMENT

Present The instrument that is used by the researchers in this study was experiment paper to gather the needed data for getting the accuracy of the system in diagnosing heart attack. This experiment paper was used to evaluate the system and compare the result of the system from the diagnosis of the experts.

In addition, the instrument is validated by consultant and doctor who is a licensed gastroenterologist for the reliability of the experiment paper.

Table 2 show the list of risk factors and symptoms of heart attack that were answered based on a patient that has been diagnose having risk of heart attack.

**Table 2. Experiment Paper**

|  |  |
| --- | --- |
| Case No. N | |
| Age | Age in Years |
| Abdominal pain, gas, bloating | Never, Sometimes, often, always |
| Weight loss | Yes or No |
| Poor growth | Yes or No |
| Diarrhea | Never, Sometimes, often, always |
| Extreme weakness | Never, Sometimes, often, always |
| Nausea, vomiting | Never, Sometimes, often, always |
| Anemia | Yes or No |
| Mood swings/depression | Never, Sometimes, often, always |
| Constipation | Never, Sometimes, often, always |
| Eczema | Yes or No |
| Bone/joint pain | Never, Sometimes, often, always |
| Mouth ulcers | Never, Sometimes, often, always |
| Muscle cramps | Never, Sometimes, often, always |
| Easy bruising | Never, Sometimes, often always |

.

## DATA GENERATION/GATHERING PROCEDURE

**Preliminaries**

For the data gathering, the researchers use the research, Celiac Disease: Evaluation of the Diagnosis and Dietary Compliance in Canadian Children (Rashin, M. et al, 2005), to get the symptoms of celiac disease.

**Experimental**

For the experimentation, the researchers conducted an interview from licensed Gastroenterologist of East Avenue Medical Center to serve as expert to evaluate the system.

The researchers used 50 medical cases that was formulated with the expert to be used as the input to the system. From that, the diagnosis of the experts and the system was compared to be the basis of the system’s accuracy on predicting the risk of having Celiac Disease.

## STATISTICAL TREATMENT OF DATA

The objective of the study is to obtain the accuracy of the system in diagnosing if a person has celiac disease and show the degree if the person is probable of having it. The following formula are used to calculate the system’s accuracy.

The formula for accuracy is used to provide the reliability of the system. It is determined using this formula:

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 = 100% − 𝐸𝑟𝑟𝑜𝑟 𝑃𝑒𝑟𝑐𝑒𝑛𝑡𝑎𝑔𝑒

Equation 1. Accuracy Formula

Equation 2. Error Percentage Formula

Where:

𝐶𝐷 : Total number of Cases that were correctly diagnosed by the system

(The system’s diagnosis and Expert’s Diagnosis are the same).

𝑇𝑜𝑡𝑎𝑙 : Total number of Cases diagnosed

# REFERENCES

|  |  |
| --- | --- |
| [Gonzales, 2017] | Gonzales, Eduardo MD. Pros and con of a gluten-free diet (http://www.pchrd.dost.gov.ph/index.php/news/library-health-news/6140-pros-and-con-of-a-gluten-free-diet(2015)) |
| [Wikipedia 2008] | Wikipedia. Gradient Boosting (https://en.wikipedia.org/wiki/Gradient\_boosting (2008)) |
| [Celiac Disease Foundation 1998] | Celiac Disease Foundation. What is Celiac Disease? (https://celiac.org/celiac-disease/understanding-celiac-disease-2/diagnosing-celiac-disease/diagnosis (1998)) |
| [Podgorelec, V. et al 2002] | Podgorelec, V. Decision Trees: An Overview and Their Use in Medicine (2002). |
| [Kokol, P. et al. 2012] | Kokol, P. et al. Evolutionary design of deciosion trees for medical application (2012). |
| [Buckhart, A. 2013] | Buckhart, A. Six reasons to test for celiac disease before you start a gluten-free diet  (http://theceliacmd.com/six-reasons-to-test-for-celiac-disease-before-starting-a-gluten-free-diet/(2013)) |
| [Norris, J. et al 2005] | Norris, J. et al. How important is the timing of gluten introduction for children with celiac disease? (2005). |
| [Green, Peter H. R. 2005] | Green, P H. R. The Many Faces of Celiac Disease: Clinical Presentation of Celiac Disease in the Adult Population (2005). |
| [Tayefi, M. et al 2016] | Tayefi, M. et al. The application of a decision tree to establish the parameters associated with hypertension (2016). |
| [Sonti, R. et al 2013] | Sonti, R. et al. Men with celiac disease are shorter than their peers in the general population (2013). |
| [Vlahou, A. et al 2002] | Vlahou, A. et al. Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data (2002). |
| [Tanner, L. et al 2008] | Tanner, L. et al. Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness (2008) |
| [Jong, M. 2014] | Jong, M. The clinical decision analysis using decision tree (2014) |
|  |  |

# APPENDIX

## APPENDIX A: SAMPLE RESEARCH INSTRUMENT

The tool/s to be used for collecting information. (Experiment paper or Survey Questionnaire)

## APPENDIX B: COMMUNICATIONS

Letters

## APPENDIX C: SCREENSHOTS

Screen capture of the proposed technology.