RESEARCH ARTICLE OPEN ACCESS

Predict Future Diseases using Grocery List

Nirasha Chamudini*, Amasha Palpita**,BuddiS. Lakshan**,Hansi De Silva**,Dilani Lunugalage**,Sanvitha Kasthuriarachchi**

*(Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email: it19204062@my.sliit.lk
** (Department of Information Technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email: it19171920@my.sliit.lk)
**(Department of Information Technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email :it19060972@my.sliit.lk)
**(Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email :hansi.d@my.sliit.lk)
**(Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email : dilani.l@my.sliit.lk)
**(Department of Information Technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
Email : sanvitha.k@my.sliit.lk)

----------**********----------

Abstract:

At present, people are experiencing a slew of medical issues due to poor dietary habits. It is very important to understand how the foods and beverages they consume affect their future health. People will be able to avoid unhealthy eating habits if there is a way to determine how healthy their diet is and what diseases they may get in the future as a result of the foods they eat. The grocery list is a key factor to get an idea of the foods and beverages that people consume. In this research, novel approaches are introduced to detect and predict the future diseases of the respective grocery list users by analyzing the amount of nutrients like protein, carbs, fat, and fiber in the grocery list. The research consists of a mobile application that uses a trained machine learning model to predict diseases. The system is capable of predicting three specific diseases, namely high blood pressure, cholesterol, and diabetes. The results obtained show that a high accuracy of 84% on Decision Tree model. Th proposed system is incredibly useful for users, as they can be aware of the risk of contracting diseases and take appropriate measures to prevent those diseases. Users can get many benefits such as eliminating health disparities, saving money and improving quality of life.

Keywords **—Machine Learning, Grocery List, Predicting Future Diseases**

----------**********----------

I. INTRODUCTION

Health is the most important asset of human life because living a healthy life directly affects human's well-being. Therefore, it is essential for everyone to take care of their own health. Food is one of the most significant factors that directly affects human health, and these days humans suffer from a variety of diseases as a result of consuming various types of foods in inappropriate amounts. According to the WHO, obesity, malnutrition, diabetes, heart disease and stroke [1] are some of the diseases that can occur if humans eat an unhealthy diet. Eating a high fat diet, eating a high sugar diet and eating starchy foods can lead to these diseases. If a person gets such diseases, they have to spend a lot of time on their treatment. Therefore, it really affects to waste time. In addition, the cost of

treatment for such diseases can be very high. According to the WHO, 41 million people die each year due to chronic diseases [1]. Decreased efficiency and weakness are another disadvantage of chronic illness.

Food contains major nutrients such as protein, carbohydrates, fats and fiber. Therefore, there are appropriate amounts that people should take of those nutrients based on certain factors such as age, gender, physical activity and weight. If people do not get appropriate amounts of food, they may suffer from the above-mentioned diseases.

Diabetes is a major disease that affects people due to unhealthy food and improper food consumption. It is a chronic disease that affects how your body turns food into energy. After a healthy person gets foods, his body converts those foods into glucose and releases it into the person's bloodstream. When the person's blood glucose level increases, his pancreas releases insulin to convert that glucose to energy. But, if a person gets diabetes, his pancreas doesn't release enough insulin [2]. Therefore, eating a high-sugar diet will be a major cause of diabetes.

High cholesterol is another disease that people get due to improper diet. High cholesterol means having too much fat in a person's blood. It's also called hypercholesterolemia. A person's body needs the right amount of fat to function. If a person has too much fat, his body cannot use all of it. Then that condition is considered high cholesterol [3]. Therefore, eating a high-fat diet will be a major cause of high cholesterol.

A person can get high blood pressure as a result of eating large amounts of salt, fatty meat, vegetable oil, margarine and bread. It's also called hypertension. When a person's blood pressure is higher than normal, it is considered as high blood pressure.

People will be able to get a proper diet if they have a way of knowing in advance if they are at risk of getting a disease in the future as a result of their bad eating habits. Also, they will be able to adopt appropriate preventive measures. In addition, early detection of future diseases can prevent time, money and bad effects. Unfortunately, methods of early detection of diseases caused by an unhealthy diet are extremely rare nowadays.

Usually, if we are buying foods and other beverages from a shop, we use a grocery list. If not, at least we will get a bill from the shop. Hence, a grocery list is ideal for capturing the details of users' food consumption over a period of time. Therefore, the aim of this research is to develop a future disease prediction system by using the grocery list. Additionally, a speech recognition input option is provided to enter food lists for users who do not have a proper grocery list.

II. LITERATURE REVIEW

Researchers have been conducting various researches with the similar approaches. But none of them had used the grocery list to predict diseases.

Machine Learning Models to Predict Childhood and Adolescent Obesity [4] has reviewed a large number of models to predict childhood/adolescent obesity. They have grouped all models into two types as statistical ones and machine learning ones. They have reviewed those two types of models using various types of variables. Parental BMI, sex and birth weight, smoking mother during gestation, weight gain at some previous period, parental education, exclusive breastfeeding during some initial period [4] are some variables which used for review statistical models. Dietary habits and physical activity related measures are some variables which are used for machine learning models. Finally, they have concluded that ML/DL approaches offer extraordinary advantages and new insights for childhood and adolescent obesity prediction and prevention over statistical methods. Also, they have especially mentioned ML/DL models using EHR.

Another research on disease prediction using machine learning through big data from healthcare communities [5] has used machine learning algorithms to effectively predict the prevalence of chronic diseases in disease-rich communities. They tested the modified prediction models based on real-life hospital data collected from a hospital.

They have proposed a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospitals and it has shown 94.8% accuracy than the CNN-based unimodal disease risk prediction algorithm.

Considering the relationship between health and the grocery list, some research has been done. Some research papers have suggested that shopping with a list may be a useful tool to improve diet or reduce BMI [6] and reduce obesity and its related health conditions [7]. Therefore, such things prove that it is more effective to predict future diseases using the grocery list.

Dhiraj Dahiwade, Prof. Gajanan Patle and Prof. EktaaMeshram conducted research on "Designing Disease Prediction Model Using Machine Learning Approach" to predict diseases based on symptoms of the patient [8]. KNN model and CNN model were trained to identify the most accurate model. The data set was collected from the UCI machine learning website and their proposed system is able to display the level of risk as low risk, normal risk and high risk after providing predictions. Finally, they have identified that CNN model is the most accurate and efficient model as it showed 84.5% accuracy score which is more than accuracy score of the KNN.

In one of the reviewed papers [9], all the supervised ML models were compared for disease prediction. Several algorithms were compared by analyzing several previous researches and they have identified the decision tree, random forest and support vector machine models showed a high percentage for the number of times this algorithm showed superior accuracy.

There are so many disease prediction research papers available and most of the research describe and evaluate systems that predict diabetes diseases [10], predict heart diseases [11] and other chronic diseases [12] which use various machine learning algorithms. Research has been conducted using several key factors such as blood test results as well as external factors such as age and weight [10]. Also, there are some research papers available for disease prediction based on symptoms [13].

As discussed above, many current systems and research proposed systems have the ability to predict diseases, but none of them describe disease predictions using the details of human food consumption. There are currently no other mobile applications for predicting future diseases and the proposed system has many additional features compared to other health related mobile applications.

Although there are several research papers on health using the grocery list [7], [6], most of them focus only on things like healthy diet, lower BMI and obesity. But the proposed system is expected to predict more chronic diseases and it will include some special features like Sinhala and English speech recognition.

When observe the previous project based on mobile health applications most of the applications predict the future diseases based on the symptoms. Symptomate is a mobile application where the user can enter the symptoms and some basic information about the user. Then the application will look in to the symptoms given and ask several follow-up questions. After analyzing the answers using advanced AI the application will predict how serious the symptoms, the cause for the symptoms, the actions that should take and then suggest certain lab test accordingly. SickPredict is another mobile application developed using ML algorithms to predict the daily sick number using fitness metrics.

III. METHODOLGY

The main objective of this research is to develop a mobile application that can predict future diseases using a grocery list. That main objective consists of few sub objectives.

a) Grocery list insertion
b) Calculating total nutrients amount of the grocery list
c) Predicting diseases that users may get in the future

The user should enter their grocery list into the mobile application using the form input, text

recognition, or speech recognition options. The entered grocery list is sent to a created API, which calculates the amount of protein, carbohydrates, fats, and fiber included in that list. Then, the calculated nutrients amount and other details such as age, gender and number of dates are fed into the trained machine learning model, which provides an output with predicted diseases. All of the processes are summarized in Figure 1.

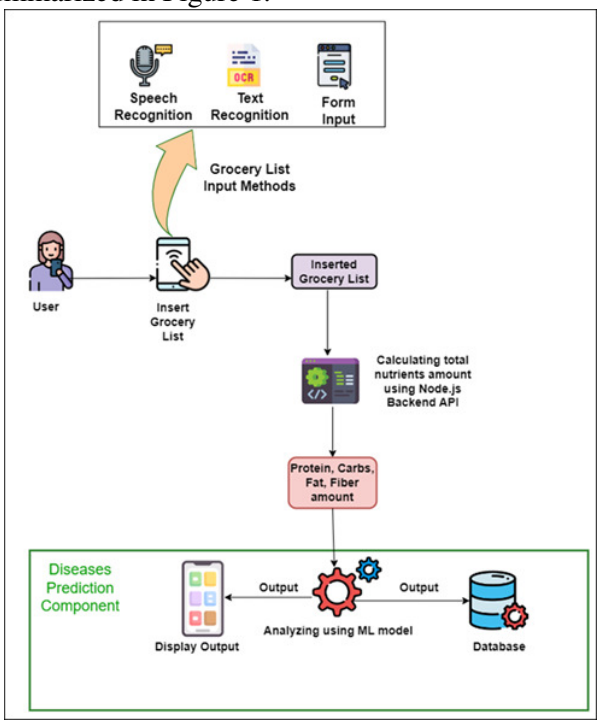

Figure 1 - System architecture diagram

A. Data Collection

Dataset was collected from Kaggle online community platform [14] and was created using data compiled from the National Health and Nutrition Examination Survey of the US in 2013-2014 [15]. The complete dataset consists 9813 records. Also, it includes five csv files called Demographics Data, Dietary Data, Examination Data, Laboratory Data and Questionnaire Data. But only Demographic data, Dietary data and Questionnaire data files were used to train the model. Each file's columns are named using a list of variables [16] [17] [18].

B. Machine Learning Models Deployment

Machine models were developed using Jupyter Notebook platform in python. Additionally, pandas, numpy, seaborn and scikit-learn were used for the ML model development.

Data preprocessing is a mandatory step when developing a machine learning model, as completing that process ensures that the data will be cleaned and made suitable for a machine learning model, raising the model's accuracy and effectiveness.

When implementing models for this research there are several data preprocessing techniques followed by me. First, all unnecessary columns are removed. Then, all existing null values were replaced with the median of each column.

After data preprocessing, “protein”, “carbs”, “fiber”, “fat”, “gender”, “age” columns were selected as features and “hypertension”, “cholesterol”, “diabetes” columns were selected as labels. Then, the data set is divided into two parts. The first part is 80% of the dataset for training the ML model and the second part is 20% of the dataset for testing purpose.

Multi-output classification model types were identified as the appropriate model as there were three output columns and all the output columns have an output of 0 (at the risk of getting disease) or 1 (not at the risk of getting the disease). So, Support Vector Machine, Logistic Regression and Decision Tree models were identified as suitable models and Two models have been trained and evaluated in order to determine which model gives the best and most accurate predictions.

A Logistic Regression model was trained by using divided data and MultiOutputClassifier was used to get predictions. Logistic Regression is a supervised learning algorithm which is mostly popular for classification problems. It predicts dependent variables by analyzing the relationship between one or more independent variables. Additionally, MultiOutputClassifier is a strategy which is imported from scikit-learn to help the Logistic Regression model to predict multiple outputs simultaneously.

A Decision Tree model was also trained for the same dataset in order to compare with the above trained Logistic Regression model. Decision Tree is a supervised learning model that is used for classification and regression issues. But it is often used to solve classification problems.

All the trained models, as well as the accuracy levels attained by each model, was reviewed and compared in order to determine the most accurate model. So, the Decision Tree model was selected as the most accurate model as it showed an accuracy of about 84%. The accuracy of the Logistic Regression model was about 70%.

An application was created using the Flask framework to use the most accurate model chosen. Then, it is deployed on the Heroku cloud platform to be used in the mobile app to get the outputs by providing the inputs.

C. Total Nutrients Calculating API Development

A dataset was collected from Kaggle online community platform was used to create an API to calculate total protein, carbs, fat and fiber amount of an inserted grocery list. The complete dataset consists 336 food items records. It includes ten columns namely, “Food”, “Measure”, “Grams”, “Calories”, “Protein”, “Fat”, “Sat.Fat”, “Carbs” and “Category”. Only “Food”, “Measure”, “Grams”, “Protein”, “Fat” and “Carbs” were used to calculation purpose.

The API was developed using Node.js and Express.js. A file reading package called “FS” was used to read the dataset. Then, created API was deployed on the Heroku cloud platform to be used to get the outputs by providing the inputs.

D. Mobile Application Development

The mobile application design is done by creating prototypes using Figma tool. Then implementations are done using flutter as it is a cross platform framework.

The user's grocery list will be used to collect data and generate the output. The speech recognition option is one of the input methods to insert the grocery list. Speech recognition feature can be used by selecting any language from Sinhala and English. The system will then convert the speech to text. If the user has selected Sinhala as the insert option, the system should convert Sinhala text to English text and then the system will display the inserted grocery list to the user. The figure 2 describes the speech recognition feature.

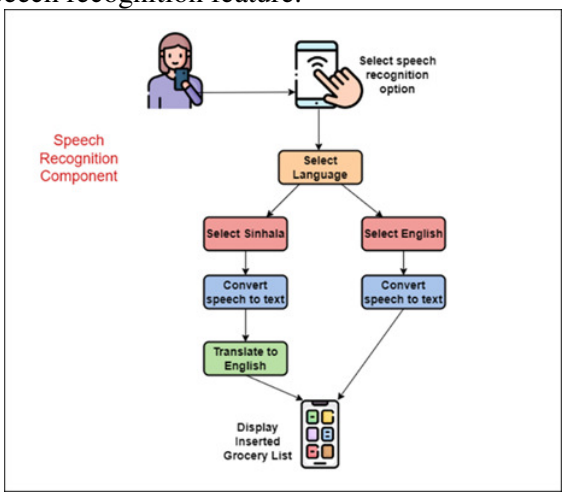

Figure 2 - System diagram for speech recognition input

When developing the speech recognition feature, flutter package called “speech_to_text” was used to get and identify the speech input. To convert Sinhala text to English a google API was used.

The text recognition option is another input method to insert the grocery list. Users can insert an image of the printed grocery list. Then, system will analyze the image and extract the relevant information from the image. Google's ML Kit Text Recognition was used to develop the text recognition feature.

The form input method is the final option to insert the grocery list. Users will be able to insert their grocery list by typing the name of the food item and the amount.

After the grocery list insertion, the system will ask to insert some additional details like number of days, number of users, ages of users and genders of users. Afterward, the system will send a request with the grocery list details to the created API and it will respond with the total nutrients amounts of the grocery list. Then those total nutrients details with users’ details will be sent to the API which was created using the trained machine learning model. It

will send the result with the predicted diseases. Finally, the predicted result will be displayed in the result user interface of the mobile application.

IV. RESULT AND DISCUSSION

To prove the accuracy and the methods suggested in research, thorough testing methods have been carried. The below section will provide descriptive information and results accordingly.

After training two machine learning models using Logistic Regression and Decision Tree algorithms to Predict future diseases using grocery list, those models were tested to check the accuracy using the testing dataset. The following table illustrate the test result.

Table 1 - Machine Learning Models' Accuracy

Algorithm	Accuracy
Logistic regression alorithm	0.7040
Decision tree algorithm	0.8368

According to the above table Decision tree algorithm was selected as the most accurate algorithm, as it showed a high accuracy than logistic regression algorithm.

The created API using node.js, express.js and a csv file which includes nutrients details of foods to get total nutrients amount of the grocery list was able to provide a most reliable and accurate result. It will be provided protein, carbs, fat and fiber amounts separately.

Some unit testing also done to test the mobile application and backend APIs.

There are some significant benefits with comparison to the existing mobile applications which are related to health in this industry.

- The proposed solution predicts future illnesses by considering grocery list item details
- The proposed solution provides the ability to input the grocery list using speech recognition option
- The proposed solution provides the ability to input the grocery list using Sinhala speech recognition option

As the future development plan, the model can be trained using more data in the future. Additionally, different machine learning algorithms can be tested to get more accuracy.

V. CONCLUSION

Several researches have revealed that people suffer from various diseases due to unhealthy eating habits. Diabetes is a main disease that people suffer a lot these days. Eating foods that include lots of glucose is a main factor to get diabetes. Some disadvantages such as kidney damage, eye damage, foot damage, nerve damage and skin conditions can be caused by diabetes. In addition, high blood cholesterol is another main disease that people suffer nowadays. Eating foods with lots of fats is a main reason to get cholesterol. Heart attack and stroke are main disadvantages of getting high cholesterol. High blood pressure is another main disease that people get due to unhealthy food consumption. It also causes lots of disadvantages in human life. Further, time wasting, money wasting and being unhealthy are disadvantages of getting those chronic diseases. When purchasing food people tend to use a grocery list. A grocery list is simply a list of groceries to be bought. Therefore, grocery lists are an important asset to get details of people's food consumption in day to day life. The proposed system will use people's grocery list and predict future diseases that they may get due to improper diets in the future. The proposed system will predict three diseases namely diabetes, high blood pressure and cholesterol. Hence, the proposed system helps in early awareness of diabetes, people may able to get prevent from those diseases by controlling their diets. It would be extremely beneficial if people could be aware about all these risks in advance. If someone knows that they are going to get a chronic disease in the near future, they can adopt preventive measures. As discussed above, bad eating habits can lead to chronic diseases like diabetes and cholesterol. Therefore, if there is a risk of such diseases, people can eat only healthy food. Additionally, if they need any consultant services to prevent disease, they can see a consultant before

they get infected. Today even small children get diseases like diabetes. Therefore, if parents are able to know their children's risk of disease early, they can limit their children's unhealthy food intake and direct their children to appropriate physical activities.

REFERENCES

[1] "World Health Organization," [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/healthy-diet.

[2] "Centers for Disease Control and Prevention," [Online]. Available: https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=With%20diabetes%2C%20your%20body%20doesn,vision%20loss%2C%20and%20kidney%20disease.

[3] "Cleveland Clinic," [Online]. Available: https://my.clevelandclinic.org/health/articles/11918-cholesterol-high-cholesterol-diseases#:~:text=High%20cholesterol%20is%20a%20condition,can't%20use%20them%20all..

[4] B. a. B. Unit, "Machine Learning Models to Predict Childhood and Adolescent Obesity," 2020.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE, 2017.

[6] M. S. Tamara Dubowitz, M. M. Deborah A. Cohen and M. Christina Y. Huang, "Using a Grocery List Is Associated With a Healthier Diet," 2015.

[7] N. Au, G. Marsden, D. Mortimer and P. Lorgelly, "The cost-effectiveness of shopping to a predetermined grocery list," 2013.

[8] P. G. P. P. E. M. Dhiraj Dahiwade, "Designing Disease Prediction Model Using," India, 2019.

[9] A. K. M. E. H. M. A. M. Shahadat Uddin1, "Comparing different supervised machine," 2019.

[10] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," IEEE, 2018.

[11] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," IEEE, 2019.

[12] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease".

[13] Y. Deepthi, K. P. K. Vyas, K. RadhikaD, K. Babu and N. V. K. Rao, "Disease Prediction Based on Symptoms Using Machine Learning," 2020.

[14] "Kaggle,"[Online].Available:https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey?resource=download&select=labs.csv

[15] "National Center for Health Statistics," [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013

[16] "National center for Health Statistics," [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DEMO_H.htm.

[17] "National center for Health Statistics," [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DR1IFF_H.htm.

[18] "National center for Health Statistics," [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DIQ_H.htm

