Exploration of Machine Learning to develop a Low-cost screening method with Global Diet Quality Score to detect Vitamin D deficiency

TUHINA PATRA

(20PFN027)

THESIS SUBMITTED TO

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATION FOR WOMEN, COIMBATORE – 641043

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR DEGREE OF MASTER OF SCIENCE IN FOOD SCIENCE AND NUTRITION

MAY 2022

Exploration of Machine Learning to develop a Low cost screening method with Global Diet Quality Score to detect Vitamin D deficiency

TUHINA PATRA

(20PFN027)

THESIS SUBMITTED TO

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATION FOR WOMEN, COIMBATORE – 641043

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR DEGREE OF MASTER OF SCIENCE IN FOOD SCIENCE AND NUTRITION

MAY 2022

25/5/22

Signature of the Supervisor

25/5/22

Signature of the Head of the Department

DECLARATION

I hereby declare that the dissertation entitled **"Exploration of Machine Learning to develop a Low cost screening method with Global Diet Quality Score to detect Vitamin D deficiency"**, submitted to the Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, in partial fulfillment of the requirement for the award of the **Degree of Master of Science in Food Science and Nutrition** is a record of original research work done by me under the supervision and guidance of Prof.(Mrs.).C.A. Kalpana, Professor, Department of Food Science and Nutrition, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore and that it has not formed the basis for the award of any Degree/Diploma/Associateship/Fellowship or similar title to any candidate of any other University and it represents entirely an independent work on the part of the candidate.

25/5/22

Signature of the Supervisor

Tuhina Patra
25/5/22

Signature of the Candidate

ACKNOWLEDGEMENT

First and foremost the investigator places her humble salutations at the feet of God **Almighty** who has given sound wisdom, knowledge, strength and opportunity to do the investigation effectively.

The investigator owes her respectful gratitude to **Late Dr. (Thiru) T.S. Avinashilingam**, Founder and First Chancellor of Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing this temple of learning to do this research.

The investigator records her reverential gratitude to **Late Hon. Colonel Dr. (Tmt.) Rajammal P. Devadas, M.A., M.Sc., Ph.D(Ohio State), Hon. D.Sc.(Madras, Hon D.H.L.(Oregon State), Hon. D.H.L.(Ohio State), D.S.C.(Kanpur), Hon.D.Sc.(Northern Ireland)**, Former Chancellor, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for being a perennial source of inspiration for conducting the study.

The investigator is grateful to **Prof. S.P.Thyagarajan, Ph.D, M.D, D.Sc, FAMS, FNASc, FIMSA, FABMS, FFTM (Glasgow, UK)**, **Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing an opportunity to conduct the study.

The investigator offers her deepest gratitude to **Dr. T. S. K. Meenakshisundaram. M.A., M.Phil., Ph.D., Managing Trustee,** Sri-Avinashilingam Education Trust Institutions, Coimbatore for providing the opportunity to conduct the study.

The investigator owes her special thanks to **Dr. V. Bharathi Harishankar Ph.D., FRSA**, **Vice Chancellor**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for facilitating to complete the study.

The investigator expresses her thanks and gratitude to **Prof. (Mrs). S. Kowsalya, M.Sc., M.Phil., Ph.D., Registrar**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all help for the smooth conduct of the study.

She expresses her heartfelt thanks to **Dr. (Mrs.) N. Vasugi, M.Sc., M.B.A., M.Phil., Ph.D.,** Dean, School of Home Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her help to carry out the research.

It gives the researcher an immense pleasure and proudness to offer profound gratitude to **Prof. (Mrs.) M. Sylvia Subapriya, M.Sc, M.Phil., B.Ed., Ph.D**, Professor and Head of the Department of Food Science and Nutrition, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her suggestions and constant encouragement throughout the research study to make this project fruitful and successful learning experience.

It is investigator's pleasure and privilege to express her deep sense of gratitude to professor and her guide **Prof. (Mrs). C.A. Kalpana, M.Sc, B.Ed, M.Phil, Ph.D., SLET., NET.,** Associate Professor, Department of Food Science and Nutrition, Avinashilingam, Institute for Home Science and Higher Education for Women, Coimbatore, for her dynamic, excellent and awesome guidance , by which she has been able to execute her research and complete it successfully.

Investigator expresses her humble gratitude to **Ms. Rama Mercy S., Temporary Teaching and Research Fellow**, Department of Computer Science, Campus 2, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her help in the Machine Learning aspects of the study with her advanced experience for proceeding this study.

Out of deep sense of indebtedness, the investigator expresses her sincere thanks to all the **Staff Members** of the Department of Food Science and Nutrition, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore who helped her with their constructive suggestions at all times.

The investigator also wishes to express her deep indebtedness and gratefulness to her Parents**, Sisters, Friends and All Well Wishers** for their patience, motivation and constant support throughout the course of the study.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

INTRODUCTION

I. INTRODUCTION

Vitamin D is one of the fat soluble vitamins which are also known as sunshine vitamin. It has many pivotal roles in the human beings. Vitamin D is found in two forms in the nature: (i) Vitamin D_3 (cholecalciferol) – animal origin (ii) Vitamin D_2 (ergocalciferol) - plant origin. Until 1992 it was unknown that precursor of Vitamin D is 7-dehydro-cholesterol in animals and ergosterol in plants. Measurements of vitamin D are made with their active metabolites of Vitamin D which includes 25-hydroxy vitamin D, 24, 25-dehydroxy vitamin D and 1,25 dihydroxy vitamin D.

Synthesis of cholecalciferol depends on sun exposure that takes place in the lower layers of skin epidermis through a chemical reaction. Both cholecalciferol and ergocalciferol can also be taken from the diet or supplement. Foods which contain vitamin D are certain cereals, cow' milk, fortified milk or fortified breakfast cereals, mushroom exposed to UV light.

However, vitamin D get absorbed either from diet or it gets synthesized under the skin, but both are present in their inactive form. It gets activated in two steps which involve two protein enzyme hydroxylation steps in the liver and kidney respectively.

In humans, serum 25 hydroxyvitamin D or [25(OH) D] considered best biomarker for estimating Vitamin D level as it is the most reliable indicator of storage in the human body. Vitamin D deficiency (VDD) can be defined as having a serum [25(OH) D] level <20 ng/mL (50 nmol/L), and severe deficiency is defined as [25(OH) D] <10 ng/mL (25 nmol/L) (Holick, 2009).

Endocrine Society Task Force under the clinical practice guildelines defined deficiency as a cutoff level of 50 nmol/L [Holick *et al*, 2011]. The Institute of Medicine (IOM) considers the cut-off values of serum vitamin D as 25(OH)D < 12 ng/mL (30 nmol/L) indicates the risk of vitamin D deficiency along with risk of metabolic bone disease (Cashman, 2018). The Endocrine Society and Scientific Societies as European Calcified Tissues Society (ECTS), International Osteoporosis Foundation (IOF) or American Geriatrics Society (AGS) established the threshold of deficiency at ≤ 20 ng/mL (50 nmol/L), based on the increasing levels of circulating parathormone and designate levels ≤ 10 as severe deficiency. According to the IOM, 25(OH)D serum levels of 12 to 20 ng/mL corresponds to a risk of insufficiency

of Vitamin D, and levels ≥20 is taken as sufficiency. ECTS has different criteria for this. Levels ≥30 ng/mL considered normal and levels of 21 to 29 ng/mL as insufficient (Duarte *et al.*, 2020).

Vitamin D deficiency is the most common and unnoticed nutrient deficiency prevalent all over the world. People with VDD are unaware that they are deficient of Vitamin D due to its overlapping symptoms such as weakness, joint pain and fatigue with other nutrient deficiencies. Hence, people ignore it and consider it mild. This may be the reason for the wide prevalence of vitamin D deficiency. Vitamin D insufficiency causes the ineffective utilisation of calcium which is essential for bone mineralisation. Along with that, phosphorus has crucial role in bone mineralisation. So it can be said that vitamin D deficiency is associated with disorders of calcium and phosphorus metabolism, which in turn causes defective chondrocyte differentiation, mineralization in growth plate and defects in osteoid mineralization among children. Defect in bone metabolism causes rickets, osteomalacia, osteoporosis, and others bone density disorders. Deficiency of vitamin D also plays a massive role in other physiological disorders viz. diabetes mellitus, immune disorders, certain cancers, dementia, heart diseases, high blood pressure or preeclamsia, covid-19, obesity, dental caries.

Insufficiency of vitamin D or its deficiency is trending as one of the global health issues that affect more than one billion population of adults as well as children worldwide. It is now recognised as a major public health concern and pandemic.

Prevalence of Vitamin D deficiency is measured by using various kinds of screening methods in different period of time in different countries. Prevalence rates of vitamin D deficiency is defined as 25(OH)D <30 nmol/L or (12 ng/L), of 5.9 percent in US, 7.4 percent Canada and 13 percent Europe (Amrein *et.al*., 2020). The prevalence of 25(OH)D levels <50nmol/L (or 20ng/L) is found 24 percent US, 37 percent Canada and 40 percent Europe. Countries such as India, Pakistan, Tunisia and Afghanistan shown the high Vitamin D deficiency prevalence i.e., 25(OH)D levels or <30nmol/L in more than 20 percent of the population. In India, around 450 million people are deficient of vitamin D (Cashman *et.al*., 2019). Socio-demographic, lifestyle, and clinical factors were found to associate with vitamin D deficiency in these different studies (Duarte, 2020).

Specific categories of patient are highly prevalent of Vitamin D deficiency. They are generally characterised by insufficient level of serum Vitamin D or failure of the organs involved in Vitamin D

metabolism (Amrein *et.al*., 2020). 85 to 99 percent patients on hemodialysis or with chronic renal failure, affected liver in renal transplant recipient or liver transplantation are deficient of vitamin D (Courbebaisse *et.al.*, 2009). A study in Germany was conducted by Micheal F. Holick showed that 25 percent to 35 percent are deficient of vitamin D along with osteomalcia and osteoidosis respectively. A latest observational data have suggested that about 40 percent of Europeans are vitamin D deficient, and 13 percent are severely deficient (Cashman *et.al*.,2019).

Vitamin D estimating methods are very expensive methods. Its active metabolites, especially 25-(OH)D, are useful while estimating serum Vitamin D level in the body. There are various screening methods which are used for estimating Vitamin D are:

- **25(OH)D Assay** – Vitamin D binding protein (DBP), Radioactive immunoassay (RIA), High performance liquid chromatography (HPLC) ; and advanced liquid chromatography tandem with mass spectroscopy (LC-MS) (Holick, 2009).
- **1,25-dihydroxy Vitamin D assay** – Bovine VDR binding protein to estimate 1,25-(OH)D, Radioimmunoassay, Diasorin assay, IDS assay (Holick, 2009).

Among all these active metabolites, only 25(OH)D gives the right determination for levels of vitamin D in the serum. It measures the sufficiency and deficiency of vitamin D in the circulating blood. Poor quality diets are the reason why both undernutrition and overnutrition are prevalent in the world. It may link to adverse health effects and may directly or indirectly cause any kind of diseases and disorders. In 2018, Intake – Center for Dietary Assessment launched a 2-year research initiative to support a consortium of researchers at the Harvard T.H. Chan School of Public Health Department of Nutrition and the National Public Health Institute (INSP), Mexico, to develop and validate metrics of diet quality that would be appropriate for collecting data through routine population-based surveys and that would be fit for purpose for inclusion in global monitoring frameworks (Harvard T.H Chan, School of Public Health, Mexico, 2021).

GDQS is a food group-based metrics that accounts the quantity of consumption in a distinctive scoring method. An overall metric of diet quality from these analyses is identified i.e., the Global Diet Quality Score (GDQS). The GDQS can be formulated by using both primary and secondary data.

GDQS comprises of 25 food groups: (i) 16 healthy foods (ii) 7 unhealthy food groups (iii) 2 food groups (high-fat dairy, red meat) that are hazardous when consumed in extremely high amounts

(Harvard T.H Chan School-Intake). Two different types of quantity of consumption (grams/day) of scoring methods are used for these food groups. 24 of the GDQS food groups use scoring in three ranges of quantity of consumption: low, medium, and high. While the one food group (high-fat dairy) uses four ranges of quantity of consumption: low, medium, high, and very high. It is used because its quality and quantity affects the health adversely (Harvard T.H Chan, School of Public Health, Mexico, 2021).

During the 24-hour reference period each participant receives points for each GDQS food group depending upon the amount of food consumption from each food group. Certain scores have been assigned to each food groups and scoring for healthy and unhealthy food groups is different. Its food groups range exists between 0 to 49 (Harvard T.H Chan, School of Public Health, Mexico, 2021). It is the outline of overall diet quality at population which is taken in reference to nutrient adequacy and diet-related non-communicable disease risk. GDQS scores signify: (i) ≥23 - low risk (ii) scores ≥15 and <23 indicate - moderate risk (iii) <15 - high risk (Harvard T.H Chan, School of Public Health, Mexico, 2021).

GDQS data is collected for tabulation of two GDQS sub-metrics: the GDQS positive sub-metric (GDQS+) and the GDQS negative sub-metric (GDQS–).Total score range of the 16 healthy GDQS food groups of GDQS+ is 0 to 32. Total score range of the 7 unhealthy GDQS food groups and the 2 GDQS food groups that are unhealthy when consumed in excessive amounts are designated in the GDQS− is 0 to 17 (Harvard T.H Chan, School of Public Health, Mexico, 2021). The GDQS+ and GDQS− provide appropriate information about the contribution made by both healthy and unhealthy food group consumption to wholesome diet quality in a specific setting. The GDQS is sensitive both undernutrition and overnutrition.

An American Computer Scientist named John McCarthy coined the term “artificial intelligence” in 1955 during a research project proposal (MaCarthy, 2006). Artificial intelligence (AI) is a branch of computer science known to imitate certain actions of humans such as learning abilities, thoughts and way of knowledge using (Sak and Suchodolska, 2021). In much experimental and clinical medicine it has shown many of its application. In recent years artificial intelligence proved much of its effectiveness in the field of medical diagnosis and detecting risk.

AI is the broader field of computer science while Machine learning (ML) is a sub category of AI which comprises algorithms to perform certain functions in a specific way. Computers are known to do complex tasks and it gave huge success in the world. Machine learning is the scientific discipline which

focuses how computers learn from raw data and process to give result. It is the bridge between statistics and computing algorithms. It made this possible to use millions of millions of data points by building statistical models from massive data sets to derive to certain conclusion. In machine learning, performance of each individual participant is judged on the basis of their performance from common data sets and recurrent supervised learning problems. The computer is programmed to approximate human performance. The machine learning can be sub-classified into supervised learning and unsupervised learning. Machine learning (ML) algorithms are widely used in studies on the influence of nutrients on the functioning of the human body in health and disease (Deo, 2015).

The working of objective of machine learning lies in learning subjects from the input data which in general refers to training data and make future predictions with the new available data.

ML algorithms are responsible to make mathematical models for decision making. These models are designed by using a large amount of data which can be collected from the random population. Machine learning and statistical methods co-ordinately discover patterns present in data. It may use many expensive imaging technology or laboratory measurements. Some low-cost scoring methods using a combination of ML techniques and questionnaire data can be used in effective screening for Vitamin D deficiency.

Sun exposure is a strong risk factor for estimating the vitamin D level, examining the predictive performance of low-cost dietary assessment tools with ML techniques. Adults are most susceptible to Vitamin D deficiency especially in country like India. They expose to sunlight but most of them cover themselves by wearing long sleeves to protect from UV-B sun rays. UV-B sunrays are mostly available during mid-day time i.e., from 10 am to 4 pm. This UV-B has the only ability to synthesise Vitamin D under the epidermis. The Global Diet Quality Score (GDQS) used for measuring diet to interpret health status and vitamin D level in an individual and has been tested in a number of LMIC settings. GDQS alone can be used to assess risk of several chronic diseases but for detecting the vitamin D level we need serum vitamin D level from a group of population to form algorithm for ML.

Various methods for machine learning are Gaussian Naïve Bayes, Linear Regression, k-nearest neighbor, Random Forest classifier, Boosting Classifier, Decision Tree, AdaBoost Classifier, Support Vector Machine, Stochastic Gradient Classifier, Multilayer Perceptron, Linear Discriminant Analysis, and Gradient Boosting classifier. An appropriate method for detecting VDD is selected for getting the correct result with most accuracy (Sambashivam *et al.*, 2020).

Objectives of the study:

To:

i. Collect serum Vitamin D databases
ii. Conduct online survey using Global Diet Quality Score (GDQS) method
iii. Assess vitamin D rich food intakes
iv. Select appropriate technique for Machine Learning and apply algorithm
v. Explore in machine learning to detect Vitamin D deficiency with real time dataset

REVIEW OF LITERATURE

II. REVIEW OF LITERATURE

The review of literature pertaining to the study entitled "**Exploration of Machine Learning to develop a Low cost screening method with Global Diet Quality Score to detect Vitamin D deficiency**" is presented under the following heads:

(i) Prevalence of Vitamin D deficiency

(ii) Screening of Vitamin D

(iii) Global diet quality score (GDQS)

(iv) Machine Learning

1. PREVALENCE OF VITAMIN D DEFICIENCY

Vitamin D deficiency is prevalent worldwide. It was earlier thought that vitamin D deficiency causes only bone related disorder viz. osteoporosis, rickets. But later it was mentioned in "The vitamin D deficiency pandemic: approaches for diagnosis, treatment and prevention" by Dr. Holik MF that its deficiency is due to myriad of many diseases such as s sarcopenia, multiple sclerosis, cardiovascular disease, diabetes, infections, auto-immune diseases, and cancer, besides the former (Holick, 2003). In certain cases it has been found that vitamin D deficiency and mortality are correlated. It is especially applicable with specific causes such as cardiovascular and certain kind of cancer (Chowdhary *et al.*, 2017).

For detecting the concentration of vitamin D, the serum 25(OH) D is used to assess vitamin D status because it reflects the contribution from both diet and dermal synthesis. About 90 percent of the required Vitamin D is synthesized in the skin under sun exposure (Holick, 2003). Several studies have been conducted across the world to assess the deficiency in different parts of the world. Even though it is worldwide deficiency, still it remains untreated and undiagnosed. From various studies it has been found that Vitamin D is not directly related to age, sex and geography. Different organization has different cut off values for vitamin D deficiency, so there is no such As there is no standard guideline which can be followed all over the world to classifying the Vitamin D status. However, majority of these studies mentioned serum 25(OH) D level of <20 ng/ml as cut off value for vitamin d deficiency. Vitamin D deficiency related to nutritional rickets and osteomalacia, the majority of expert bodies have suggested that serum 25(OH)D concentrations <25 nmol or <30 nmol/L indicates increased risk (Institute of Medicine, 2011).

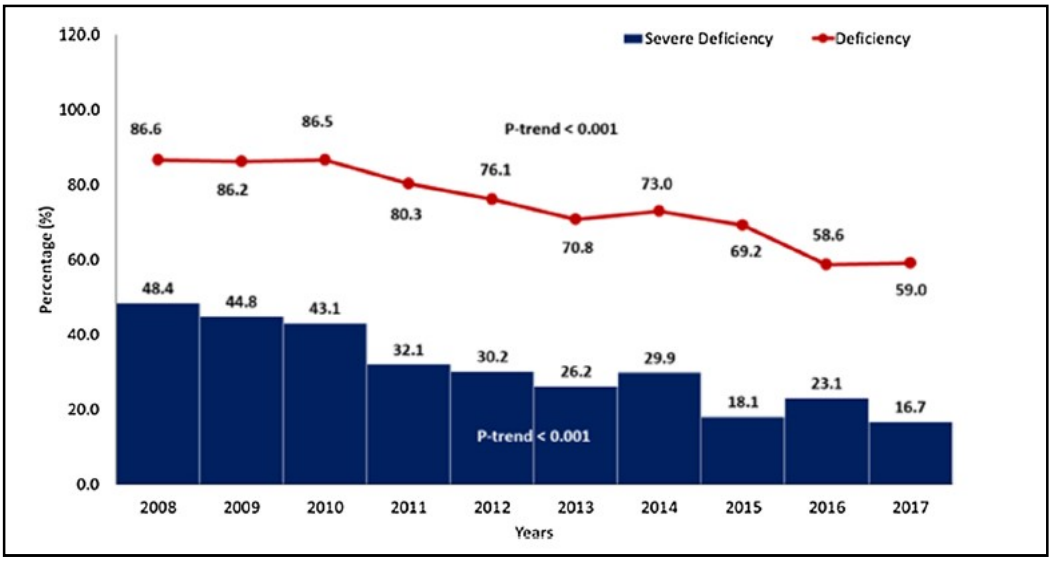
a. PORTUGAL

A study was conducted among 5459 people in Portugal population. Among which 60 percent showed their 25-(OH)D level less than 20ng/ml (Santosh *et al*., 2015). According to a retrospective study conducted by Duarte et al. among Portuguese population, it was found that 66 percent adults in Portugal is either having vitamin D deficiency or insufficiency. They considered various parameters for this study: (i) Demographic and socioeconomic profiles – gender, age (categorized into eight age groups ≥ 18–29, 30–39, 40–49, 50–59, 60–69, 70–74, and ≥ 75 years old), ethnicity, educational level, and household income of the last month; (ii) Seasonality (iii) Lifestyle- it includes alcohol consumption, smoking practice and physical exercise and (iv) Anthropometric datas are height, weight, BMI. Based on EpiReumaPt the estimated national prevalence recorded in Portugal was 25(OH)D ≤ 10 ng/mL and < 20 ng/mL is 21.2 percent and 45.4 percent respectively (Duarte *et al.*, 2020). According to IOM only 33.4 percent of the adult population present normal values (≥ 20 ng/mL). While Endocrine society reported only 3.6 percent present normal values (≥ 30 ng/mL). These data are recorded among adult Portuguese population which raised many publich health concerns (Duarte *et al.*, 2020).

b. SAUDI ARABIA

Saudi Arabia is one of the Western Asian countries where vitamin D deficiency is recognized as a serious health issues. It was found that vitamin deficiency is quite common there in all type of age group – men, women, pregnant, lactating, toddler, children, adolescent (Al-Dahgiri *et al.*, 2021). Al-Dahgri *et al*. conducted a meta-analysis of 13 studies published from 2011 to 2016, which revealed a pooled prevalence of vitamin D deficiency (< 50 nmol/l) as 81.0 percent. To prevent vitamin D deficiency in Saudi Arabia, various attempts has been taken to create awareness among public. Parameters taken for estimating Vitamin D deficiency are as follows: (i) anthropometric factors: height, weight, waist to hip ratio; (ii) Diastolic pressure and systolic pressure. In the study by Al-Alyani found that overall mean 25(OH) D was significantly higher in males (41.6 ± 21.6) than females (39.5 ± 25.6, $p < 0.001$). In the case of both sexes, participants aged over 40 years had higher mean 25(OH) D concentration than younger participants. The overall vitamin D deficiency in Saudi Arabia was 73.2 percent (Al-Alyani *et al*., 2018).

Al-Mogbel documented that females are prone to vitamin D deficiency because of the absence of outdoor activities, however both male and female participants in this study had similar vitamin D deficiency affecting 72.8 percent and 73.5 percent of the subjects, respectively (Al Mogbel, 2012).



(Source : Al . Al-Daghri *et al.*)

Fig 1: Vitamin D trend in Saudi Arabia from 2008 to 2017

c. EAST ASIAN COUNTRIES

China is a large country with large geographic variation. Various studies have been conducted in the certain part of the country. A study conducted by Strand MA in 2009 in Yuci in China among 18 to 23 years of adults showed 33 percent deficiency of vitamin D. While the insufficiency was found 65.3 percent among 18 years old adults and 2.9 percent among 23 years adults (Strand, 2009).

In another study conducted by Yu S et al. in China where 2173 individuals were included (male:female=0.98:1). The mean 25(OH)D level was recorded as 19.4±6.4 ng/ mL, with 109 (5.0 percent) participants having 25(OH)D2 level >2.5 ng/ mL. Of these 109 participants, 98 (90.0 percent) had a level of 2.5 to 10 ng/mL (22.4 ng/mL). The distribution of participants with detectable levels of $25(OH)D_2$ was found in the following pattern: 19 (4.6 percent) in Beijing, 23 (5.3 percent) in Hangzhou, 54 (12.3 percent) in Guangzhou, 5 (1.0 percent) in Dalian, and 8 (2.0 percent) in Urumqi (Yu *et al.*, 2015).

A study among Japanese population, age from 20-69 years was conducted in which 107 active participants found. A survey was done in Shakotan Town and Kumamoto City both during summer and winter to find how the sunlight exposure, diet influence serum Vitamin D, 25(OH)D. The participants who gave consent for this study, a UV measurement device, two questionnaires, and an activity diary were distributed to them. They wore the device for 10–14 days, and kept the activity diary and focused

on factors which are related to UV exposure viz., clothing, sunscreen use, duration of outdoor activity. In the given time period they also completed the dietary assessment and lifestyle questionnaires (Asakura *et al.*, 2020).

An observational study of seven years was conducted in South Korea by Korea National Health and Nutrition Examination Surveys (KNHANES) from 2008 to 2014. A total of 39,759 patients were included for the analyses. Radioimmunoassay was used to measure serum 25-hydroxyvitamin D (25 (OH)D levels. The overall mean serum level of 25 (OH)D was 45.7 nmol/L in males and 40.9 nmol/L in females was recorded (Park *et al*., 2018).

d. UNITED KINGDOM

United Kingdom is made up of mixed population, mostly South Asian (Indian, Bangladeshii and Pakistani) and Black American-Caribbean. A cross-sectional sub-study of the Ethnic-Echocardiographic Heart of England Screening study (E-ECHOES) was conducted among South Asian (SA) and Black African-Caribbean (AC). There are 5408 active participants for the study. The data for serum vitamin D were available for 1904 participants. Overall, many more SA men and women had severe vitamin D deficiency at 42.2 percent (39.2–45.1) than ACs at 12.5 percent (10.2–14.9). Adequate levels of vitamin D were seen in 8.5 percent (6.8–10.1) of SAs and in 15.4 percent (12.8–17.9) of ACs. More female ACs had adequate vitamin D compared to men. On age and gender matched comparison of ethnic groups with severe deficient or deficient vitamin D, vitamin D was lowest in South Asians, who also had lower lipids (Patel *et al*., 2012).

e. INDIA

In India the deficiency of Vitamin D was detected in all the parts of the country and among all the age groups. The scenario of this is similar in both rural and urban. Fish is the staple diet in Bengal (eastern India) and fish is known to be rich in vitamin D. Still Vitamin D nutrition status of eastern population is not comparable to other states (Ritu and Gupta, 2014). In some countrywide studies it has reported vitamin D deficiency in as high as 70 to 100 percent of ostensibly healthy individuals (Ritu and Gupta, 2014).

Kapil *et al.* conducted a study among children (6-18 years), found that 93 percent of them are vitamin D deficient (Kapil *et al.*, 2017). Kumar *et al.* conducted a study on 106 labour mothers in Bangaluru, suggesting 70.7 percent deficiency among them and 83 percent deficiency in cord blood of

new born (Kumar *e.al.*, 2015). A study conducted in Kolkata around 300 participants of the age group of 1-16 years to find the deficiency of Vitamin D; the result was noted 52.9 percent deficiency (Basu *et al*., 2015).

In India to understand the Vitamin D level in the children, many school based studies has been done. In 2011 a study on 214 premenarchal girls in Pune was conducted by Kadam *et al*. It showed a prevalence of 34.2 percent of Vitamin D (Kadam *et al.*, 2011). Kapil *et al*. back to 2017 conducted another school based study on 1222 school children aged 6–18 years in Kangra and Kullu districts of Himachal Pradesh. It reported the prevalence of 81 percent and 80 percent respectively (Kapil *et al.*, 2017).

In 2010 a study conducted by Borkar *et al.* among the North Indian population recorded the mean levels of vitamin D were lower in the cases (20.02 ± 10.63 ng/mL, 50.05 ± 26.57 mmol/L) as compared to controls (26.16 ± 12.28 ng/mL, 65.4 ± 30.7 mmol/L). In the study there were twenty-nine children which constituted 58 percent of the study group were vitamin D deficient while 16 children from control group (32 percent) had deficiency. Fourteen children in the study group (28 percent) and 22 children in the control group (44 percent) had insufficient vitamin D levels. Only seven (14 percent) in the study group and 12 (24 percent) in the control group had sufficient vitamin D concentrations. Majority of subjects (81 percent), both in study (86 percent) and in control (76 percent) were either vitamin D deficient or insufficient (Borkar *et al*., 2010). There are various reasons why Indians are deficient of Vitamin D:

- Due to modernization in the urban population, people prefer to indulge in indoor activities, which are reducing their exposure to the sunlight.
- Increase in industrialization causes pollution which can hamper the synthesis of Vitamin D in the skin by UV rays (Babu *et al.,* 2010)
- Lifestyle modification led to changes in food habits that in turn led to lower intake of dietary calcium and Vitamin D
- Fiber rich diet contains phytates and phosphates that can deplete Vitamin D stores and increase calcium requirement (Harinarayana *et al.,* 2007)
- Increased application of sunscreens
- Wearing full or long sleeve shirts or tops
- Un-spaced and unplanned pregnancies in women with dietary deficit can lead to worsening of Vitamin D status in both mother and child.

2. SCREENING OF SERUM VITAMIN D DEFICIENCY

A. OVERVIEW

Vitamin D, as mentioned, found in two forms – Vitamin D_2 and Vitamin D_3. Vitamin D cannot directly function in the body. It functions via its three major active metabolites - 25-hydroxy vitamin D [25(OH)D], 24, 25-dehydroxy vitamin D and 1,25 dihydroxy vitamin D [1,25 $(OH)_2D_3$]. The levels of this specific active metabolite are considered while assaying Vitamin D serum.

Before assessing the vitamin D in the serum, it is essential to know the cut off value to determine if the individual is deficient of vitamin D or having sufficiency of vitamin D. Various international and national organizations such as IOM, Endocrine society has been work upon to find the exact cut off values. The cut off value for studies of Indian population with a desirable and safe range of serum 25(OH)D levels has been set at 30–100 ng/mL. Vitamin D deficiency is defined when the serum levels of 25(OH)D is < 20 ng/mL (50 nmol/L) with consequent and consistent elevation of PTH and reduction in intestinal calcium absorption. Vitamin D insufficiency is defined as serum 25(OH)D levels in the range of 20–29 ng/mL. Vitamin D sufficiency is defined as serum levels of 25(OH)D 30–32 ng/mL (Ritu and Gupta, 2014). These ranges would be sufficient for most known effects of vitamin D and helps to prevent deficiency and toxicity.

In vitamin D there is requirement of hydroxylation on its carbon 25 to produce 25-hydroxyvitamin D [25(OH)D] which led to develop a binding protein assay using the vitamin D binding protein (DBP) to measure circulating levels of 25(OH)D in the circulation (Holick, 2009).

The prevalence of vitamin D deficiency is high enough to achieve multiple beneficial outcomes with screening population and supplementation to those who are at risk.

In the serum both Vitamin D_2 and Vitamin D_3 are present. Together they are represented as 25(OH)D, is an useful indicator for screening. Serum concentration of 25 hydroxyvitamin D [25(OH)D] is the optimal clinical indicator of vitamin D metabolism. Moderate to severe vitamin D deficiency (25(OH)D<25nmol/ml) is casually associated with osteomalacia and rickets in children. Milder degrees of vitamin D deficiency (25(OH)D<50nmol/ml). Controversy exists regarding the 25(OH)D target of therapy for musculoskeletal benefit, when to measure 25(OH)D and the performance of current methods used to measure 25(OH)D (Cashman, 2020).

B. METHODS USED FOR SCREENING SERUM VITAMIN D

- **25(OH)D Assay** – Vitamin D binding protein (DBP), Radioactive immunoassay (RIA), High performance liquid chromatography (HPLC) ; and advanced liquid chromatography tandem with mass spectroscopy (LC-MS) (Holick M.F., 2009).
- **1,25-dihydroxy Vitamin D assay** – Bovine VDR binding protein to estimate 1,25-(OH)D, Radioimmunoassay, IDS assay (Holick, 2009).

The active metabolites for measuring vitamin D serum level are 25(OH)D and 1, 25-dihydroxy Vitamin D. Two characteristics of 25(OH)D makes it an ideal marker for monitoring short and long changes in vitamin D status are significantly higher in serum and longer shelf life as compared to other vitamin D metabolites (Farrell *et al.*, 2012). Advantages always come with disadvantages. Lack of standardization of 25(OH)D assays leads to errors in assessment of vitamin D status. Some authors mentioned that current vitamin D analysis methods often lack sensitivity and specificity. The analytical uncertainties in measuring serum 25(OH)D was recently emphasized at the NIH workshop ‘‘Nutrient Biomarkers Analytical Methodology: Vitamin D Workshop’’ held on December 16, 2009 (Adamec *et al.*, 2010).

Some of the most popular and common methods are discussed below:

1. **Vitamin D binding protein (DBP)**- In 1970s, initially the vitamin D assays based on competitive binding principle (Jafri *et al.*, 2011). Bioavailable 25-hydroxyvitamin D was defined as circulating 25-hydroxyvitamin D not bound to vitamin D–binding protein, which is analogous to the definition of bioavailable testosterone (Powe *et al.*, 2013). Intake of vitamin D in the human body occurs in the form of vitamin D2 (plants) and Vitamin D3 (animals). It binds to vitamin D binding protein (DBP) in plasma and is transported to the liver where, both are hydroxylated to form 25OHD (Jafri *et al.*, 2011). Even though approximately 85 to 90 percent of 25(OH)D is bound to D binding protein (Powe *et al.*, 2013), measurement of serum vitamin D binding protein (DBP) complicates the analysis of vitamin D status. Vitamin D binding protein (DBP) is important for the bioavailability of active 1,25-dihydroxyvitamin D [$1,25(OH)_2D$] and its precursor 25-hydroxyvitamin D [25(OH)D], however it is difficult to distinguish between DBP-bound and free 25(OH)D and $1,25(OH)_2D$ (Chun *et al.*, 2010). The major disadvantage of these assays is long incubation period and tiresome extraction procedures (Jafri *et al.*, 2011).

There were three forms of DBP called polymorphic forms which produces six allelic combinations (Chun *et al.*, 2012). These allelic forms of DBP circulate in serum at different concentrations. Because of the various genetic polymorphisim, and exhibit their binding affinities for 25(OH)D and 1,25$(OH)_2$D (Powe *et al.*, 2013). Both of these variables have the potential to influence the bioavailability of vitamin D. DBP as a determinant of free vitamin D and vitamin D function (Chun *et al.*, 2012). The levels of vitamin D–binding protein were determined using a commercial enzyme-linked immunosorbent assay (R&D Systems) that employs a sandwich of two monoclonal antibodies (Powe *et al.*, 2013).

2. **Radio Immuno Assay (RIA):** In 1985 Food and Drug Administration made available RIA for the first time for clinical use (Hollis, 2000). Many novel RIA and enzyme immunoassays have recently been created, with ease in automation. Despite some advantages, there remain some shortcomings of immunoassay include: (i) variability between immunoassay batches (ii) deviation of analyte concentrations over a linear calibration range. Immunoassay can be performed in various methods which unable to differentiate the two major form of vitamin D. It may be influenced by the variability in Vitamin D binding protein, thereby reduces selectivity (Jafri *et al.*, 2011).

3. **High Performance Liquid Chromatography (HPLC):** In 1978, HPLC was developed over Vitamin D binding protein. HPLC is considered as gold standard for vitamin D assessing as it can quantify separately vitamin D_2 and vitamin D_3 (Jafri *et al.*, 2011). One of the most recent and developed screening method to quantify vitamin D serum level is HPLC. There are two ways in which quantitative HPLC assays are being done: ultraviolet detection HPLC and phase separation HPLC. Phase separation is of three kinds: (i) normal phase separation (ii) combined use of normal- and reversed-phase separations (iii) reversed-phase separation. Reversed-phase HPLC methods for 25$(OH)D_3$ in human plasma have been developed recently with normal-phase prepurification of the sample or liquid extraction only. In the earlier times HPLC methods was designed for measuring 25$(OH)D_3$ in serum especially for research purposes and for the routine use it was difficult to perform. Then it is modified in easy way which is easy to use, sensitive and rapid with simple sample preparation. Isocratic elution Separation and quantification of 25$(OH)D_3$ from 25$(OH)D_2$ are achieved with an isocratic elution (Turpeinen *et al.*, 2003).

Some problems been encountered while using the normal HPLC. So reverse HPLC has taken that place. It was found that HPLC tandem with UV is the most convenient way to make HPLC to work, but

analysing vitamin D metabolites in biological samples such as blood, serum posses greater risk to keep an ideal sample clean to keep it away from the interfering compounds (Hymøller and Jensen, 2011). Sample cleanup is done by three processes: (i) solid phase extraction (SPE) (ii) preparative HPLC (iii) liquid–liquid extraction with or without alkaline saponification. Lipid content of the sample matrix influences the process of saponification. Soaps formed during the saponification step increase the risk of soap micelles formation, which reduces the polarity of the aqueous phase, making it difficult to extract even slightly polar compounds into the organic phase, which can compromise the extraction of hydroxylated vitamin D metabolites into the organic phase in the case of vitamin D. The most common column of HPLC for vitamin D analysis is Reversed phase C_{18} hydrocarbon (Hymøller and Jensen, 2011). Even though it has advantages and used in recent times, it has several disadvantages include tedious technique, high budget, large sample volume and specialist knowledge to perform this type of analysis (Jafri *et al.*, 2011). For complex biological matrices, HPLC lacks the appropriate level of specificity due to spectrum interferences. (Adamec *et al.*, 2010). In 2011, a comparative study was conducted between RIA and HPLC for assaying serum vitamin D. It has been found the median serum 25OHD was 51.1 nmol/L (IQ= 12.5–187.2 nmol/L) and 50.1 nmol/L (IQ= 17.7–199.4 nmol/L) using RIA and HPLC, respectively (Jafri *et al.*, 2011).

4. **LC-MS**: LC–MS/MS allows more accurate way to quantify serum 25-hydroxyvitamin D [25(OH)D] for indicating vitamin D status (Satoh *et a.l*, 2016). Biologically active vitamin D metabolite is 1α, $25(OH)_2D_3$ that is synthesised by hydroxylation of 25(OH)D in the kidney. The 1α, $25(OH)_2D_3$ metabolite is present in serum at low concentration ranges with half life time of 4 hours. These low concentration ranges led to the development of less sensitive older generation LC-M/MS instruments. LC-MS is superior to RIA as it can distinguish between analyte through chromatographic separation and differences in mass transitions. LC-MS/MS is the most efficient method for quantification of multiple vitamin D analytes in the clinical setting for measuring vitamin D status. During sample preparation for LC-MS analysis it is essential to avoid matrix effects and to concentrate the sample. There are various methods for preparing samples for HPLC and choosing the ideal open which is inexpensive and rapid method is crucial (Jenkinson *et al.*, 2016).

One method is Liquid-liquid (LLE) and supported liquid-liquid extraction (SLE). Liquid-liquid extraction (LLE) is a routine and inexpensive method. However, because of its time taking method for larger samples it is not appropriate (Jenkinson *et al.*, 2014). While supported liquid-liquid extraction

(SLE) used to reduce sample preparation time compared to LLE. Both of these extraction methods are effective at removing matrix effects and avoiding extraction of any ionized compounds that are associated with protein precipitation for vitamin D analysis. SLE is preferred over LLE for the development of high throughput assay for vitamin D LC-MS/MS analysis (Jenkinson *et al.*, 2014).

5. **HPLC tandem with LC-MS**: HPLC tandem with LCMS enables for the analysis of all relevant vitamin D metabolites from a range of biological materials, such as serum or a dried blood spot, in a particular and selective manner. It develops a fast, low cost and high-output method of serum sample preparation using protein precipitation. Organic solvent is served for this purpose. Several substances were examined and among them only acetonitrile, methanol and their mixtures with zinc sulfate were chosen. The highest recovery values for the vitamin D metabolites were obtained for acetonitrile, with an organic solvent to serum ratio of 8:1 spot. Vitamin D metabolites are bound to proteins, the relationship between the content of organic solvent in the sample preparation process and their release from the protein complex was examined. The results indicate that the organic solvent content should be 30-70 percent in order to completely release the tested compounds from the proteins. The developed chromatographic method has eliminated false positive signals for the $24,25(OH)_2D_3$ metabolite. (Rola *et al.*, 2020).

Vitamin D metabolism is too complex. So the best method for sample preparation is that which maintains within low expenses and less time consumption. There are two stages in the method of vitamin D sample preparation: extraction and purification. Extraction undergoes two more steps: saponification (SN) and protein precipitation (PP). While purification includes two sub-phases: liquid-liquid extraction (LLE), solid phase extraction (SPE) or supported liquid extraction (SLE). LLE is expensive, time-consuming, multi-stage method and difficult to automate. For reliable quantitative analysis release of the studied analytes from the carrier compounds is essential. For this purpose, PP is used, which involves the destruction of the tertiary structure of proteins. Organic solvent such as acetonitrile aggregates proteins due to the increased electrostatic interaction between the charged molecules. It has been found that extraction of vitamin D is 50 percent in SN method in contrast to compared to PP. Compared to LLE, PP needs smaller volume of less hazardous organic solvents, and is cheap, simple and versatile method, especially adapted to routine LC-MS/MS analysis (Rola *et al.*, 2020). HPLC coupled with MS (LC-MS) offers both increased sensitivity and selectivity and minimizes interferences commonly seen from complex food matrices (Adamec *et al*, 2010).

3. GLOBAL DIET QUALITY SCORE (GDQS)

A. OVERVIEW OF GLOBAL DIET QUALITY SCORE

Global Diet Quality Score (GDQS) is a new word for many people. It has recently been defined by Harvard TH Chan, School of Public health. It determines the quality of food intake by an individual. It identify if the consumption pattern of the person is healthy or unhealthy. GDQS is not absolute intake of foods; rather it is an average intake of foods.

In 2017, Global Burden of Diseases mentioned that poor quality diets are associated with adverse health outcomes related to both undernutrition and overnutrition and can be the leading cause of disease globally (Global Burden of Disease, 2017). The consequences are due to lack of awareness, vicious circle of poverty, social stigma, misconceptions, imbalance food intake etc. Very fewer studies have been carried out to conclude the standard, simple and validated method to routinely measure the quality of diet in population based surveys, and therefore have fewer means to assess and track this critical dimension of health and well-being.

The concept of GDQS was launched by the researchers of Harvard TH Chan, School of Public Health, Department of Nutrition and Public Health (INSP), Mexico. In 2018 to support this concept as consortium, Intake – Center for Dietary Assessment launched a 2-year research initiative. To develop and validate metrics of diet quality, data were collected through routine population-based surveys for inclusion in global monitoring frameworks (Harvard T.H Chan, School of Public Health, Mexico, 2021).

GDQS is food based matrices with quality and quantity of food types and foods intakes with specific scoring method. Secondary food frequency questionnaire (FFQ) and quantitative 24-hour dietary recall datasets across different regions of the world were analyzed over the course of the 2-year research initiative to examine the association of each candidate metric with a range of diet quality outcomes related to nutrient adequacy and non-communicable disease (NCD) risk (Harvard T.H Chan, School of Public Health, Mexico, 2021). The inclusion of two cohort datasets in the analyses (one from Mexico and one from the United States) allowed for the evaluation of the responsiveness of outcomes to changes in metric score over time and provided a rigorous design for examining the association of the candidate diet quality metrics developed with NCD risk-related outcomes. From these analyses, an overall metric of diet quality — the Global Diet Quality Score (GDQS) — was identified (Harvard T.H Chan, School of Public Health, Mexico, 2021).

B. SCORING PATTERN

(i) Design for Global Diet Quality Score (GDQS) Metrics

GDQS metrics is a food based measurement system which consists of 25 food groups: 16 healthy food groups, 7 unhealthy food groups, and 2 food groups (red meat, high-fat dairy) that are unhealthy when consumed in excessive amounts. The terms "Healthy", "Unhealthy" and "Unhealthy if consumed in excessive amounts" deliver information to individuals about their quality of food intake and how it could impact the health. The food consumption will be shown in grams/day. The scores are divided into three categories: low, medium and high; except on one food group i.e., high fat dairy. It has four categories: low, medium, high and very high (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Under low, medium, high and very high (for high fat dairy only) categories certain scores has been assigned. Under each food groups scores are different and amount of consumption varies for each food groups (Harvard T.H Chan, School of Public Health, Mexico, 2021).

The GDQS is calculated by summing all the points gained in each 25 groups in GDQS food chart (Harvard T.H Chan, School of Public Health, Mexico, 2021). It was documented the total range of GDQS is 0 to 49 with respect to both nutrient adequacy and diet-related NCD risk. For the 16 healthy food groups more scores are given for higher intake. For 7 unhealthy food groups more points were assigned for lower intake. While for the remaining 2 food groups that classified as unhealthy when consumed in excessive amounts increasing points are given until specific amounts have been consumed, after which no points are given (Bromage *et al.*, 2021). Population-based cutoffs of 15 and 23 have been identified for the GDQS to indicate the quality of diet (Harvard T.H Chan, School of Public Health, Mexico, 2021). It has also been found that GDQS <15, indicates percent of the population high risk for poor diet quality outcomes indicates high risk of nutrient adequacy and NCD risk, scores ≥15 and <23 indicate moderate risk and GDQS ≥23 percent of the population at low risk for poor diet quality outcomes i.e., low risk of nutrient adequacy and NCD risk (Bromage *et al.,* 2021). The 25 food groups which are included in GDQS are as follows:

Table I: GDQS food groups (Harvard T.H Chan) with Indian Food names (from IFCT-2010)

Sl.No.	Food groups	Food names
1.	Citrus Fruits	Lemon juice, sweet lime pulp, orange pulp, grapefruit, tangerine,

		gooseberry, pummelo
2.	Deep orange fruits	Apricot, Papaya, ripened mango, musk melon (orange flesh), palm fruit, plum
3.	Other fruits	Apples, guava, pineapple, dates, sapota, avocado, banana, grapes, jack fruit, lichi, pear, pomegranate, raisins, star fruit, water melon, wood apple
4.	Dark green leafy vegetables	Agathi leaves, Green Amaranth leaves, bathua leaves, fenugreek leaves, drumstick leaves, goku leaves, mustard leaves, pumpkin leaves, radish leaves, spinach
5.	Cruciferous vegetables	Cabbage, broccoli, Brussels sprouts, cauliflower, collard greens, kale, kohlrabi, turnip, rutabaga, rape, napa cabbage, bok choy, brown mustard
6.	Deep orange vegetables (Retinol >100g)	tomato, pumpkin, carrot
7.	Other vegetables	Field beans, French beans, onion stalk, bamboo shoot, bitter gourd, bottle gourd, ash gourd, brinjal, capsicum, cho-cho,cucumber, drumstick, raw jack fruit, ladies finger, raw mango, raw, peas, plaintain green, plaintain fruit, ridge gourd, snake gourd, zucchini,beet roots, lotus root, beans sprouts
8.	Legumes	Legumes and foods derived from legumes, such as tofu and soymilk.
9.	Deep orange tubers	Yam, sweet potato
10.	Nuts and seeds	Almonds, cashew nuts, gingelly seed/sesame, ground nut, mustard seeds, linseeds, pistachio, safflower seeds, walnut, peanut butter or other seed related butter
11.	Whole grains and its products (not added significant amount of sugar or artificial sweetness)	Millet, wheat, brown rice, cornflakes, oatmeal, popcorn, wheat pasta or noodles, Quinoa, sorghum, rice, bulgur, corn, rye, brown bread, multigrain bread Whole grains and whole grain products.
12.	Liquid oils	Mustard oil, coconut oil, groundnut oil, olive oil, sesame oil
13.	Fish and shellfish	Both marine and fresh water fishes and shell fishes, also their

		organs
14.	Poultry meat	Chicken, pigeon, duck, include organ meat
15.	Low-fat dairy (≤2% milk fat)	flavored milk, and milk added to coffee or tea, skim milk, double toned milk
16.	Eggs	Hen egg, duck eggs, turkey eggs, quail eggs
17.	High-fat dairy (>2% milk fat)	Flavored milk, milk shakes, and milk or cream added to coffee or tea. [Does not include butter, ice cream and whipped cream]
18.	Red meat	Mutton, pork and lamb.
19.	Processed meat	Processed red meat, poultry, or game, including organs [excluding fish or seafood]
20.	Refined grains and baked goods [no significant amounts of added sugar]	Maida, semolina, white bread, biscuits, cornflour, vermicilli
21.	Sweets and ice cream	Non-beverages, commercial and homemade sweets, Whipped cream, ice cream; sugar and other caloric sweeteners added to other foods, jam, jelly, marmalade
22.	Sugar-sweetened beverages	Soft drinks, sodas, energy drinks, and sports drinks, diet sodas/cokes.
23.	Juice	Unsweetened or sweetened drinks composed of fruit juice, fruit smoothies made from whole fruit, fruit squash made of whole fruit
24.	White roots and tubers	Radish, water chestnut, potato, cassava/tapioca, cassava flour
25.	Purchased deep fried foods	Fries like patato fries, pakodas, bhajis, fried street foods

(Source: Harvard T.H Chan, School of Public Health, Mexico, 2021)

(ii) Design for Sub-Metrics

Along with the metric score, there are sub-metric scores. It means GDQS is divided into positive GDQs (GDQS+) and negative GDQS (GDQS-). The GDQS+ and GDQS- quantify the collective contribution of healthy foods and unhealthy, respectively, to overall diet quality (Bromage *et al.*, 2021). The GDQS+ is the total score across the 16 healthy GDQS food groups, with a possible range of 0 to 32. The total score for GDQS− lies between 0 to 17 for the 7 unhealthy GDQS food groups and the 2 GDQS food groups that are unhealthy when consumed in excessive amounts. The GDQS has so much potential

that its submetrics can be subdivided even further to provide more precise information on the influence of smaller groupings of food groups or individual food groups to diet quality in populations (Harvard T.H Chan, School of Public Health, Mexico, 2021).

The scores for each food groups are as follows:

Table II: Score assigned to certain score range for each food groups

FOOD GROUPS	**CATEGORIES OF AMOUNT OF FOOD CONSUMED (g/day)**				**POINT VALUES FOR EACH CATEGORY**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
GDQS+ (healthy)								
Citrus fruits	<24	24-69	>69		0	1	2	
Deep Orange fruits	<25	25-123	>123		0	1	2	
Other fruits	<27	27-107	>107		0	1	2	
Dark GLV	<13	13-37	>37		0	2	4	
Cruciferous vegetables	<13	13-36	>36		0	0.25	0.5	
Deep orange vegetables	<9	9-45	>45		0	0.25	0.5	
Other vegetables	<23	23-114	>114		0	0.25	0.5	
Legumes	<9	9-42	>42		0	2	4	
Deep orange tubers	<12	12-63	>63		0	0.25	0.5	
Nuts and seeds	<7	7-13	>13		0	2	2	
Whole grains	<8	8-13	>13		0	1	2	
Liquid oil	<2	2-7.5	>7.5		0	1	2	
Fish and shellfish	<14	14-71	>71		0	1	2	
Poultry and game meat	<16	16-44	>44		0	1	2	
Low fat dairy	<33	33-132	>132		0	1	2	
Eggs	<6	6-32	>32		0	1	2	
GDQS- (unhealthy when consumed in excessive amount)								
High fat dairy	<35	35-142	142-734	>734	0	1	2	0
Red meat	<9	9-46	>46		0	1	0	

GDQS- (unhealthy)								
Processed Meat	<9	9-30	>30		2	1	0	
Refined grains and baked goods	<7	7-33	>33		2	1	0	
Sweat and ice cream	<13	13-37	>37		2	1	0	
Sugar sweetened beverages	<57	57-180	>180		2	1		
Juice	<36	36-144	>144		2	1	0	
White roots and tubers	<27	27-107	>107		2	1	0	
Purchased deep fried foods	<9	9-45	>45		2	1	0	

(Source : Bromage S et al, 2021, Development and Validation of a Novel Food-Based Global Diet Quality Score GDQS)

Table III: Distribution of Sub-Metrics

	GDQS	**GDQS+**	**GDQS-**
Number of metric components	25	16	9
Food composition data needed for analysis	No	No	No
Number of consumption amount or consumption frequency categories used	3 or 4	3	3 or 4
Scoring method	Points assigned to each food group based on g/day categories of consumption amount associated will each food group	Positive points for the healthy foods	Negative points for unhealthy foods and healthy if taken in large amounts.
Score range	0-49	0-32	0-17
Vegetables			
Whole fruits			
Dark-green leafy vegetables	↑	↑	

Vitamin A rich fruits and vegetables			
Deep orange fruits	↑	↑	
Deep orange vegetables	↑	↑	
Deep orange tubers	↑	↑	
White roots and tubers	↓		↓
Cruciferous vegetables	↑	↑	
Other vegetables	↑	↑	
Citrus fruits	↑	↑	
Other fruits	↑	↑	
Meat, poultry and fish			
Fish and shellfish	↑	↑	
Poultry and game meat	↑	↑	
Red meat			
Processed meat	↓		↓
Legumes	↑	↑	
Pulses			
Nuts and seeds	↑	↑	
Dairy			
Low-fat dairy	↑	↑	
High-fat dairy			
Eggs	↑	↑	
Grains, white roots, tubers and plaintains			
Whole grains	↑	↑	
Refined grains and baked goods	↓		↓
Sugar sweetened beverages	↓		↓
Juice	↓		↓
Sweets and ice cream	↓		↓
Purchased deep fried foods	↓		↓
Liquid oils	↑	↑	
PUFA (no EPA or DHA)			

Trans fatty acids			
Alcohol			
Omega-3 fatty acids (EPA and DHA)			
Sodium			

(Source:Bromage S et al, 2021, Development and Validation of a Novel Food-Based Global Diet Quality Score GDQS)

C. DATA COLLECTION TOOLS

There are many ways which can be used for calculating the GDQS of each individual. According to the availability of the data, purposive, time period of the study methods can been selected. These methods include: (i) Food Frequency Questionnaire (FFQ), (ii) 24-hour dietary recall (Harvard T.H Chan, School of Public Health, Mexico, 2021). Among these two methods, quantitative 24-dietary recall is the most appropriate method.

(i) 24- hours dietary recall Method

This method involves six steps for data collection.

Step 1: Process the dietary data to reflect gram consumption of each single food/ingredient

At first the data should be processed so that the total consumption amount (in grams) for each food consumed during the 24-hour recall period is available, by respondent. Mixed dishes should be disaggregated to each of their ingredients using recipe information. Weight will be done depending on what form it has been consumed (e.g., raw, cooked). Certain foods such as breads, cakes, and biscuits should be considered as "single foods" for the analysis and processing of data (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 2: Assign each single food/ingredient reported as consumed to its correct GDQS food group

Now each component of the food needs to be classified into the corresponding GDQS food group. There are certain food which are not classified in any of the food groups such as alcohol, semi-solid and solid fats, insects, coconuts, and coconut products (except liquid coconut oil) (Harvard T.H Chan, School of Public Health, Mexico, 2021).

- **High-fat dairy**

In GDQS high-fat dairy food group includes all high-fat dairy products except ice cream and whipped cream. It is essential during data processing to subdivide the high-fat dairy group into a high-fat "hard cheese" and a high-fat "other dairy" sub-category (Harvard T.H Chan, School of Public Health, Mexico, 2021).

- **Liquid oil**

It is difficult to get information on the actual amount of liquid oil consumption by each respondent. If a food that is deep fried and purchased is deep fried using liquid oil, the oil is not classified in the liquid oil food group. If the food is deep fried at home using liquid oil, the oil is classified in the liquid oil group (Harvard T.H Chan, School of Public Health, Mexico, 2021).

- **Purchased deep fried food**

The purchased deep fried food category includes foods which are bought exclusively from outside and fried. Some of these foods are "double-classified": purchased deep fried and normal food group. For instance, deep fried white potatoes should be classified both in the purchased deep fried food category and in the white roots and tubers category (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 3: Sum the quantity of consumption (in grams) of all foods reported as consumed for each GDQS food group, by respondent

Respondents should have a total consumption value in grams for each GDQS food group (except the liquid oil GDQS food group). If any group remains un-responded the respondent should be given a value of 0 grams for quantity of consumption for that specific food group (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 4: Assign each respondent to a quantity of consumption category per GDQS food group

Each respondent is assigned to a category of quantity of consumption for each GDQS food group based on the quantity of consumption values per GDQS food group.

After implementation Step 4, each individual in the dataset should be classified into one of three categories of quantity of consumption for 24 of the GDQS food groups and into one of four categories of quantity of consumption for the high-fat dairy GDQS food group (i.e., the "high-fat dairy " food group) (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 5: Assign points to each GDQS quantity of consumption category and sum, by respondent

Each respondent got certain points for each GDQS food group. The GDQS variable is tabulated by summing points across all 25 GDQS food groups, by respondent. For the GDQS+, summation of points across the 16 healthy GDQS food groups by respondent is required. For GDQS–, summation of points across the 7 unhealthy GDQS food groups and the 2 GDQS food groups that are unhealthy when consumed in excessive amounts is done (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 6: Calculate the population-based value of the GDQS

To produce a population-based statistic, the GDQS data for each individual must be examined. The mean GDQS, the mean GDQS+, and the mean GDQS– can be calculated for the population. Then a cutoff value of <15 and ≥ 23 can be applied to the GDQS variable. (i) GDQS ≥ 23 - low risk of nutrient inadequacy and NCD-related outcomes (ii) ≥15 and <23 - moderate risk (iii) <15 - high risk. The percent consuming low, middle, and high (very high for high-fat dairy) categories of quantity of consumption in the reference 24- hour period should be reported for each GDQS food group (Harvard T.H Chan, School of Public Health, Mexico, 2021).

(ii) Food Frequency Questionnaire

It consists of seven steps, as follows:

Step 1: Process the dietary data to reflect gram consumption of each single food/ingredient

For FFQ, the data should record the total consumption amount (in grams) for each available food for the reference period of data collection. Mixed dishes should be disaggregated to each of their ingredients using recipe information. Weight will be done depending on what form it has been consumed (e.g., raw, cooked). Certain foods such as breads, cakes, and biscuits should be considered as "single foods" for the analysis and processing of data (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 2: Process the food frequency data to reflect gram consumption for a 24-hour reference period

Then the data should be converted into quantities consumed (in grams) for the reference period. For this, consumption amounts are divided for each food by the number of days that constitute the reference, by respondent.

If the FFQ used a 7-day reference period, then the amount for consumption each respondent should be divided by 7 to get an average gram consumption amount for a 24-hour period. While converting a

reference period of 1 month into a 24-hour period, a conversion factor of 30.4 days should be used (i.e., 365 ÷ 12 = 30.4) (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 3: Assign each single food/ingredient as consumed to its correct GDQS food group

Now each component of the food needs to be classified into the corresponding GDQS food group. There are certain food which are not classified in any of the food groups such as alcohol, semi-solid and solid fats, insects, coconuts, and coconut products (except liquid coconut oil) (T.H Chan, School of Public Health, Mexico). These are given below:

- **High-fat dairy**

In GDQS high-fat dairy food group includes all high-fat dairy products except ice cream and whipped cream. It is essential during data processing to subdivide the high-fat dairy group into a high-fat "hard cheese" and a high-fat "other dairy" sub-category (T.H Chan, School of Public Health, Mexico).

- **Liquid oil**

It is very difficult to get exact information on how much liquid oil was consumed by each respondent.

- **Purchased deep fried food**

The purchased deep fried food category includes foods which are bought exclusively from outside and fried. Some of these foods are "double-classified": purchased deep fried and normal food group. For instance, deep fried white potatoes should be classified both in the purchased deep fried food category and in the white roots and tubers category (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 4: Sum the quantity of consumption (in grams) of all foods/ingredients reported as consumed for each GDQS food group, by respondent

After the foods have been reported as consumed they have been classified into the correct GDQS food group and are summed. If any GDQS food group not reported by the respondent then 0 grams for quantity of consumption will be assigned for that food group (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 5: Assign each respondent to a quantity of consumption category per GDQS food group

The quantity of consumption values (in grams) per GDQS food group are used to assign each respondent to a category of quantity of consumption for each GDQS food group. After completing Step

5, each individual in the dataset should be classified into one of three categories of quantity of consumption for 24 of the GDQS food groups and into one of four categories of quantity of consumption for the high-fat dairy GDQS food group (i.e., the "high-fat dairy" food group) (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 6: Assign points to each GDQS quantity of consumption category and sum, by respondent

Each respondent got certain points for each GDQS food group. The GDQS variable is tabulated by summing points across all 25 GDQS food groups, by respondent. For the GDQS+, summation of points across the 16 healthy GDQS food groups by respondent is required. For GDQS–, summation of points across the 7 unhealthy GDQS food groups and the 2 GDQS food groups that are unhealthy when consumed in excessive amounts is done (Harvard T.H Chan, School of Public Health, Mexico, 2021).

Step 7: Calculate the population-based value of the GDQS

To produce a population-based statistic, the GDQS data for each individual must be examined. The mean GDQS, the mean GDQS+, and the mean GDQS– can be calculated for the population. Then a cutoff value of <15 and ≥ 23 can be applied to the GDQS variable to create a population-level categorical indicator. (i) GDQS ≥ 23 - low risk of nutrient inadequacy and NCD-related outcomes (ii) ≥15 and <23 - moderate risk (iii) <15 - high risk. The percent consuming low, middle, and high (very high for high-fat dairy) categories of quantity of consumption in the reference 24- hour period should be reported for each GDQS food group (Harvard T.H Chan, School of Public Health, Mexico, 2021).

C. IMBALANCE IN GLOBAL DIET QUALITY SCORE AND RELATED DISORDERS

Increase and decrease in GDQS is related to various kinds of diseases such as undernutrition, obesity, and type 2 diabetes mellitus. A survey was conducted in China under China National Nutrition and Health Survey to evaluated the double burden among individuals of >18 years. Metabolic syndrome and nutritional deficiency were identified as a double burden. This survey was conducted from 2010-2012. A total of 35,146 individuals were taken for the survey among whom 14,978 were men and 20,168 were women. GDQS score was calculated by using average intakes of 25 food groups for consecutive 3 days through 24-hour dietary recalls. It was found a higher GDQS score was inversely linked to metabolic syndrome and nutrient deficiency, or both. And GDQS was significantly higher in urban areas than the rural areas (Yuna. *et al*., 2021).

A study was conducted among Ethiopian adults to find the relation between Higher Nutrient Adequacy, Midupper Arm Circumference, Venous Hemoglobin, and Serum Folate and GDQS. Secondary analyses of FFQ and 24-hour recall (24HR) data from a population-based cross-sectional survey of nonpregnant, nonlactating women of reproductive age and males (15–49 years) in Addis Ababa and five predominantly rural regions yielded a comparison metrics. From the FFQ analysis a correlations between the GDQS and aggregate measure of dietary protein, fiber, calcium, iron, zinc, vitamin A, folate, and vitamin B12 adequacy were 0.32 in men and 0.26 in women were measured. The following results were found: (i) inverse relation between GDQS and folate deficiency in men and women (ii) inversely associated with underweight (iii) low midupper arm circumference (iv) inverse relation to anemia in women (v) positively associated with hypertension in men (Bromage *et al*., 2021).

Another study was conducted among 8967 Mexican women of reproductive age (non-pregnant, non-lactating) from 2006 to 2008 among, aged 25–49 years. It designed a regression models to understand 2-years changes in the GDQS and each food group with weight and waist circumference changes within the same period. It found that women with the largest increase in the GDQS had less weight and waist circumference gain. While women with the largest decrease in the GDQS had more weight and waist circumference gain (Angulo *et al*., 2021).

4. MACHINE LEARNING

A. DEFINITION OF MACHINE LEARNING

Machine Learning has been defined in various ways. Ethem Alpaydin defined machine learning in his textbook as "Programming computers to optimize a performance criterion using example data or past experience" (Alpaydin *et al.*, 2014). The sole purpose of machine learning is to imitate the characteristics and behavior of human beings and learn to process sensory (input) signals to achieve a goal (Sak and Suchodolska, 2021).

B. HISTORY OF MACHINE LEARNING

Machine learning is the sub type of Artificial Intelligence. In 1955 the term "artificial intelligence" was by John McCarthy, an American computer scientist in a research project proposal. It was carried out the following year at Dartmouth College in Hanover, New Hampshire (MaCarthy *et al*., 2006).

Machine learning is not the new concept. It was back in seventh century by Pascal and Leibniz to develop machines that can emulate human ability (Ifrah, 2001). The term Machine learning was coined by Arthur Samuel from IBM in 1959. This led to the development of neural network architectures (perceptron) in 1958 by Rosenblatt. Later many perceptron were developed with many advantages and limitations as well. These led to development of perceptron from decades to decades. A massive breakthrough was accomplished with the advancement of multilayer perceptron (MLP) by Werbos in 1975. The popularization of the use of ML algorithms took place in the last decade of the 20th century in search engine applications (Sak and Suchodolska, 2021).

C. CONCEPT OF MACHINE LEARNING

ML is a sub category of AI area which is based on algorithms that develop through experience by its own. It emulates human intelligence by learning from the surrounding environment without any teacher to derive a certain outcome. These algorithms are soft coded and the output is hard coded. They automatically adapt the characteristics through experience/repetition so that they can work better at achieving the desired result. ML algorithms can create mathematical models which is required for decision making. ML involves a large set of trained data without being programmed (El Naqa and Murphy, 2015). It has taken ideas from various disciplines such as probability, artificial intelligence, computer science, statistics, information technology and control theory (Bishop, 2006).

The process by which data are adapted in Machine Learning is called training. For this input data taken from large samples are given along with preferred results. The algorithm optimally configures to not only produce the desired outcome with the trained data, but can also simplify to construct the desired outcome from new data. This training of data is referred to as “learning” part of machine learning. Algorithm used in ML has role to determine probability distributions from the input data and use them to predict outcomes (El Naqa and Murphy, 2015).

There are two important merits of algorithm. (i) it can substitute human effort in effective way (ii) it can learn complicated and restrained patterns in the input data that the average human observer is unable to do (El Naka and Murphy, 2015).bFor the better performance of the machine learning techniques in a given dataset all the bias should be identified, removed or minimized (Caixinha and Nunes, 2017).

D. TYPES OF LEARNING APPROACHES OF MACHINE LEARNING

The way of learning by computer is divided into three parts: Supervised learning, unsupervised learning and semi-supervised learning.

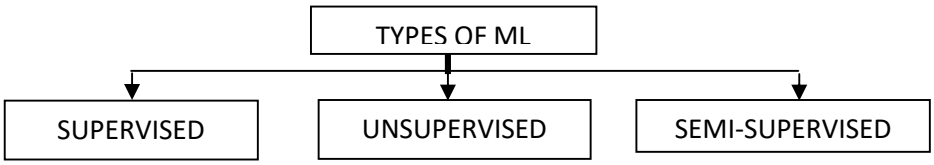
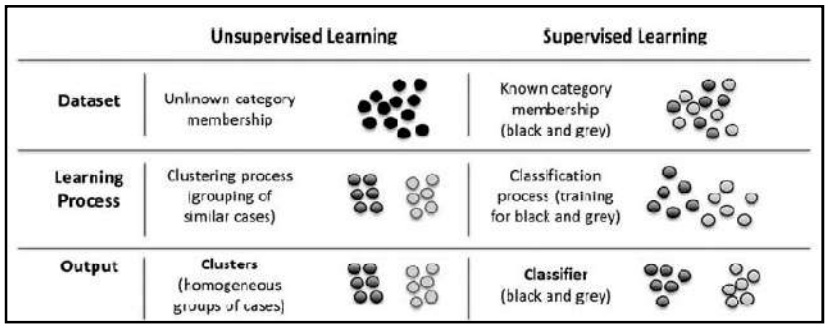


Fig. 2: Types of Machine Learning

1. **Supervised learning**: The purpose of the supervised learning is to create a function based on a trained dataset that maps data into the pre-existing groupings allowing to forecast new cases (Caixinha and Nunes, 2016). It means it interprets unknown output from known input. (El Naqa and Murphy, 2015). It functions to recognize digits, classify images and document, have regression functions and predict the category of new cases (Deo, 2015; Caixinha and Nunes, 2017). Supervised learning model is used in medicine (Deo, 2015). The most used techniques for supervised learning techniques are: linear discriminant analysis (LDA), support vector machines (SVMs), neural networks (NNs), Bayesian classifiers, k-Nearest Neighbor (kNN), and decision and classification trees (DCTs) (Caixinha and Nunes, 2016).

2. **Unsupervised learning**: In unsupervised learning no output can be predicted. Only input samples are given to the learning system. Instead naturally occurring patterns within the data are explored (El Naqa and Murphy, 2015). It also indicates that it allows the recognition of homogeneous groups in the input dataset (Caixinha and Nunes, 2016). No training is given to the labels in this type of learning. Various techniques used for unsupervised learning are fuzzy clustering, hierarchical clustering, K means clustering (Shailaja *et al.,* 2018).

3. **Semi-supervised learning:** Semi-supervised learning combines supervised and unsupervised learning by partially labeling data and using the labeled piece to infer the unlabeled portion (El Naka and Murphy, 2015). It transmits great classification performance by using unlabeled data (Shailaja *et al.,* 2018).



(Source: Caixinha and Nunes, 2017)

Fig.3: Illustration of the two machine learning approaches, the unsupervised approach and the supervised approach

E. TECHNIQUES OF ML

1. Support Vector Machine

In 1990's vector machine (SVM) is used for machine learning (ML). It is believed to be a simple and popular technique. In this technique samples are divided in group and trained data is collected. The main use of Support vector machine (SVM) is in classification and regression problems (Murphy, 2012).

2. Decision Tree

Most popular technique for classification in ML is Decision Tree (DT), contains internal node and one leaf node with a class label. It is more like a tree-like model or graphs. DT can capture the decision-making knowledge from the given data (Sambasivam *et al*., 2020).

3. K-Nearest neighbor (KNN)

When there need to frequently used approach for classification of samples then K-nearest neighbor is used. It is used to calculate distance measure from N number of training samples (Shailaja, 2018).

4. Artificial Neural Networks

Artificial neural networks (ANNs) are collection of simple processing nodes (units), interconnected with each other to increase the computational power over any single node. In this, the input nodes are the observations that are used for prediction. Other nodes are calculated from the values of the input, and are used to calculate the values for the output (Meyfroidt *et al.*, 2008).

5. Random Forest Classifier

Breiman proposes the concept of Random forest classifier (RF) in machine learning method used to solve classification problems. RF constitutes many decision trees randomly from the training set and then it aggregates the values from different decision trees and predicts final severity deficiency as the outcome (Sambasivam *et al*., 2020).

METHODOLOGY

III. METHODOLOGY

The methodology for the study of **"Exploration of Machine Learning to develop a Low cost Screening Method with Global Diet Quality Score to detect Vitamin D deficiency"** can be presented in the following ways:

Phase I – Selection of areas and sample

Phase II – Selection of research tool

Phase III – Conducting survey on GDQS (Global diet quality score) and dietary intake of Vitamin D and data segregation

Phase IV – Validation of low cost and effective screening method and Exploration of Machine Learning

Phase V – Statistical analysis of the GDQS, Vitamin D intake data and secondary serum Vitamin D levels

The Research Resign of the study is presented in the figure 4.

PHASE I

SELECTION OF AREA AND SAMPLE

i. Selection of area

The first step in this study is the selection of the area and to understand the demography needed for this study. As this study is about detecting vitamin D deficiency with the help of the food pattern of the subjects and their Vitamin D serum level, therefore, considering wide region was more beneficial.

Many studies have been done in South India and it was found that deficiency of Vitamin D in South India is not uncommon. Sunlight is the primary source of Vitamin D. Its exposure accounts to elevate or maintain the level of Vitamin D. It was found that 69 percent people were tested to be Vitamin D deficient and 15 percent had insufficient levels of Vitamin D. Some previous studies had already been conducted on Vitamin D insufficiency and obesity among young women (Gowthami and Kalpana, 2022) and Impact of Nutrition intervention and dress code on Vitamin D nutriture of Muslim women (Habeeba and Kalpana, 2022). So the area selected for this study is Tamil Nadu.

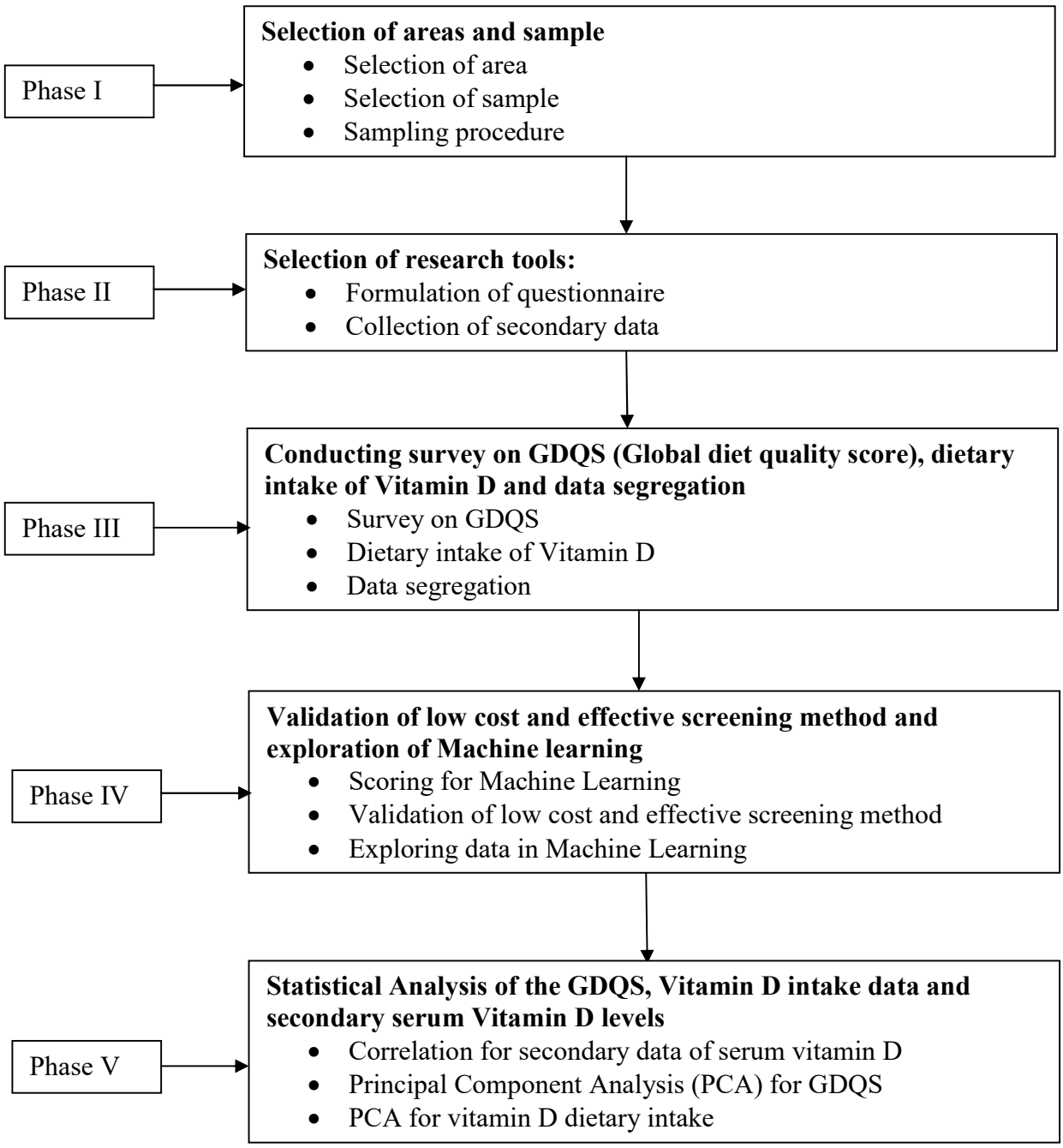


Fig. 4: Research Design

ii. Selection of sample

Sample is a subset of the larger population with similar characteristics which will be used in the study. Females are mostly deficient of Vitamin D deficiency as compared to males. Young adults who can be easily vitamin D sufficient, mostly remains with Vitamin D insufficiency. These young adults belong to age group of 18-25 years. These age group people are mainly college or university going students or working women, while some are found to be married. Bone differentiation disorders and other disorders can occur among them in later period of their life. So the study was conducted among young female adults in the range of 18- 25 years.

For GDQS, 150 numbers, which constitute adults females of age group 18-25 years, were taken. For Vitamin D dietary intake 150 sample size was chosen of age group 18-25 years of females. While for Serum Vitamin D 50 sample size was taken. All the females were University students from Tamil Nadu.

(a) Procedure of Sampling –

Selection of non probability sampling was done for its more suitability in this study. Among all the sampling methods of non-probability sampling, convenience sampling and snowball sampling is selected. For convenience sampling the target population had meet these certain criteria such as different dietary pattern and dietary habits, variation in the availability of foods, vitamin D and calcium availability, willingness of the respondents.

Age group 18-25 years is the young adults who can get deficient of Vitamin D for various factors. In general, these age group individuals go outside and get exposure to sunlight but still Vitamin D seen among them. There is certain inclusion and exclusion criteria taken in consideration for selecting the sample, these are as follows:

(b) Inclusion criteria:

- Young adults of 18-25 years
- Female respondents only
- State of Tamil Nadu
- Having smart phones with both android and iOS system or having computer or Tablet PC

(c) Exclusion criteria:

- Any persons who is having medication due to any kind of diseases

- Person younger than 18 years or older than 25 years
- Those who do not have mobile phones or computer

PHASE II

SELECTION OF RESEARCH TOOL

For this study questionnaire is the main research tools as the primary sources while some already existed databases collected from trusted sources were acted as secondary sources. Questionnaire contained mostly closed ended questions and few open ended questions.

i. **Formulation of questionnaire**

The questionnaire is formulated to find the dietary intake of specific foods on the regular basis and vitamin D intake through foods. The whole questionnaire was made by referring GDQS method which includes 25 food groups with certain scores associated with certain foods. Also questions regarding dietary intake of Vitamin D has been formulated. Fifty six questions were closed ended questions while six questions were open ended questions. This questionnaire was used in the survey to understand the distribution of adequate intake of Vitamin D solely through diet. Through GDQS we understood if the person's dietary pattern is healthy or unhealthy. Along with that, foods containing moderate to high levels vitamin D in food were also highlighted.

ii. **Collection of secondary data**

Requirement of Secondary data for this study was very crucial. Secondary data was collected from the previous studies collected among young adult women. The data were the estimated values of serum Vitamin D levels in humans and nutrient factors affecting them.

PHASE III

CONDUCTING SURVEY ON GDQS (GLBAL DIETARY QUALITY SCORE), DIETARY INTAKE OF VITAMIN D AND SEGREGATION

i. Survey on GDQS

Survey is the method which is used to collect data from a large population with the given set of questions. It can be done online or offline depending on the convenience of the researcher. This survey was conducted online by distributing Google forms that had been sent to the participants through

whatsapp, instagram and ask them to fill the survey form and to share with others with the same age group.

The questionnaire was used in the survey with the sole purpose to collect data from the subjects. There was three parts in the questionnaire – personal data, GDQS (Global Diet Quality Score) related food groups, vitamin D intake through diet.

GDQS is a type of dietary score in which the diet score is calculated from the type and amount of the food consumption by each individual. The dietary score has already been assigned for each food type and amount by the Intake – Centre for Dietary Assessment (Harvard T.H Chan, School of Public Health, Mexico, 2021).

In this GDQS method there are 25 food groups. Among them, 16 are healthy food groups, 7 unhealthy food groups, and 2 food groups (red meat, high-fat dairy) that are unhealthy when consumed in excessive amounts. Under consumption of certain amount of food in the already designed range, there are specific scores. Total GDQS range is 0-49. Total score from each food groups indicates the health risk.

ii. Dietary Intake of Vitamin D

Among these 25 food groups, five groups has been segregated which contain moderate to high levels of vitamin D. These are:

(i) Whole grains (moderate to high)
(ii) Poultry (moderate levels in liver)
(iii) Eggs (moderate levels)
(iv) Fish and shellfish (specific types)
(v) Red meat (liver)

Vitamin D levels of eggs, poultry and red meat depends on the exposure of poultry birds and goat, sheep, cow, pig to the rate of their sun exposure.

The questionnaires are enclosed in Appendix (I and II).

iii. Data segregation

After collection of data from both primary and secondary sources, datas was studied thoroughly, selected and segregated according to the requirement of the study. Unnecessary data was taken off from the main data viz., the persons who were > 25 years and <17 years were excluded from the study.

Phase IV

VALIDATION OF LOW COST AND EFFECTIVE SCREENING METHOD AND EXPLORATON OF MACHINE LEARNING

i. Scoring for Machine learning algorithm

In the standard GDQS method, the range of score lies between 0-49 where 0-32 is contributed from healthy food groups and 0-17 contributed by unhealthy food groups. If the score is <15 then it indicates high risk and nutrient inadequacy, if ≥15 and <23 indicates moderate risk and ≥23 indicates less risk.

Because of complexity of response obtained, a new scoring has been devised for easy calculation and algorithm forming. These scoring are as follows:

Table IV: Scoring of GDQS which is especially made to use in this study

	Frequency of consumption	Range of consumption	Score
NORMAL	2-4 days per week	LL-UL	1
	5-6 days per week	<LL	1
	Once in a week	>UL	1
HIGH	2-4 days per week	>UL	2
	5-6 days per week	UL-LL/>UL	2
	Everyday	UL-LL/>UL	2
LOW	2-4 days per week	>UL	0
	Once in a week	UL-LL/>UL	0
	Seasonally/monthly	Irrespective of any amount	0
LOW	No consumption	Nil	0

***UL- Upper Limit of the consumption amount, LL- Lower Limit of the consumption amount**

ii. Validation of low cost and effective screening method

There are various techniques for Machine Learning but for this study Multi-class SVM classifier was used. Multiclass classification means the Machine Learning can classify the instances as three or more categories. SVM refers to Support Vector Machines. It is a supervised Machine Learning algorithm which basically functions in classification and regression problems. Its objective is to find an optimal boundary between the possible outputs. SVM supports binary classification in its old technique. For the multiclass classification, same principle is applied after splitting the multiclass problems into multiple binary classification problems. Its idea is to amplify data points to high dimensional space to gain mutual linear separation between every two classes.

For this study multi-classifier SVM classified the variables as:

(i) Frequency of consumption in a week

(ii) Amount of consumption within a certain range per time.

This technique is applied for the GQDS questions. It especially uses the columns which consist of Vitamin D rich foods such as whole grains, poultry, eggs, fish and shell fish and meat (liver).

iii. Exploring data in Machine Learning

Machine Learning algorithms can easily identify relationships between various data variables and dataset structures to determine whether outliers exists and create data values that can highlight patterns or points of interests. During exploring data in machine learning notable patterns were identified to draw conclusions from the datasets. Machine learning allows the users to extract information in large databases quickly.

After the data were explored in Machine Learning using a unique technique the machine learning was ready to expose to new and untrained dataset to understand the pattern (predicted form trained dataset). Steps in data exploring for Machine Learning:

(a) **Decide the problem or determine what to predict**: While finding the problem what to predict; classification, clustering and categories were done. The classification which was done for this study was multiclass classification. Then the data were clustered. In clustering the algorithm it found the rules of classification and number of classes. The algorithm ranked the objects by a number of features.

(b) **Establish data collection mechanisms:** Collecting data is an important part of predictive analysis in Machine learning. Real time data has been collected to record the consumption pattern of each individual in a particular food group in this study.

(c) **Learning**: For this, nine groups have been selected out of 50 groups depending on which groups are related to Vitamin D foods. It is also called as attribute sampling.

(d) **Data cleaning:** In data cleaning missing values have been substituted with some approximate values to ease the Machine Learning.

(e) **Rescale data:** Data rescaling is a type of data normalization. It improved the quality of databases by reducing dimensions and overweighed values (if present).

(f) **Discretize data:** In this step the continuous data was converted to a finite set of intervals and some specific data values were added in interval.

(g) **Data training:** In this step the finalized data was trained under the technique multi-class SVM and a certain pattern is predicted and classified with certain level of accuracy.

(h) **Exploring untrained data to the already trained data to predict the results:** The final step was the involvement of untrained/new data in the trained data. From the pattern which has already been predicted, the untrained dataset would fit itself in the previous dataset and would predict the result.

For this the dataset included three categories of Vitamin D rich food consumption from GQS: low, moderate and low with certain values (refer to Table IV). Any new data which fits in any of the category would suit itself in that and would predict the result.

Phase V

STATISTICAL ANALYSIS OF THE GDQS, VITAMIN D INRAKE DATA AND SECONDARY SERUM VITAMIN D LEVELS

The data which were obtained were raw in nature. So they were refined in the numerical form for the further statistical analysis. There are three different types of data that which are two primary data and one secondary data; and all of these data were treated in different way. SPSS was used: (i) to find correlation for secondary data of serum vitamin D and nutrient components (ii) Principal Component Analysis (PCA) for GDQs data and primary vitamin D intake data.

RESULT AND DISCUSSION

IV. RESULT AND DISCUSSION

The result of the study entitled **"Exploration of Machine learning to develop a low cost screening method with Global Diet Quality Score to detect vitamin D deficiency"** is presented and discussed here under the following heads:

1. **Age variation and serum vitamin D variation of the population**
2. **Simplifying the intake pattern of food groups under GDQS section**
3. **Simplifying the dietary intake of vitamin D rich foods**
4. **Correlation of serum Vitamin D and nutrient intake**
5. **Interpretation of Machine Learning**

A. AGE VARIATION AND SERUM VITAMIN D VARIATION OF THE POPULATION

a. Age variation

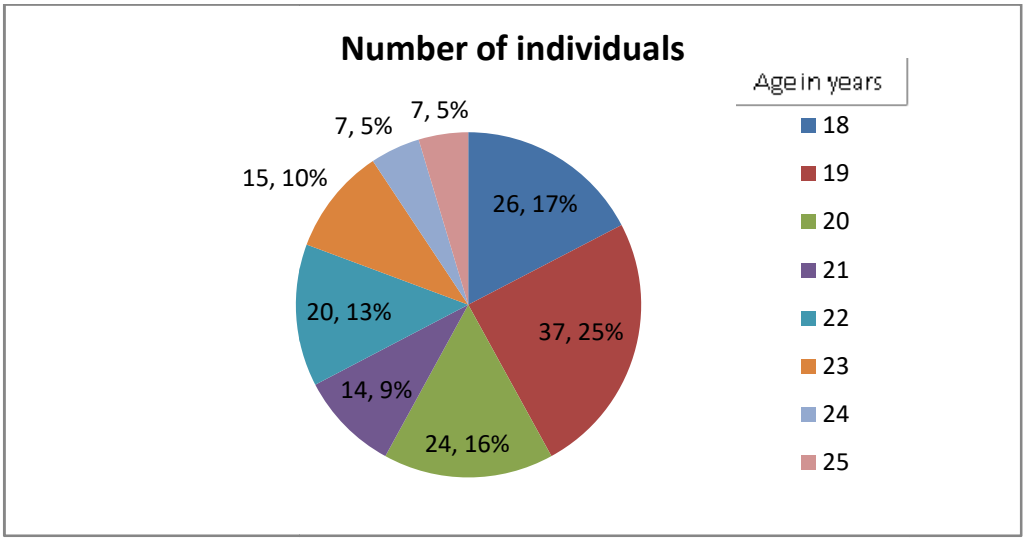


Fig. 5: Age variation of the selected population

Among 150 active participants, 26 participants belonged to age group of 18 years which constituted 17 percent. 37 participants belonged to the group of 19 years (25 percent). 16 percent constituted age group 20, i.e., 24 respondents. 21 respondents belonged to age group of 21 years which constituted 9 percent of the total study population. It was followed by 20 (13 percent) respondents of 22 years age and 15 (10 percent) respondents of 23 years age. There were 7 respondents from each of 24 years age group and 25 years age group. It formed 5 percent from both of the groups.

b. Serum Vitamin D variation

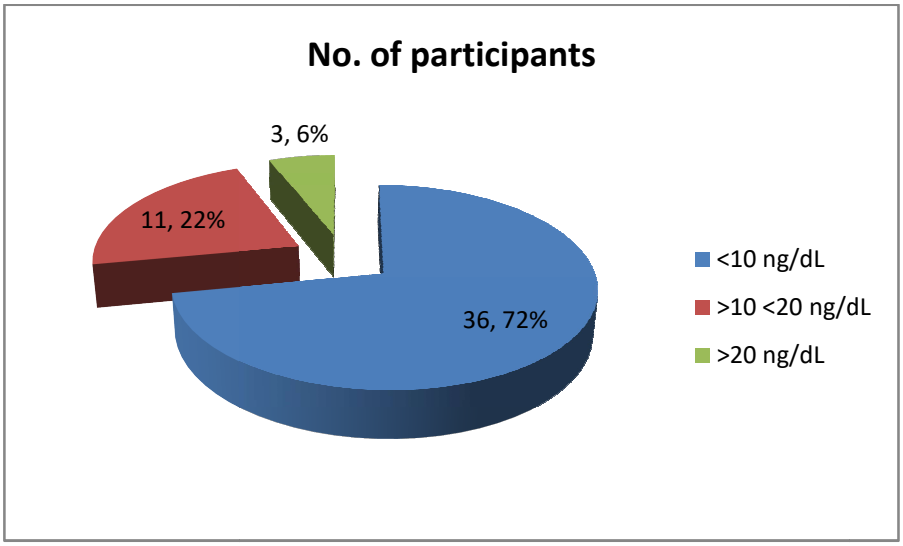


Fig. 6: Serum Vitamin D level among respondents

From the above pie chart (Fig. 6) it is clearly shown that 36 participants among 50 participants were having serum vitamin D <10 ng/dL. It constituted 72%. 11 participants were having serum vitamin D >10 ng/dL and <20 ng/dL (22 percent), followed by 3 participants from group >20 ng/dL (six percent). From this chart it can be concluded that most of the respondents in the population are deficient of Vitamin D while few participants were only having sufficient levels of serum Vitamin D.

In a study of 2007 in the North Indian Population it has been shown that 81% people are either deficient or insufficient of Vitamin D (Borkar *et.al*., 2010). In another study population it has been noted that 70% individuals in the country is deficient of vitamin D (Ritu and Gupta, 2014).

So it can be concluded from the results in the study and the results from the previous studies conducted, that majority of the populations are deficient of Vitamin D or there is insufficiency of Vitamin D.

B. **SIMPLIFYING THE COMPLEX DATA OF INTAKE PATTERN OF FOOD GROUPS UNDER GDQS SECTION**

The data which has been collected under the questionnaire of GDQS was huge in amount. It was needed to be simplified to interpret it. For this reason Principal Component Analysis (PCA) has been performed. PCA uses algorithms on the complex data to reduce it into correlated factors that gives conceptual understanding of the construct of interest. These factors are (inter-correlated factors) are calculated to give correlational matrix. These factors are extracted from the correlational matrix by the Principal component. The factors are then rotated to interpret and analyze the data. Two statistical calculation which helps to make decisions: Eigenvalues and Scree plot.

(a) Amount of variance

Amount of variance each variable accounted is referred to as Communalism. Initial communalities are estimates of the variance in each variable accounted for by all components. For principal component extraction, this is always equal to 1.0 for correlation analyses. The accepted cut-off value for communalism is ≥ 0.3. While the ideal value for communalism is ≥ 0.7.

Table V: Communalities of GDQS food group containing Vitamin D rich food groups

Communalities		
	Initial	**Extraction**
Age	1.000	.588
Whole Grains	1.000	.679
Amount of Whole grains	1.000	.707
Fish	1.000	.687
Amount of Fish	1.000	.681
Eggs	1.000	.535
Amount of eggs	1.000	.655
Meat	1.000	.733
Amount of meat	1.000	.712

From the given table (Table V) it can be noted that all the values are greater than 0.3. Three variables were above the ideal threshold value i.e., (i) Amount of whole grains (ii) Frequency of meat consumption (iii) Amount of meat. So the variance in all of the nine variables was justified.

(b) Checking the data if significant for further analysis

Eigenvalues is the total amount of variance that can be explained by a given principal component. It is essential for the eigenvalues to be greater than 1.0 for deciding if a factor could be further interpreted. They can be positive as well as negative.

Table VI: Extracted four essential Component matrix

Component Matrix[a]					
		Component			
		1	**2**	**3**	**4**
1.	Age	.014	.760	-.046	-.090
2.	Whole Grains	-.065	.545	-.610	.071
3.	Amount of Whole Grains	.138	.613	.520	.204
4.	Fish	0.13	.049	.492	.327
5.	Amount of Fish	.768	.151	-.149	-.213
6.	Eggs	-.429	.257	-.236	.479
7.	Amount of Eggs	.667	.185	.388	-.160
8.	Meat	-.527	.258	.142	-.608
	Amount of Meat	.709	-.036	.005	.456
a. 4 components extracted.					

Table VII: Total Variance for GDQS food group containing Vitamin D rich food groups

Total Variance Explained						
Component	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	2.357	26.191	26.191	2.357	26.191	26.191
2	1.444	16.042	42.233	1.444	16.042	42.233
3	1.135	12.607	54.840	1.135	12.607	54.840
4	1.040	11.553	66.393	1.040	11.553	66.393
5	.815	9.059	75.452			
6	.753	8.363	83.815			
7	.620	6.891	90.706			
8	.484	5.381	96.087			
9	.352	3.913	100.000			
Extraction Method: Principal Component Analysis.						

Nine important variables from the 50 variance have been selected depending on the presence of Vitamin D in those food groups. Four components have been extracted from the data of the GDQS. Eigenvalues is essential for getting the significance of data for further analysis. Eigenvalues for each component is different. It is calculated as: square of the summation of each component of each category (table VI):

For Component 1:

$(0.014)^2+(-0.065)^2+(0.\ 138)^2+(0.13)^2+(0.768)^2+(-0.429)^2+(0.667)^2+(-0.527)^2+(0.709)^2 = 2.357$

For component 2:

$(0.760)^2+(0.545)^2 +(0.613)^2+(0.049)^2+(0.151)^2+(0.257)^2+(0.185)^2+(0.258)^2+(-0.36)^2 =1.444$

It can be noted that values in the table VII under the column "Total" is equal to the calculated eigenvalues. So column "Total" is the eigenvalues. From the table 3, it is shown that eigenvalues of component 1 is 2.357, component 2 is 1.444, component 3 is 1.135 and component 4 is 1.040. Only these first four components having eigenvalues more than 1, so they are significant for the interpretation and analysis of the data.

(c) Deciding number of factors to be interpreted

Number of factors to be extracted depends on eigenvalues and scree plot.

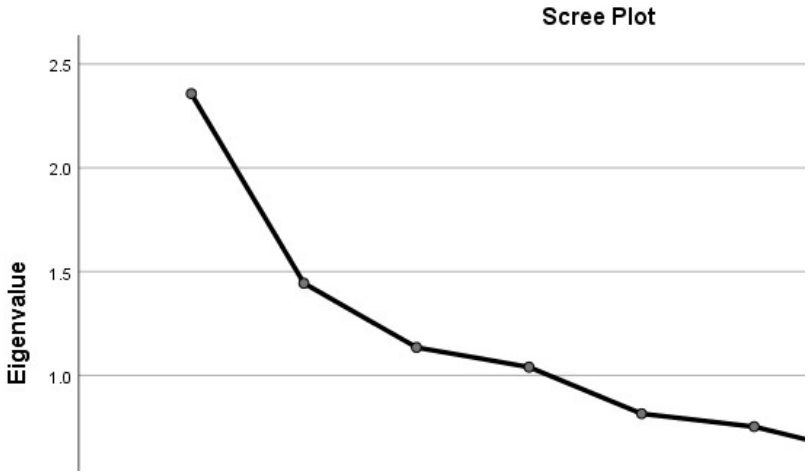


Fig. 7: Scree plot of GDQS chart containing vitamin D rich food groups

Number of component above 1.0 eigenvalue is number of factors to be extracted and interpreted. Four components are clearly visible in the Fig. 7, above 1.0 eigenvalues, so four components has been analysed.

(d) Interpreting KMO and Bartlett's Test table

Table VIII: KMO and Bartlett's test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.601
Bartlett's Test of Sphericity	Approx. Chi-Square	195.201
	df	36
	Sig.	.001

KMO and Bartlett's Test indicate the suitability of the data for structure detection. It shows the proportion of variance in the variables that might caused by the underlying factors. For the better adequacy of data, Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) needs to be at least 0.5 or with values closer to 1.0, so that it can be interpreted. Significance row of the Bartlett's Test of Sphericity shows the p-value. If the $p \leq 0.05$ then null hypotheses is rejected, that it is an identity matrix. If the $p \geq 0.05$, there would be an identity matrix and a principal components analysis should not be conducted.

For the given data the p-value (sig.) was 0.001 which was lesser than 0.05 ($p<0.05$). $p < 0.05$ is considered significant for the given data. Therefore, the values in the nine groups of the GDQS chart were statistically significant.

(e) Interpreting percent of variance

When a PCA is performed it changes the basis of the data and a new dataset with new domain is obtained. If the first two component having total variance percent of 80%, it means it retained less information of the main dataset. The more the percentage of the variance %, the data structure is less informative.

In the food chart of the GDQS the variance percentage of component 1 was 26.191 % and for component 2 was 16.042 % (refers from table VII). It signified that the data retained more information, thus making it more significant for the study. First two components from PCA showed percent of variance as they together explained most of the variability in the original nine variables. So it substituted the 10 original variables to make is simpler and efficient. So it is a good indication for the percent of variance in PCA. Amount of variance each of these factors accounts are given in the table:

Table IX: Percent of variance of the principal component

Principal component	% of Variance
Component 1	26.191
Component 2	16.042
Component 3	12.607
Component 4	11.553

(f) Pattern Matrix

The pattern matrix cannot be detected with this data. It needs more 25 iterations (repetitions) for the rotation of the matrix. So the given data is not sufficient enough for the rotation of the data.

The given variables of the GDQS chart were significant in some steps of the PCA but unable to generate matrices which were very essential for establishing the eigenvalues and scree plots with the data. Certain changes were needed within the data to make the data more accurate for the study.

A study conducted in German adult population to find the habitual intakes with different levels of intake, i.e., daily and meal level. It was observed that higher contribution of the meals to the formation of the PCA-derived habitual dietary patterns when consistency of consumption was low and the intake of foods was substantial in quantity for the respective meal (Schwedhelm *et al*., 2022).

C. SIMPLIFYIG THE DIETARY INTAKE OF VITAMIN D RICH FOODS

The data collected through a questionnaire under the head “Dietary Vitamin D Intake” contained many variables. From those variables only seven variables were selected which were more significant for the study. The statistical method used to simplify the data obtained on the dietary intake of vitamin D rich foods was Principal Component Analysis. This method uses only important factors used for the analysis of the data.

(a) Amount of Variance

For principal component extraction, communalism should always equal to 1.0 for correlation analyses. The accepted cut-off value for communalism is ≥ 0.3. While the ideal value for communalism is ≥ 0.7.

Table X: Communalities of Vitamin D rich foods

Communalities		
	Initial	**Extraction**
Age	1.000	.413
Cereals	1.000	.796
Amount of Cereals	1.000	.661
Rajma	1.000	.514
Amount of Rajma	1.000	.634
Mushroom	1.000	.629
Amount Mushroom	1.000	.673

From the given table X, one extraction was >0.07 values i.e., 0.079 (cereals). It means it was ideal. Remaining fell on the average cut-off values of communalisms i.e., ≥ 0.3.

(b) Checking the data if significant for further analysis

Eigenvalues and Scree plot are essential to find the number of principal components (factors) and it determines if the data is significant for the further analysis. It is necessary for the eigenvalues to be greater than 1.0 for its further analysis and interpretation.

Table XI: Component Matrix (Extracted three principal components)

Component Matrix			
	Component		
	1	2	3
Age	.269	-.312	.493
Cereals	.832	.320	
Amount of Cereals	-.719	-.286	.248
Rajma	.693		.181
Amount of Rajma	-.196	.467	-.614
Mushroom	.380	-.523	-.460
Amount of Mushroom	-.126	.723	.367

Table XII: Total Variance for Vitamin D rich foods

Total Variance Explained							
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	1.961	28.019	28.019	1.961	28.019	28.019	1.899
2	1.296	18.518	46.536	1.296	18.518	46.536	1.295
3	1.062	15.168	61.704	1.062	15.168	61.704	1.223
4	.921	13.161	74.864				
5	.740	10.576	85.441				
6	.660	9.433	94.874				
7	.359	5.126	100.000				
a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.							

From both the table XI and XII it is seen that three principal component has been extracted from the whole dataset of seven variables.

Eigenvalues indicated in the table XII within the column “total”. The eigenvalues for component 1 is 1.961, for component 2 is 1.296 and component 3 is 1.067. Only first three components having eigenvalues >1.0, so the data (Factor 1, 2 and 3) are significant for further analysis and interpretation.

(c) Deciding number of factors to be interpreted

Number of factors to be interpreted depends on the eigenvalues and Scree plot.

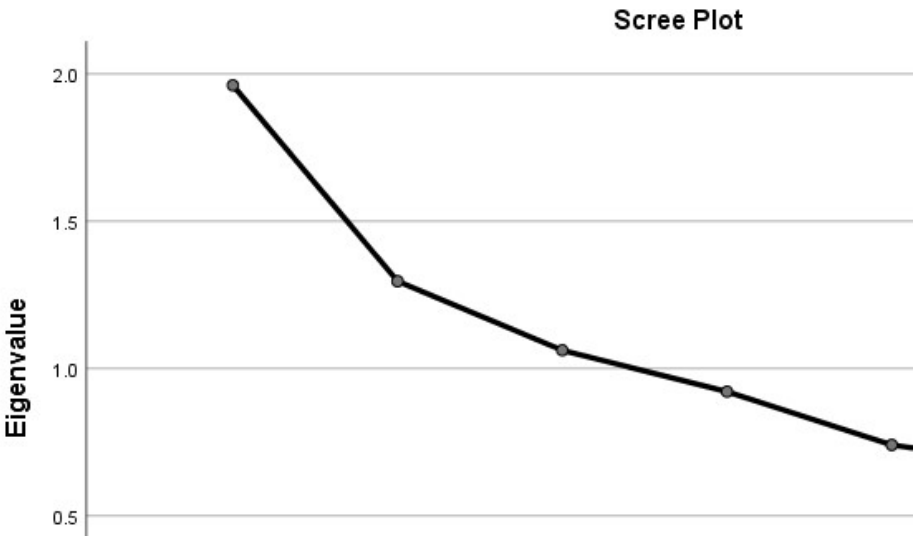


Fig. 8: Scree plot of dietary Vitamin D intake

Number of component above 1.0 eigenvalue is number of factors to be extracted and interpreted. From the fig.8, the number of component above 1.0 eigenvalues is three, so these three factors have been interpreted.

(d) Interpreting KMO and Bartlett's Test table

Table XIII: KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.563
Bartlett's Test of Sphericity	Approx. Chi-Square	120.992
	df	21
	Sig.	.001

KMO and Bartlett's Test indicates the suitability of the data for structure detection. It shows the proportion of variance in the variables due to underlying factors. KMO value over 0.5 and a significance level for the Bartlett's test below p-value <0.05 suggests a substantial correlation of the data.

From the table XIII, the value of KMO was 0.563 and the p-value was 0.001. It suggested that there was a correlational of the data. It also means that PCA could be conducted with the given data.

(e) Interpreting percent of variance

The larger the percent of variance, lesser the information the data holds. From table XII, the percent of variance for component 1 was 28.019, for component 2 were 18.518 and for component 3 were 15.168. It signified that the data retained more information, and making it more significant for the study. First three components from PCA showed percent of variance as they together explained most of the variability in the original seven variables. Hence, the following percent of variance showed good result in PCA.

Table XIV: Percent of variance for dietary vitamin D intake

Principal component	% of Variance
Component 1	28.019
Component 2	18.518
Component 3	15.168

(f) Interpreting through Matrices

There are two main types of matrices for the analysis of PCA viz., Pattern matrix and Structural Matrix. The pattern matrix holds loadings. Pattern matrix is based on the rotation. For oblique rotations, the factors are allowed to correlate (oblimin or promax), then the loadings and correlations are distinct. Each row of the pattern matrix is the regression equation where the standardised observed variable is expressed as functions of the factors. The loadings are regression coefficients. The structure matrix takes the correlations between the variables and the factors.

For interpreting a set of oblique components both pattern and structure matrices are involved. While pattern matrix is mainly used for PCA as it is easy to analyse

Table XV: Pattern Matrix

Pattern Matrix[a]			
	Component		
	1	2	3
Age			.632
Cereals	.892		
Amount of Cereals	-.799		.204
Rajma	.610		.305
Amount of Rajma			-.797
Mushroom	.167	-.765	
Amount of Mushroom	.158	.811	
Rotation Method: Oblimin with Kaiser Normalization.[a]			
a. Rotation converged in 5 iterations.			

From the table XII, in the last column of Rotation Sums of Squared Loadings, only three values are documented, suggestion only three factors have been analysed from the given dataset.

For showing the validity of the data it is essential that values present in one component should not be present in other component (Table XV). It makes the pattern confusing and invalid. From table XV it was noted that Age, Cereals and Amount of Rajma is making it valid for analysis. While the remaining two values are present in both of the component making it invalid.

Few criteria in PCA have been fulfilled with the given data. But for matrix analysis and interpretation, there is requirement to remove the invalid columns from the pattern matrix to make it more accurate and valid for further analysis. It can be interpreted from the PCA analysis is that variables (Age, Cereals and Amount of Rajma) are the important variables for this category as only this three variables can be interpreted. The remaining variables formed confusing matrix for further interpretation, so should be removed from the data.

A study conducted in 10-year-old children studied the dietary data from compared dietary patterns derived from PCA using four strategies for quantifying input variables (gram weights, energy-adjusted weight and percentage energy contribution). It documented that when continuous variables were used

(gram weights, energy-adjusted weight and percentage energy contribution), the first three components extracted had similar loadings and described similar dietary patterns: one contrasting 'more healthy' foods with 'less healthy' foods, one with high loadings on meat, potatoes and vegetables and one with high loadings on lunch and snack foods. The fourth component, present only when intake was measured as percentage energy, was difficult to interpret (Andrew *et al.,* 2013).

D. CORRELATION OF SERUM VITAMIN D AND NUTRIENT INTAKE

Secondary serum vitamin D data has been collected from the previous studies (Habeeba and Kalpana, 2022) and (Gowthami and Kalpana, 2022). That data contained serum vitamin D level and various nutrients. Correlation between serum vitamin D and different nutrients has been performed. These nutrients include protein, fats, calcium, dietary vitamin D_2 and D_3 has taken as the variables for this correlational study.

a. Correlation between serum Vitamin D and dietary fat:

Table XVI: Serum Vitamin D vs Dietary Fat

		Serum Vitamin D	**Dietary Fat**
Serum Vitamin D	Pearson Correlation	1	.331*
	Sig. (2-tailed)		.019
	N	50	50
Dietary Fat	Pearson Correlation	.331*	1
	Sig. (2-tailed)	.019	
	N	50	50
*. Correlation is significant at the 0.05 level (2-tailed).			

From the table XVI it is shown that at significance level $p<0.05$ in two tailed test the significance value is 0.019. The Pearson correlation (R) for the given data is 0.331. Hence, the correlation is significant between Serum Vitamin D and dietary Fat.

In a study it was shown that increment in plasma 25-hydroxyvitamin D was positively associated with monounsaturated fatty acids, and negatively associated with polyunsaturated fatty acids (McLarnon, 2011).

So from the previous papers and this study it can be concluded that dietary fat has significant effects in the absorption of Vitamin D.

b. Correlation between serum Vitamin D and dietary protein

Table XVII: Serum Vitamin D vs Dietary Protein

		Serum Vitamin D	**Dietary Protein**
Serum Vitamin D	Pearson Correlation	1	.228
	Sig. (2-tailed)		.111
	N	50	50
Dietary Protein	Pearson Correlation	.228	1
	Sig. (2-tailed)	.111	
	N	50	50

From the table XVII it can be noted that the two tailed value between serum Vitamin D and dietary protein is 0.111. And the Pearson Correlation coefficient is 0.228. It means there is not significant relationship between serum Vitamin D and Dietary protein. So the correlation is insignificant.

In a study among the young adult population in Brazil, a significant interaction was found between vitamin D and total protein intake (non-animal protein) on 25(OH)D, where individuals consuming a high protein diet (≥73 g/d) (Alathari *et al.*, 2022).

Although for this study an insignificant relationship between vitamin D and dietary protein is shown, few studies documented positive relation between serum Vitamin D and dietary protein.

c. Correlation between Serum Vitamin D and dietary Calcium

Table XVIII: Serum Vitamin D vs Dietary Calcium

		Serum Vitamin D	**Dietary Calcium**
Serum Vitamin D	Pearson Correlation	1	.118
	Sig. (2-tailed)		.414
	N	50	50
Dietary Calcium	Pearson Correlation	.118	1
	Sig. (2-tailed)	.414	
	N	50	50

In the table XVIII it is shown that the significance from result of two-tailed test is 0.414. The Pearson Correlation coefficient (R) is detected from the above table is 0.118. Hence, the data is insignificant to find a correlation between serum Vitamin D and dietary calcium intake.

Threshold level of vitamin D is essential to increase the efficiency of calcium absorption. Inadequate vitamin D absorbs less than or equal to 10% to 15% of dietary calcium. In the vitamin D–sufficient state, the intestinal calcium absorption increases to 30% to 40%. It emphasizes to maintain the calcium:Vitamin D level to prevent dizziness, cardiac arrest, muscoskeletal health (Khazai, 2004).

There is no such relation found which establishes effect of calcium on serum vitamin D level of the body. And this study also failed to find a relation Vitamin D on calcium.

d. Correlation between Serum Vitamin D and dietary Vitamin D_2 (plant origin)

Table XIX: Serum vitamin D vs dietary vitamin D_2

		Serum Vitamin D	**Dietary Vitamin D_2**
Serum Vitamin D	Pearson Correlation	1	.361*
	Sig. (2-tailed)		.031
	N	37	36
Dietary Vitamin D_2	Pearson Correlation	.361*	1
	Sig. (2-tailed)	.031	
	N	36	36
*. Correlation is significant at the 0.05 level (2-tailed).			

From the given table XIX it is noted that the significance of two tailed test at $p<0.05$ is 0.31. The Pearson Correlation coefficient for the given data came out as 0.361. Hence, the correlation between Serum Vitamin D and Dietary Vitamin D_2 is significant.

e. Correlation between serum Vitamin D and dietary Vitamin D_3 (animal source)

Table XX: Serum Vitamin d vs Dietary Vitamin D_3

		Serum Vitamin D	**Dietary Vitamin D_3**
Serum Vitamin D	Serum Vitamin D	1	.357*
	Sig. (2-tailed)		.011
	N	50	50
Dietary Vitamin D_2	Pearson Correlation	.357*	1
	Sig. (2-tailed)	.011	
	N	50	50
*. Correlation is significant at the 0.05 level (2-tailed).			

From the above table it has been found that the significance by two-tailed test is 0.011 at the significance level of $p<0.05$. The Pearson Correlation coefficient is 0.357 for this correlation. Hence, a significant correlation between serum Vitamin D and dietary vitamin D3 can be established.

E. INTERPRETATION OF MACHINE LEARNING

Machine learning (ML) is one of the sub groups of Artificial Intelligence. Three different types of data have been collected as follows:

(i) Secondary of serum Vitamin D along with the nutrients intake of each individuals

(ii) Primary data on Global diet quality score of 150 respondents. 50 variants are present in this data. Those food groups and its amount of intake are taken for the study of ML which is rich in Vitamin D.

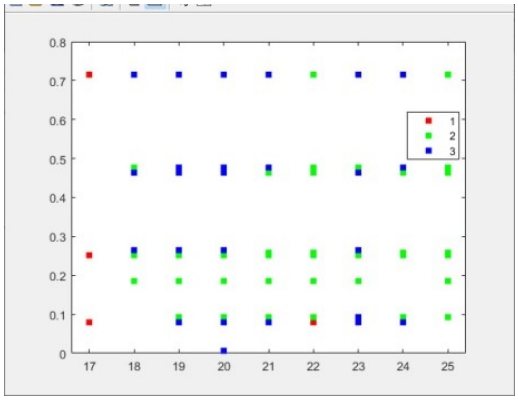
(iii) Vitamin D intake from diet which contains Vitamin D rich foods

For the ML second category of the data from the above (ii) is used to interpret the pattern and categories them. Three category have been made depending upon the consumption pattern i.e., low, moderate (normal) and high.

After coding the consumption pattern of all the food groups (vitamin D rich) and amount of consumption per time in a week (see table IV), a class has been made for this. This class consists of certain numbers which is equivalent to specific consumption pattern in a week. Many repeating values is visible of each category, showing that during the multi-class SVM technique it has changes certain numbers according to its odd.

As all the food groups has different upper limit and lower limit of consumption to show if it is high, low or normal; different threshold values has to be put. All the data has been trained in different way for different food groups.

Following are some of the images of the output (classes) and accuracy.

a. 

b.
```
Class = 1
Accuracy = 62% (31/50) (classific
Class = 2
Accuracy = 58% (29/50) (classific
Class = 3
Accuracy = 80% (40/50) (classific
==========Confusion Matrix==========
     9    10     0
     5    13     3
     2     4     4

**********
Class = Normal
Precision = 56.25
Recall  = 47.3684
FSCore  = 51.4286
Over all Accuracy  = 66
**********
**********
Class = High
Precision = 48.1481
```

Fig. 9 (a): Output of Whole grains (Multi-class SVM technique showing its way of classification)

Fig 9 (b): Output (Accuracy level of Whole grains classes after exploring ML)

Figure 9 (b) shows the accuracy level of whole grains after exploring ML through multi class SVM. It is found that moderate/normal class has accuracy level of 62%. The accuracy level for high class is 56% and for low class the accuracy is 80%. It is essential for the data to be accurate of 90% or more.

a. 

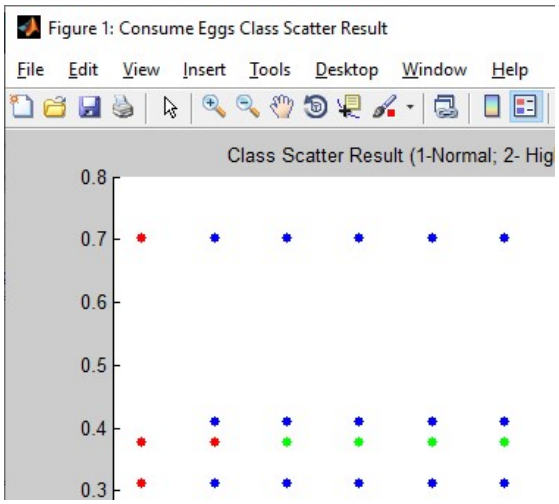
b.
```
========Consume Fish and Shellfish==========
============================================================
Class = 1
Accuracy = 70% (21/30) (classification)
Class = 2
Accuracy = 90% (27/30) (classification)
Class = 3
Accuracy = 60% (18/30) (classification)
========================================
=========Confusion Matrix==========
     4     1     4
     1     2     0
     4     6     8

****************************************
Class = Normal
Precision = 44.4444
Recall  = 44.4444
FSCore  = 44.4444
Over all Accuracy  = 66.6667
****************************************
****************************************
Class = High
Precision = 22.2222
Recall  = 66.6667
FSCore  = 33.3333
Over all Accuracy  = 73.3333
****************************************
****************************************
Class = Low
Precision = 66.6667
Recall  = 44.4444
FSCore  = 53.3333
Over all Accuracy  = 53.3333
****************************************
```

Fig. 10 (a): Output of Fish and shellfish (Multi-class SVM technique showing its way of classification)

Fig 10 (b): Output (Accuracy level of Fish and shellfish classes after exploring ML)

Figure 10 (b) shows the accuracy level of fish and shell fish after exploring ML through multi class SVM. It is found that moderate/normal class has 66.6% accuracy. The accuracy level for high class is 73% and for low class the accuracy is 53%.

a.



b.

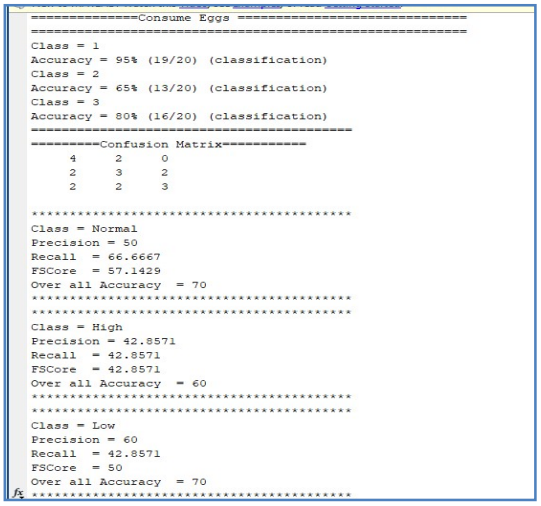


Fig. 11 (a): Output of Egg (Multi-class SVM technique showing its way of classification)

Fig 11 (b): Output (Accuracy level of Eggs classes after exploring ML)

Figure 11 (b) shows the accuracy level of eggs after exploring ML through multi class SVM. It is found that moderate/normal class has 70% accuracy. The accuracy level for high class is 60% and for low class the accuracy is 70%.

a.

b.

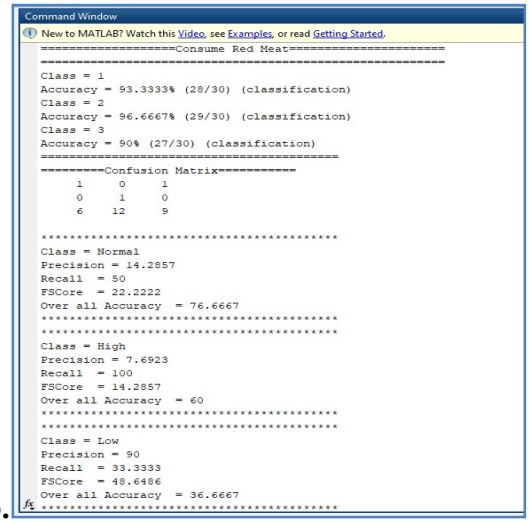


Fig. 12 (a): Output of meat (Multi-class SVM technique showing its way of classification)

Fig 12 (b): Output (Accuracy level of meat classes after exploring ML)

Figure 12 (b) shows the accuracy level of eggs after exploring ML through multi class SVM. It is found that moderate/normal class has 76% accuracy. The accuracy level for high class is 60% and for low class the accuracy is 36%.

For a machine learning technique to be efficient the accuracy level should be 90%. Approximately accuracy level is found to be only 60% in this case (GDQS with Vitamin D) which is far from 90% accuracy. Although the accuracy level is low with the real time databases, still it can be used for predicting the future data to be used in this ML data. It can predict or classify the data to make understand if the intake of vitamin D from the given food groups is low, normal or high for the humans.

SUMMARY AND CONCLUSION

V. SUMMARY AND CONCLUSION

Vitamin D deficiency is one of the hidden nutritional deficiencies found widespread among individuals irrespective of age. In the Indian Subcontinent the situation is more severe due to demography, cultural variations and diet as well. From the old times to recent times various techniques have been devised which measure serum either of the Vitamin D metabolites (1,25 dihydroxy Vitamin D_3 or 25-hydroxy vitamin D). Both of these have their own advantages and disadvantages. In recent times using the metabolite, 25-hydroxy Vitamin D, for screen serum Vitamin D level to detect the Vitamin D status of an individual is considered more effective and accurate than using 1, 25-dihydroxy Vitamin D_3.

Assaying Vitamin D in clinical settings costs 800 to 2000 INR. The Indian population mostly consists of low to moderate economic background individuals. Expending that amount of money will be burdensome for each of the people. Keeping that in mind, this study entitled "Exploration of Machine Learning to develop a Low-cost Screening Method with Global diet quality score to detect Vitamin D deficiency" develops an economical technique to assay the Intake of Vitamin D through diet and mentions its classification if it is low, moderate or high.

From the Global diet quality score (GDQS) chart only those food groups have been taken which contain moderate to high amounts of vitamin D and those groups who have positive influence in Vitamin D absorption (fats and protein). This includes five groups viz., whole grains, poultry, eggs, meat and fish and shellfish.

Salient findings of the study:

(i) Most of the people belong to the age group of 19 years constituting 25 percent in 150 respondents. Second largest categories were 18 years old, forming 17 percent.

(ii) From the secondary data of serum level of Vitamin D it was found that 36 individuals out of 50 (72 percent) individuals have total serum Vitamin D [25(OH)D] less than 10 ng/dL (<10ng/dL). It shows that approx three-fourth of the study participants were severely Vitamin D deficient. 22 percent individuals were moderately deficient of vitamin D. Their serum Vitamin D is >10 ng/dL and <20ng/dL. Only 3 respondents (6 percent) had sufficient levels of serum Vitamin D. This shows that most of the individuals are deficient in Vitamin D in a given population of a specific area.

(iii) When Principal Component Analysis (PCA) was performed on the Global diet quality score (GDQS) chart containing Vitamin D rich foods it was found four principal components after

reducing the complex data to simpler one. It also showed eigenvalues (more than 1.0) and scree plot also formed. The total data was also significant at the level $p < 0.001$. The percent of variance was also less, suggesting it retained less information from the original data, which is a good indication for the dataset. But the matrix formation, which is essential part of the PCA, did not form because of lack of iterations (repetitions) of data. To get accurate results, some changes have to be done in the original dataset. For this study any changes cannot be performed.

(iv) PCA was also performed on the dietary intake of Vitamin D rich foods. In the GDQS a whole food group which consists of Vitamin D rich foods was considered while in this, only specific foods have been taken into consideration. While doing PCA for the dataset under this it found three principal components which hold eigenvalues >1.0 and formed the scree plot. In KMO and Bartlett's test the significance level which it found was 0.001. Even the percent of variance showed low variance, means less information from the raw data has been retained, which is a good indication for the further analysis. With the given data it interpreted pattern matrices too making the data significant for the study. In the pattern matrix only the three variants are most significant while the remaining four variants made confusion matrix or invalid. These three groups include: (a) age (b) cereals (c) amount of Rajma. So these three categories are most significant.

(v) A significant and positive relation has been found between serum Vitamin D and dietary fat. Dietary fat intake has a positive influence in serum vitamin D levels.

(vi) An insignificant difference has been established in this study between dietary protein and vitamin D. So intake of protein has no influence in the serum vitamin D level.

(vii) It was difficult to find a significant relationship between serum Vitamin D levels and dietary calcium

(viii) A positive correlation has been found between serum vitamin D and dietary Vitamin D_2 and D_3. It means intake of vitamin D through diet can affect positively on the serum vitamin D level.

From this study major finding has been identified:

(i) First model to perform Mac

(ii)

(iii) hine Learning with GDQS food groups especially for Vitamin D rich food groups.

(iv) From the vitamin D intake foods three variables were found valid (according to statistics) age, cereals and amount of rajma

(v) Using the secondary data of serum vitamin D and nutrient intake, a positive influence has been drawn from dietary fat, dietary Vitamin D_2 and Vitamin D_3.

(vi) Developing a machine learning to detect the low, moderate and high intake of dietary vitamin D to show if dietary intake is sufficient for maintaining serum vitamin D level. Basic Machine Learning is suitable for small range of data. They have been trained and untrained data can be used for future prediction related to dietary Vitamin D consumption and serum Vitamin D.

Recommendations for future study

(i) Large population size is important to form more accurate machine learning for detecting vitamin D adequacy through diet

(ii) Usage of GDQS for other nutritional deficiencies and disorders.

(iii) Prediction of Vitamin D deficiency with the help of trained dataset using GDQS.

(iv) Taking large sample size for the study approximately 1000 to 10,000 can give more accurate result and lead to deep learning too.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adamec, J., Jannasch, A., Huang, J., Hohman, E., Fleet, J. C., Peacock, M., Ferruzzi, M. G., Martin, B., & Weaver, C. M. (2011). Development and optimization of an LC-MS/MS-based method for simultaneous quantification of vitamin D2 , vitamin D3 , 25-hydroxyvitamin D2 and 25-hydroxyvitamin D3. *Journal of separation science*, *34*(1), 11–20. https://doi.org/10.1002/jssc.201000410

Alathari, B. E., Cruvinel, N. T., da Silva, N. R., Chandrabose, M., Lovegrove, J. A., Horst, M. A., & Vimaleswaran, K. S. (2022). Impact of Genetic Risk Score and Dietary Protein Intake on Vitamin D Status in Young Adults from Brazil. *Nutrients*, *14*(5), 1015. https://doi.org/10.3390/nu14051015

Al-Alyani, H., Al-Turki, H. A., Al-Essa, O. N., Alani, F. M., & Sadat-Ali, M. (2018). Vitamin D deficiency in Saudi Arabians: A reality or simply hype: A meta-analysis (2008-2015). *Journal of family & community medicine*, *25*(1), 1–4. https://doi.org/10.4103/jfcm.JFCM_73_17

Al-Daghri, N. M., Hussain, S. D., Ansari, M., Khattak, M., Aljohani, N., Al-Saleh, Y., Al-Harbi, M. Y., Sabico, S., & Alokail, M. S. (2021). Decreasing prevalence of vitamin D deficiency in the central region of Saudi Arabia (2008-2017). *The Journal of steroid biochemistry and molecular biology*, *212*, 105920. https://doi.org/10.1016/j.jsbmb.2021.105920

Al-Daghri N. M. (2018). Vitamin D in Saudi Arabia: Prevalence,distribution and disease associations. *The Journal of steroid biochemistry and molecular biology*, *175*, 102–107. https://doi.org/10.1016/j.jsbmb.2016.12.017

Al-Mogbel E. S. (2012). Vitamin D status among Adult Saudi Females visiting Primary Health Care Clinics. *International journal of health sciences*, *6*(2), 116–126. https://doi.org/10.12816/0005987

Alpaydin E. Introduction to machine learning. 3rd ed. Cambridge, MA: The MIT Press; 2014.

Amrien K et al, (2020), Vitamin D deficiency 2.0: an update on the current status worldwide, European Journal of Clinical Nutrition, https://doi.org/10.1038/s41430-020-0558-y

Angulo, E., Stern, D., Castellanos-Gutiérrez, A., Monge, A., Lajous, M., Bromage, S., Fung, T.T., Li, Y., Bhupathiraju, S.N., Deitchler, M., Willett, W., & Batis, C. (2021). Changes in the Global Diet Quality Score, Weight, and Waist Circumference in Mexican Women. *The Journal of Nutrition, 151*, 152S - 161S.

Asakura, K., Etoh, N., Imamura, H., Michikawa, T., Nakamura, T., Takeda, Y., Mori, S., & Nishiwaki, Y. (2020). Vitamin D Status in Japanese Adults: Relationship of Serum 25-Hydroxyvitamin D with Simultaneously Measured Dietary Vitamin D Intake and Ultraviolet Ray Exposure. *Nutrients*, *12*(3), 743. https://doi.org/10.3390/nu12030743

Babu, U. S., & Calvo, M. S. (2010). Modern India and the vitamin D dilemma: evidence for the need of a national food fortification program. *Molecular nutrition & food research*, *54*(8), 1134–1147. https://doi.org/10.1002/mnfr.200900480

Basu S., Gupta R., Mitra M., & Ghosh A. (2015). Prevalence of vitamin d deficiency in a pediatric hospital of eastern India. *Indian journal of clinical biochemistry : IJCB*, *30*(2), 167–173. https://doi.org/10.1007/s12291-014-0428-2

Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

Borkar, V. V., Devidayal, Verma, S., & Bhalla, A. K. (2010). Low levels of vitamin D in North Indian children with newly diagnosed type 1 diabetes. *Pediatric diabetes*, *11*(5), 345–350. https://doi.org/10.1111/j.1399-5448.2009.00589.x

Bromage S, Batis C, Bhupathiraju SN, et al. (2021). Development and Validation of a Novel Food-Based Global Diet Quality Score (GDQS). *J Nutr*.;151(12 Suppl 2):75S-92S. doi:10.1093/jn/nxab244

Bromage S, Andersen C.T, Tadesse A.W, Passarelli S, Hemler E.C, Fekadu H, Sudfeld C.R., Worku A., Berhane H, Batis C, Bhupathiraju S.N, Fung T.T, Li Y, Stampfer M.J, Deitchler M, Willett W.C, Fawzi W.W (2021). The Global Diet Quality Score is Associated with Higher Nutrient Adequacy, Midupper Arm Circumference, Venous Hemoglobin, and Serum Folate Among Urban and Rural Ethiopian Adults. *The Journal of Nutrition*, 151(2), 130S–142S, https://doi.org/10.1093/jn/nxab264

Bromage S, Andersen CT, Tadesse AW, et al. (2021). The Global Diet Quality Score is Associated with Higher Nutrient Adequacy, Midupper Arm Circumference, Venous Hemoglobin, and Serum Folate Among Urban and Rural Ethiopian Adults. *J Nutr*.;151(2), 130S-142S. doi:10.1093/jn/nxab264

Caixinha, M., & Nunes, S. (2017). Machine Learning Techniques in Clinical Vision Sciences. Current eye research, 42(1), 1–15. https://doi.org/10.1080/02713683.2016.1175019

Cashman, K. D., Dowling, K. G., Škrabáková, Z., Gonzalez-Gross, M., Valtueña, J., De Henauw, S., Moreno, L., Damsgaard, C. T., Michaelsen, K. F., Mølgaard, C., Jorde, R., Grimnes, G., Moschonis, G., Mavrogianni, C., Manios, Y., Thamm, M., Mensink, G. B., Rabenberg, M., Busch, M. A., Cox, L., Kiely, M. (2016). Vitamin D deficiency in Europe: pandemic?. *The American journal of clinical nutrition*, *103*(4), 1033–1044. https://doi.org/10.3945/ajcn.115.120873

Cashman K. D. (2020). Vitamin D Deficiency: Defining, Prevalence, Causes, and Strategies of Addressing. *Calcified tissue international*, *106*(1), 14–29. https://doi.org/10.1007/s00223-019-00559-4

Chowdhury, R., Taneja, S., Bhandari, N., Sinha, B., Upadhyay, R. P., Bhan, M. K., & Strand, T. A. (2017). Vitamin-D deficiency predicts infections in young north Indian children: A secondary data analysis. *PloS one*, *12*(3), e0170509. https://doi.org/10.1371/journal.pone.0170509

Chun, R. F., Lauridsen, A. L., Suon, L., Zella, L. A., Pike, J. W., Modlin, R. L., Martineau, A. R., Wilkinson, R. J., Adams, J., & Hewison, M. (2010). Vitamin D-binding protein directs monocyte responses to 25-hydroxy- and 1,25-dihydroxyvitamin D. *The Journal of clinical endocrinology and metabolism*, *95*(7), 3368–3376. https://doi.org/10.1210/jc.2010-0195

Courbebaisse M, Thervet E, Souberbielle JC, Zuber J, Eladari D, Martinez F, Mamzer-Bruneel MF, Urena P, Legendre C, Friedlander G, Prié D. (2009) Effects of vitamin D supplementation on the calcium-phosphate balance in renal transplant patients. *Kidney Int.*;75(6), 646-51. doi: 10.1038/ki.2008.549.

Deo R.C (2015), Machine Learning in Medicine, Basic Science for Clinicians, 132:1920–1930, DOI: 10.1161/CIRCULATIONAHA.115.001593

Duarte C, Carvalheiro H, Rodrigues AM, Dias SS, Marques A, Santiago T, Canhão H, Branco JC, da Silva JAP. (2020). Prevalence of vitamin D deficiency and its predictors in the Portuguese population: a nationwide population-based study. *Arch Osteoporos.*;15(1), 36. doi: 10.1007/s11657-020-0695-x.

Duarte, C., Carvalheiro, H., Rodrigues, A. M., Dias, S. S., Marques, A., Santiago, T., Canhão, H., Branco, J. C., & da Silva, J. (2020). Prevalence of vitamin D deficiency and its predictors in the

Portuguese population: a nationwide population-based study. *Archives of osteoporosis*, *15*(1), 36. https://doi.org/10.1007/s11657-020-0695-x

El Naqa, Issam & Murphy, Martin. (2015). What Is Machine Learning?. 10.1007/978-3-319-18305-3_1.

Farrell, C. J., Martin, S., McWhinney, B., Straub, I., Williams, P., & Herrmann, M. (2012). State-of-the-art vitamin D assays: a comparison of automated immunoassays with liquid chromatography-tandem mass spectrometry methods. *Clinical chemistry*, *58*(3), 531–542. https://doi.org/10.1373/clinchem.2011.172155

Geert Meyfroidt, Fabian Güiza, Jan Ramon, Maurice Bruynooghe, (2009) Machine learning techniques to examine large patient databases, Best Practice & Research Clinical Anaesthesiology, 23(1), 127-143, https://doi.org/10.1016/j.bpa.2008.09.003.

Global Burden of Disease, 2017

Gowthami.K and Kalpana C.A., May 2022 : " Obesity and Vitamin D insufficiency among young women – Prevalence, Association and mHealth intervention"

Habeeba.B and Kalpana C.A., January 2022, " Impact of Nutrition Intervention and Dress Code on Vitamin D nutriture of Muslim Women"

Harinarayan, C. V., Ramalakshmi, T., Prasad, U. V., Sudhakar, D., Srinivasarao, P. V., Sarma, K. V., & Kumar, E. G. (2007). High prevalence of low dietary calcium, high phytate consumption, and vitamin D deficiency in healthy south Indians. *The American journal of clinical nutrition*, *85*(4), 1062–1067. https://doi.org/10.1093/ajcn/85.4.1062

Harvard T.H. Chan (Mexico) School of Public Health, Department of Nutrition, Intake, 2021

Hazra A., Mandal S., Gupta A., Mukherjee A & Mukherjee A. (2017). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology. 10. 2137-2159.

Holick M. F. (2009). Vitamin D status: measurement, interpretation, and clinical application. *Annals of epidemiology*, *19*(2), 73–78. https://doi.org/10.1016/j.annepidem.2007.12.001

Holick M. F. (2003). Vitamin D: A millenium perspective. *Journal of cellular biochemistry*, *88*(2), 296–307. https://doi.org/10.1002/jcb.10338

Hollis B. W. (2000). Comparison of commercially available (125)I-based RIA methods for the determination of circulating 25-hydroxyvitamin D. *Clinical chemistry*, *46*(10), 1657–1661.

Holick MF, Binkley NC, Bischoff-Ferrari HA, Gordon CM, Hanley DA, Heaney RP, et al. (2011). Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab*. 96:1911–30. https://doi.org/10.1210/jc.2011-0385

Holick M.F., (2009), Vitamin D Status: Measurement, Interpretation, and Clinical Application, Elsevier, Vol 19, No. 2, 2-4

Hymøller, L., & Jensen, S. K. (2011). Vitamin D analysis in plasma by high performance liquid chromatography (HPLC) with C(30) reversed phase column and UV detection--easy and acetonitrile-free. *Journal of chromatography. A*, *1218*(14), 1835–1841. https://doi.org/10.1016/j.chroma.2011.02.004

Ifrah G. The universal history of computing: from the abacus to the quantum computer. New York: John Wiley; 2001.

Institute of Medicine, 2017

Jafri, L., Khan, A. H., Siddiqui, A. A., Mushtaq, S., Iqbal, R., Ghani, F., & Siddiqui, I. (2011). Comparison of high performance liquid chromatography, radio immunoassay and electrochemiluminescence immunoassay for quantification of serum 25 hydroxy vitamin D. *Clinical biochemistry*, *44*(10-11), 864–868. https://doi.org/10.1016/j.clinbiochem.2011.04.020

Jenkinson, C., Taylor, A. E., Hassan-Smith, Z. K., Adams, J. S., Stewart, P. M., Hewison, M., & Keevil, B. G. (2016). High throughput LC-MS/MS method for the simultaneous analysis of multiple vitamin D analytes in serum. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, *1014*, 56–63. https://doi.org/10.1016/j.jchromb.2016.01.049

Kadam, N. S., Chiplonkar, S. A., Khadilkar, A. V., Fischer, P. R., Hanumante, N. M., & Khadilkar, V. V. (2011). Modifiable factors associated with low bone mineral content in underprivileged premenarchal Indian girls. *Journal of pediatric endocrinology & metabolism : JPEM*, *24*(11-12), 975–981. https://doi.org/10.1515/jpem.2011.405

Kapil, U., Pandey, R. M., Goswami, R., Sharma, B., Sharma, N., Ramakrishnan, L., Singh, G., Sareen, N., Sati, H. C., Gupta, A., & Sofi, N. Y. (2017). Prevalence of Vitamin D deficiency and associated

risk factors among children residing at high altitude in Shimla district, Himachal Pradesh, India. *Indian journal of endocrinology and metabolism*, *21*(1), 178–183. https://doi.org/10.4103/2230-8210.196031

Khazai, N., Judd, S. E., & Tangpricha, V. (2008). Calcium and vitamin D: skeletal and extraskeletal health. *Current rheumatology reports*, *10*(2), 110–117. https://doi.org/10.1007/s11926-008-0020-y

Kumar, P., Shenoi, A., Kumar, R. K., Girish, S. V., & Subbaiah, S. (2015). Vitamin D Deficiency Among Women in Labor and Cord Blood of Newborns. *Indian pediatrics*, *52*(6), 530–531.

MaCarthy J, Minsky M. Rochester N., and Shannon C.E, (2006), A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, *AI Magazine*, 27(4), https://doi.org/10.1609/aimag.v27i4.1904

McLarnon A, (2011), Dietary fat might influence serum vitamin D level, *Nature Reviews Endocrinology* . 7 (562), https://doi.org/10.1038/nrendo.2011.150

Mika Matsuzaki, Sabri Bromage, Carolina Batis, Teresa Fung, Yanping Li, Megan Deitchler, Meir Stampfer, Walter Willett, Sanjay Kinra, Shilpa Bhupathiraju, (2020). Validation of a New Instrument for Assessing Diet Quality and Its Association with Undernutrition and Non-Communicable Diseases for Women in Reproductive Age in India, *Current Developments in Nutrition*. 4(2), 1451, https://doi.org/10.1093/cdn/nzaa061_079

Murphy K .P, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

Niramitmahapanya, S., Harris, S. S., & Dawson-Hughes, B. (2011). Type of dietary fat is associated with the 25-hydroxyvitamin D3 increment in response to vitamin D supplementation. *The Journal of clinical endocrinology and metabolism*, *96*(10), 3170–3174. https://doi.org/10.1210/jc.2011-1518

Park, J. H., Hong, I. Y., Chung, J. W., & Choi, H. S. (2018). Vitamin D status in South Korean population: Seven-year trend from the KNHANES. *Medicine*, *97*(26), e11032. https://doi.org/10.1097/MD.0000000000011032

Patel, J. V., Chackathayil, J., Hughes, E. A., Webster, C., Lip, G. Y., & Gill, P. S. (2013). Vitamin D deficiency amongst minority ethnic groups in the UK: a cross sectional study. *International journal of cardiology*, *167*(5), 2172–2176. https://doi.org/10.1016/j.ijcard.2012.05.081

Powe, C. E., Evans, M. K., Wenger, J., Zonderman, A. B., Berg, A. H., Nalls, M., Tamez, H., Zhang, D., Bhan, I., Karumanchi, S. A., Powe, N. R., & Thadhani, R. (2013). Vitamin D-binding protein and vitamin D status of black Americans and white Americans. *The New England journal of medicine*, *369*(21), 1991–2000. https://doi.org/10.1056/NEJMoa1306357

Ritu G., & Gupta A. (2014). Vitamin D deficiency in India: prevalence, causalities and interventions. *Nutrients*, *6*(2), 729–775. https://doi.org/10.3390/nu6020729

Rola, Rafał & Kowalski, Konrad & Bieńkowski, Tomasz & Studzińska, Sylwia. (2020). Improved sample preparation method for fast LC-MS/MS analysis of vitamin D metabolites in serum. Journal of Pharmaceutical and Biomedical Analysis. 190. 113529. 10.1016/j.jpba.2020.113529.

Sak, J., & Suchodolska, M. (2021). Artificial Intelligence in Nutrients Science Research: A Review. Nutrients, 13(2), 322. https://doi.org/10.3390/nu13020322

Sambasivam G., Amudhval J. and Sathya G (2020), A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods, IEEE Access, Vol 8, 16

Santos, M. J., Fernandes, V., & Garcia, F. M. (2015). Carência de Vitamina D numa População Hospitalar: Uma Fotografia pela Perspetiva Laboratorial [Vitamin D Insufficiency in a Hospital Population: A Photograph from the Laboratory Perspective]. *Acta medica portuguesa*, *28*(6), 726–734.

Satoh M., Ishige T., Ogawa S., Nishimura M., Matsushita K., Higashi T. & Nomura F. (2016). Development and validation of the simultaneous measurement of four Vitamin D metabolites in serum by LC-MS/MS for clinical laboratory applications. *Analytical and bioanalytical chemistry,* 408(27), 7617-7627.

Schwedhelm, C., Iqbal, K., Knüppel, S., Schwingshackl, L., & Boeing, H. (2018). Contribution to the understanding of how principal component analysis-derived dietary patterns emerge from habitual data on food consumption. *The American journal of clinical nutrition*, *107*(2), 227–235. https://doi.org/10.1093/ajcn/nqx027

Shailaja, K. & Seetharamulu, B. & Jabbar, M.. (2018). Machine Learning in Healthcare: A Review. 910-914. 10.1109/ICECA.2018.8474918.

Strand, M. A., Perry, J., Zhao, J., Fischer, P. R., Yang, J., & Li, S. (2009). Severe vitamin D-deficiency and the health of North China children. *Maternal and child health journal*, *13*(1), 144–150. https://doi.org/10.1007/s10995-007-0250-z

Turpeinen, U., Hohenthal, U., & Stenman, U. H. (2003). Determination of 25-hydroxyvitamin D in serum by HPLC and immunoassay. *Clinical chemistry*, *49*(9), 1521–1524. https://doi.org/10.1373/49.9.1521

Yuna He, Yuehui Fang, Sabri Bromage, Teresa T Fung, Shilpa N Bhupathiraju, Carolina Batis, Megan Deitchler, Wafaie Fawzi, Meir J Stampfer, Frank B Hu, Walter C Willett, Yanping Li, (2021), Application of the Global Diet Quality Score in Chinese Adults to Evaluate the Double Burden of Nutrient Inadequacy and Metabolic Syndrome, *The Journal of Nutrition*, 151(2), 93S–100S, https://doi.org/10.1093/jn/nxab162

Yu, S., Fang, H., Han, J., Cheng, X., Xia, L., Li, S., Liu, M., Tao, Z., Wang, L., Hou, L., Qin, X., Li, P., Zhang, R., Su, W., & Qiu, L. (2015). The high prevalence of hypovitaminosis D in China: a multicenter vitamin D status survey. *Medicine*, *94*(8), e585. https://doi.org/10.1097/MD.0000000000000585

APPENDIX

APPENDIX - I

QUESTIONNAIRE FOR (Global diet quality score) GDQS

GDQS QUESTIONS

1. **Frequency of consumption of Lemon juice/ sweet lime pulp/ orange pulp/other citrus fruits**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

2. **Amount of its consumption per day/at each time?**

a) <24 g/day b) 24-69 b/day c) >69 g/day

3. **How often you eat the following – Apricot/ Papaya/ ripened mango/ musk melon (orange flesh)**?

a) Everyday b) 2-4 days per week c) 5-6 days per week d) Seasonally

4. **Amount of its consumption per day/at each time?**

a) <25 g/day b) 25-123 g/day c) >123 g/day

5. **How often do you take these – Apples/guava/pineapple/dates/sapota**?

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Seasonally

6. **Amount of its consumption per day/at each time?**

a) <27 g/day b) 27-107g/day c) >107g/day

7. **How often do you take Dark green leafy vegetables (spinach/fenugreek/agathi leaves/drumstick leaves/amaranth leaves/raddish leaves/mustard leaves)**?

a) Everyday b) Once in a week c) 2-4 days per week

d) 5-6 days per week e) Rarely

8. **Amount of its consumption per day/at each time?**

a) <13g/day b) 14-37g/day c) >37g/day

9. **How frequent do you consume cruciferous vegetable – Cabbage/ broccoli/Brussels sprouts/ cauliflower/ collard greens/ kale/ kohlrabi/turnip/ rape/ brown mustard**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

10. **Amount of its consumption per day/at each time?**

a) <13g/day b) 13-36g/day c) >36g/day

11. **How often do you consume deep orange vegetables – orange pumpkin/orange squash** ?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

12. **Amount of its consumption per day/at each time?**

a) <23g/day b) 23-114g/day c) >114g/day

13. **How often do you consume other vegetables- Field beans, French beans, onion stalk, bamboo shoot, bitter gourd, bottle gourd, ash gourd, brinjal, capsicum, cho-cho,cucumber, drumstick, raw jack fruit, ladies finger, raw mango, raw, peas, plaintain green, plaintain fruit, ridge gourd, snake gourd, zucchini,beet roots, lotus root, beans sprouts**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Few days in a month

14. **Amount of consumption per day/at each time?**

a) <23 g/day b) 23-114 g/day c) >114 g/ day

15. **How often do you consume legumes**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

16. **Amount of its consumption per day/at each time?**

a) <9g/day b) 9-42g/day c) >42g/day

17. **How often do you consume deep orange tubers – yam/sweet potato?**

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

18. **Amount of its consumption per day/at each time?**

a) <12g/day b) 12-63g/day c) >63g/day

19. **How often do you consume nuts and seeds?**

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Not regular, sometimes

20. **Amount of its consumption per day/at each time?**

a) <7g/day b) 7-13g/day c) >13g/day

21. **How often do you consume Whole grains**?

a) Everyday b) 2-4 days per week c) 5-6 days per week

22. **Amount of its consumption per day/at each time?**

a) <8 gday b) 8-13 g/day c) >13 g/day

23. **How often do you consume liquid oils?**

a) Everyday b) 2-4 days per week c) 5-6 days per week

24. **Amount of its consumption per day/at each time?**

a) <2 g/day b) 2-7.5 g/day c) >7.5g/day

25. **How often do you consume fish and shellfish?**

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Not at all

26. **Amount of its consumption per day/at each time?**

a) <14 g/day b) 14-71 g/day c) >71 g/day

27. **How often do you consume chicken/duck/pigeon**?

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Not at all

28. **Amount of its consumption per day/at each time?**

a) <16 g/day b) 16-44 g/day c) >44 g/day

29. **How often do you consume low fat dairy – skim milk/double toned milk/toned milk/skim milk products?**

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

30. **Amount of its consumption per day/at each time?**

a) <33g/day b) 33-132 g/day c) >132 g/day

31. **How often do you consume Eggs?**

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month d) Not at all

32. **Amount of its consumption per day/at each time**?

a) <6 g/day b) 6-32 g/day c) >32 g/day

33. **How often do you consume high fat dairy or dairy products - buffalo milk/ghee/butter/cheese/condensed milk?**

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Not at all

34. **Amount of its consumption per day/at each time?**

a) <35 g/day b) 35-142 g/day

c) 142-734 g/day d) >734 g/day

35. **How often do you consume red meat?**

a) Everyday b) 2-4 days per week c) Once in a week

d) Once or twice in a month e) Not at all

36. **Amount of its consumption per day/at each time?**

a) <9 g/day b) 9-46 g/day c) >46/day

37. **How often do you consume hot dogs/canned meat/sausages/salami?**

a) 2-4 days per week b) 5-6 days per week

c) Once or twice in a month d) Rarely e) Not at all

38. **Amount of its consumption per day/at each time?**

a) <9 g/day b) 9 -30 g/day c) >30 g/day

39. **How often do you consume refined grains and baked goods - Maida/white bread/semolina/refined wheat biscuits/cakes?**

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

40. **Amount of its consumption per day/at each time?**

a) <7 g/day b) 7-33 g/day c) >33 g/day

41. **How often do you consume Sweet/ ice-cream**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Few days in a month e) Rarely

42. **Amount of its consumption per day/at each time?**

a) <13 g/day b) 13-37 g/day c) >37 g/day

43. **How often do you consume sugar sweetened beverages/soft drinks/squashes**?

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Rarely

44. **Amount of its consumption per day/at each time?**

a) 57 ml/day b) 57-180 ml/day c) >180 g/day

45. **How often do you consume fruit and vegetable juice**?

a) Everyday b) 2-4 days per week c) 5-6 days per week

d) Once or twice in a month e) Not at all

46. **Amount of its consumption per day/at each time?**

a) <36 g/day b) 36-144 g/day c) >144 g/day

47. **How often do you consume white roots and tubers – raddish, potato, cassava,tapoica**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

48. **Amount of its consumption per day/at each time?**

a) <27 g/day b) 27-107 g/day c) >107 g/day

49. **How often do you consume purchased deep fried foods-chips, pakodas**?

a) Everyday b) 2-4 days per week

c) 5-6 days per week d) Once or twice in a month

50. **Amount of its consumption per day/at each time?**

a) <9 g/day b) 9-45 g/day c) >45 g/day

APPENDIX - II

QUESTIONNAIRE FOR VITAMIN D RICH FOODS

VITAMIN D DIETARY INTAKE

1. How frequent do you eat: Amaranth/maize/ragi?
    a. Once in a week
    b. 2-3 days in a week
    c. 4-6 days in a week
    d. Everyday
    e. Once or twice in a month
    f. Not at all
2. Amount of its consumption per day/at each time (in grams)?
3. How often do you consume rajma?
    a. Once in a week
    b. 2-3 days in a week
    c. 4-6 days in a week
    d. Everyday
    e. Rarely
    f. Not at all
4. Amount of its consumption per day/at each time (in grams)?
5. How often do you consume soyabean?
    a. Once in a week
    b. 2-3 days in a week
    c. 4-6 days in a week
    d. Everyday
    e. Once or twice in a month
    f. Rarely
    g. Not at all
6. Amount of its consumption per day/at each time (in grams)?
7. Frequency of consumption gingelly/sesame seeds:
    a. Once in a week

b. 2-3 days in a week
c. 4-6 days in a week
d. Everyday
e. Once or twice in a month
f. Not at all

8. Amount of its consumption per day/at each time (in grams)?
9. Frequency of consumption of mustard seeds and walnuts seeds?
    a. Once in a week
    b. 2-3 days in a week
    c. 4-6 days in a week
    d. Everyday
    e. Once or twice in a month
    f. Rarely
    g. Not at all
10. Amount of its consumption per day/at each time (in grams)?
11. How frequent do you consume mushroom?
    a. Once in a week
    b. 2-3 days in a week
    c. 4-6 days in a week
    d. Everyday
    e. Once or twice in a month
    f. Rarely
    g. Not at all
12. Amount of its consumption per day/at each time (in grams)?

APPENDIX III

ETHICAL CLEARANCE CERTIFICATE

INSTITUTIONAL HUMAN ETHICS COMMITTEE

Avinashilingam

Institute for Home Science and Higher Education for Women
(Deemed to be University under Category 'A' by MHRD, Estd. u/s 3 Of UGC Act 1956) RE-Accredited with 'A++' Grade by NAAC.
Recognised by UGC Under Section 12 B
Coimbatore – 641 043, Tamil Nadu

26th February 2022

To
Ms.Tuhina Patra
Department of Food Science and Nutrition
Avinashilingam Institute for Home Science and
Higher Education for Women
Coimbatore – 641 043

Dear Tuhina Patra,

Ref: Your proposal No. IHEC/21-22/FSN-27 entitled "Exploration of Machine Learning in Development of Low-Cost Screening Method Featuring the Global Diet Quality Score for Detecting Vitamin D Deficiency" submitted for approval of IHEC on 23.11.2021.

The Institutional Human Ethics Committee of our University hereby grants approval to your research proposal No. IHEC/21-22/ FSN-27 entitled "Exploration of Machine Learning in Development of Low-Cost Screening Method Featuring the Global Diet Quality Score for Detecting Vitamin D Deficiency" submitted by you. The Approval number for the same is AUW/IHEC/ FSN-21-22/XPD-27.

We wish you all the best in your research endeavours.

Regards,

S.Uma Mageshwari
Dr.S.Uma Mageshwari
Member Secretary

Institutional Human Ethics Committee