An Efficient Machine Learning Approach to Predict the Dietary Fiber Content of Packaged Foods

Anitha E (Senior Member, IEEE) Assistant Professor, Artificial Intelligence and Data Science Sri Eshwar College of Engineering Coimbatore, India suriya13.ms@gmail.com	Preethie K Artificial Intelligence and Data Science Sri Eshwar College of Engineering Coimbatore, India Corresponding author kpreethie@gmail.com	Amitha M Artificial Intelligence and Data Science Sri Eshwar College of Engineering Coimbatore, India amithaamitha816@gmail.com

Dhaanus K Artificial Intelligence and Data Science Sri Eshwar College of Engineering Coimbatore, India Dhaanus.k2020aids@sece.ac.in	GohulPrasath T Artificial Intelligence and Data Science Sri Eshwar College of Engineering Coimbatore, India Gohulprasath.n2020aids@sece.ac.in

***Abstract*—** Dietary fibre low consumption is common across the world and is related to a number of undesirable health issues. Considering the significance of fibre, most countries are not requiring the labelling of the amount of fibre in packaged foods and beverages, making it difficult for consumers and policymakers to keep track of their fibre intake. Here, using commonly available nutrient data on packaged products, we developed a machine learning approach for programmed and structured prediction of fibre content. Training and test datasets were created using a dataset of Australian packaged foods with known fibre content information. More variance in fibre content was explained by the use of a random forest machine learning system than by a traditional individuals fibre prediction approach. Our results shows the possibilities of machine learning to reliably and precisely forecast the fibre content of packaged items.

Keywords: *Fibre, Diet, Accuracy, Machine learning, packaged foods, food, beverages.*

I. INTRODUCTION

Fibre can be added to food items in a refined form (like insulin), in addition to being naturally present in veggies, grains, fruits, and other grain-based foods. Fibre is an integral part of a balanced diet since it raises a sense of stuffed up while regulating cholesterol and blood sugar levels. The World Health Organisation(WHO) suggests consumers to consume no fewer than 25g of fibre daily to help prevent disease, as consumption below this level are linked to type 2 diabetes, cardiovascular disease, and a number of other health problems. Although underconsumption of fibre is prevalent around the world, including in Australia, where the majority of people fall under of target level.

In contrast with unrefined or minimally processed foods, packaged food items typically include less fibre and more contaminants such added sugar, salt, and saturated fat. In Australia, packaged foods comprise about two-thirds of all meals and beverages that are sold by retailers, and this number appears to be increasing. In Australia, unless a fibre content claim is made, marking a quantity of fibre in a product's nutrition section is currently optional. As a result, consumers may make informed decisions about what to invest in, and policymakers can watch enhancements in the fibre content of packaged foods.

Using machine literacy styles would be an intriguing approach that could eliminate the requirement for customised fibre prediction. Machine literacy is a key area of artificial intelligence that enables a computer system to create an algorithm that can combine input data (like on-pack product characteristics) with a certain problem (like fibre content) based on training data. These patterns have been used in a wide range of contexts, such as food item recognition, transcriptome subtype categorization, antioxidant protein identification. shows the capability for predicting the sodium content, protein, and carbohydrate content of ingested foods working with machine learning algorithms based on their ingredients. Nevertheless, despite the importance of fibre as an essential part of a nutritious food plan, computerised approaches for figuring out the amount of fibre in packaged foods and beverages are yet to be developed..

Thereby, the intent of this project is to

(i) formulate a machine literacy approach for figuring out the measure of fibre in packaged foods and beverages based on openly accessible nutrient information.. (ii) use the approach to thoroughly evaluate fibre situations throughout

the Australian packaged food force for the first time for a sample of packaged items that fail to offer fiber levels.

II. LITERATURE SURVEY

Underconsumption of dietary fiber is prevalent worldwide and is associated with multiple adverse health conditions. Despite the importance of fiber, the labeling of fiber content on packaged foods and beverages is voluntary in most countries, making it challenging for consumers and policy makers to monitor fiber consumption. Here, we developed a machine learning approach for automated and systematic prediction of fiber content using nutrient information commonly available on packaged products. Utilization of a k-nearest neighbors machine learning algorithm explained a greater proportion of variance in fiber content than an existing manual fiber prediction approach [1]was discovered. The added-sugar prediction algorithm was developed using *k*-nearest neighbors (KNN) and packaged food information from the US Label Insight dataset (n = 70,522). A synthetic dataset of Australian packaged products (n = 500) was used to assess validity and generalization. Performance metrics included the coefficient of determination (R^2), mean absolute error (MAE), and Spearman rank correlation (ρ)[2].

The model used in this study [3] was based on the framework is focused on implementing both machine and deep learning algorithms like, logistic regression, naive bayes, Recurrent Neural Network (RNN), Multilayer Perceptron (MLP), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM). The medical dataset collected through the internet and hospitals consists of 30 patient's data with 13 features of different diseases and 1000 products. Product section has 8 features set. The features of these IoMT data were analyzed and further encoded before applying deep and machine and learning-based protocols. The performance of various machine learning and deep learning techniques was carried and the result proves that LSTM technique performs better than other scheme with respect to forecasting accuracy, recall, precision, and F1 -measures. We achieved 97.74% accuracy using LSTM deep learning model. Similarly 98% precision, 99% recall and 99% F1 -measure for allowed class is achieved, and for not-allowed class precision is 89%, recall score is 73% and F1 Measure score is 80%.[3].

In this study [4], evaluation studies judges the approach to have considerable potential to improve the daily routine of hospital dietitians as well as to improve the average quality of the dietary advice given to patients within the limited available time for dietary consultations. Our approach opens up a new avenue towards building highly specialised CBR systems in a more cost-effective way. Hence, our approach promises to allow a significantly more widespread development and practical deployment of CBR systems in a large variety of application domains including many medical applications..

The model used in this study [5] The dataset for the empirical analysis of the developed system was performed with the data set of the patients collected over the internet as well as hospitals, information's of about 50 patients were collected with thirteen features of various disease and thousand products with eight feature set. All these features were encoded and grouped into several clusters before applying into the deep learning classifiers. The better preciseness and the accuracy observed for the developed system experimentally is compared with the machine learning techniques such as logistic regression and Naïve Bayes and other deep learning classifiers such as the MLP and RNN to demonstrate the proficiency of the K-clique deep learning classifier The proposed system integrates the data mining techniques of Case-based Reasoning, Rule-based Reasoning and Genetic Algorithm. Case-based Reasoning is used to suggest a set of diet plans taken from the cases existing in the system, whereas Rule-based Reasoning is used to filter out irrelevant cases from the system and select the most appropriate case to be suggested to the patient. The Genetic Algorithm technique ensures that the diet menus suggested are customized according to each patient's personal health conditions. The output of the diet plan system is in the form of a list of specific nutritional values to be taken daily, and a menu recommendation suggesting actual dishes for the patient.

V. PROJECT DESCRIPTION

In this sections let's see about the project in detail.

A. METHODOLOGY

For our goal of estimating fibre content, we prefer to use Random forest. Random forest has been approved in prior studies to be advantageous at discerning nutritional traits and kids becoming obese based on nutrient issues which raises the possibility that a procedure may also be useful for predicting fibre content based on nutrient variables. Other than that, the random forest's internal structure is intuitive and facilitates interpretation.

The distance of Manhattan(d) between a query product and each training product was estimated at first using the Random forest algorithm via calculating the mean differences between each pair of normative nutrient values (Equation (1))..,

$$d(x, y)=\sum i=16|xi-yi|d(x, y)=\sum i=16|xi-yi| \quad (1)$$

where the i-th normalised nutritional characteristic of the query point and training point, respectively, is denoted by the letters xi and yi.

A product's fiber content (Q) was then predicted by calculating an inverse distance weighted average of the eight nearest product's fiber values (Equation (2)),

$$Q(q, w)= \sum 8j=1 wjqj \sum 8j=1 wjQ(q, w)= \sum j=18 w jqj \sum j=18 wj \quad (2)$$

where w_j is the weight of the *j*-th closest product (calculated using $1/d_j$) and q_j is the fiber content of the *j*-th closest product.

Supplementary Table S2: Descriptive statistics on nutrients for included products

Nutrient (per 100g/mL)	*n*	Mean	SD	Min	25th quartile	Median	75th quartile	Max	Spearman correlation with fiber
Sugar (g)	21,246	12.0	15.4	0.0	1.9	4.8	16.7	94.0	0.09
Starch (g)	21,246	26.8	23.4	0.0	4.6	21.5	44.9	90.0	0.29
Saturated fat (g)	21,246	3.4	5.3	0.0	0.3	1.2	4.3	66.8	0.32
Unsaturated fat (g)	21,246	8.4	12.0	0.0	0.8	3.3	10.8	67.4	0.47
Protein (g)	21,246	7.4	6.2	0.0	2.6	6.4	10.1	81.1	0.64
Sodium (mg)	21,246	301	517	0	16	225	400	1800	-0.02
Fiber (g)	11,441	5.2	5.9	0.0	1.7	3.6	7.2	90.1	1.00

SD is standard deviation

Machine Learning:
Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed for each specific task. The primary goal of machine learning is to allow computers to improve their performance on a given task as they are exposed to more data.

In traditional programming, developers write explicit instructions to instruct computers on how to perform specific tasks. However, in machine learning, instead of providing explicit instructions, the computer learns from historical data to identify patterns, relationships, and insights that help it generalize and make predictions on new, unseen data.

There are different types of machine learning approaches, but some of the most common ones include:

Supervised Learning:
In this approach, the model is trained on labeled data, where each input example is paired with the corresponding correct output. The model learns to map inputs to outputs, allowing it to make predictions on new, unseen data accurately.

Unsupervised Learning:
This approach involves training the model on unlabeled data, where the algorithm attempts to identify patterns or structures within the data without explicit guidance. Unsupervised learning is often used for tasks like clustering and dimensionality reduction.

Random forest:
An approach for machine learning that uses supervised learning is called Random Forest. It can be applied to machine learning tasks including classification and regression. It is built on the idea of ensemble learning, which is a method of implementing various classifiers to address challenging issues and enhance model performance.

IMPLEMENTATION:

The project's use the Random Forest approach for estimating fibre qualities led to appreciable advancements in that area. Using this ensemble learning approach, we were able to predict results more effectively than we could with traditional statistical methods. The Random Forest model, which correctly depicted tangled relationships between input factors and fibre characteristics, allowed for robust and trustworthy predictions.

By using feature importance analysis, it was possible to pinpoint important variables that significantly affect fibre qualities, providing information that may be used to improve processes. The scalability and adaptability of the Random Forest algorithm allowed for simple integration into real-time monitoring and quality control systems, which made it possible for the active fault identification and modification activities. Overall, the Random Forest algorithm's use demonstrated that it was successful in enhancing the precision of fibre prediction and providing helpful data for the fibre trade.

Statistical Analysis:

Python 3.7 and the tools scikit-learn, numpy, pandas, and seaborn were used for data analysis. The performance of the algorithm on the test dataset was assessed by comparing predicted values with the statistic outlined on the package. The Spearman rank correlation, mean standard deviation, and coefficient of reliability were all calculated. These metrics were determined as a whole, by food and beverage category, and by percentile of all first neighbour distances (i.e., all d1 values).

We tested the algorithm's ability to pick out commodities with low fibre density (0.9-3.7 g per 100 g or 100 mL), trivial fibre density (0.9 g per 100 g or 100 mL), and none at all. After using the RF fibre prediction approach on the test dataset, two fibre densities were determined: medium (3.7-7.3 g per 100 g or 100 mL), and high (>7.3 g per 100 g or 100 mL). These cut-offs have been chosen using the nutrient profile scoring standards that serve as the cornerstone of the Australian front-of-pack Health Star Rating system. They associate to 0, 1-3, 4-7, and 8-15 fibre points, respectively. Next, we derived evaluations for recall, precision, and classification accuracy (CA).

The manually executed recipe-based dietary prediction method devised by Ng et al. (2015), which is momentarily the only method stated that can predict fibre, was also applied to the test dataset. In a nutshell, this process comprised (i) manually coordinating each constituent on an ingredient list from a product to an ingredient in the previously outlined database of nutrient composition. It is necessary to (i) use a mathematical optimisation technique to predict the proportion of each ingredient in the product, and (ii) add the amount of fibre in each ingredient in order to calculate the product's overall fibre content because components instances are frequently not listed on the packaging.

After performing the fibre prediction approach on products that did not specify their fibre content, we calculated a median (interquartile range) fibre values for all included supplies. Global and food division levels of calculation were used. The disparities between all reported and all non-reported (but predicted) fibre values were assessed using a the Wilcoxon signed-rank test, with a statistical significance

threshold of p 0.05. This test was done to measure the disparities in fibre levels between commodities that indicate fibre content and those that do not.

Sklearn:
A free machine learning library for Python is called scikit-learn. It includes several classification, regression, and clustering methods, such as support-vector machines.

Pandas:
For the purpose of manipulating and analysing data, the programming language, Python, has a software package called pandas. It includes specific data structures and procedures for working with time series and mathematical tables.

Random forest regressor:
A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting many classification decision trees to different dataset subsamples.

Gradient boosting Classifier:
In order to minimise a loss function, the functional gradient method known as Gradient Boosting continually chooses a function that points in the direction of a weak hypothesis or a negative gradient. A powerful prediction model is created using the gradient boosting classifier by combining many weak learning models.

RESULT S AND DISCUSSION

Dataset:

The evaluation was performed on the Australian datasets. More than 80,000 packaged goods and necessities that have been distributed in Australia since January 2013 are included in this database's product information. The majority of the information (60) is gathered by skilled data collectors during routine in-store inspections at four supermarkets in downtown Sydney, one from each of the four primary supermarket groups. A trained data- collection battalion takes pictures of each product's packaging using a customised smartphone operation and retrieves all appropriate data from these pictures, including the barcode, brand name, product name, nutrition, and assets. The quality of every product's data is checked. According to the Global Food Tracking Group category scheme, products are divided into 67 food and beverage orders.

Performance metrics:
Precision:
The Accuracy is defined as the proportion of properly classified data instances over all data instances.

Recall:
It is determined by splitting the total lot of Positive samples by the proportion of Samples tested that were correctly classified as Positive. The model's recall measures its capacity for identifying positive samples. The recall increases with the number of positive samples found.

F1 score:
Using this number, we may determine the natural log of recall and accuracy; Theoretically speaking, is known weighted combination of the correctness, recall.

Accuracy:
Accuracy aids to determine the correlation and patterns between the variables in a dataset which helps to assess which model is more effective.

The performance measure of packaged food ingredients of the model is given below:

The fiber predicting project using machine learning aimed to develop a model capable of predicting fiber properties based on various input variables. The project utilized machine learning techniques to analyze a dataset containing information about fibers and their corresponding properties. By training the model on this data, it was able to learn patterns and relationships between the input variables and fiber properties, enabling it to make accurate predictions on unseen data.

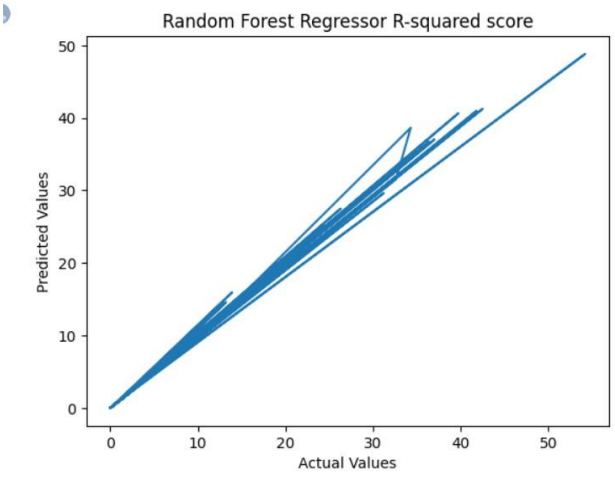

Fig : Random forest Regressor R score

The results of the study indicate that the proposed machine learning approach achieves a high level of accuracy in predicting the dietary fiber content of packaged foods. This breakthrough has significant implications for the food industry, consumers, and public health. Food manufacturers can utilize this technology to streamline quality control processes, ensure compliance with regulatory standards, and develop healthier food products. Additionally, consumers can make informed dietary choices by accessing accurate information about the fiber content of packaged foods, promoting healthier lifestyles and well-being.

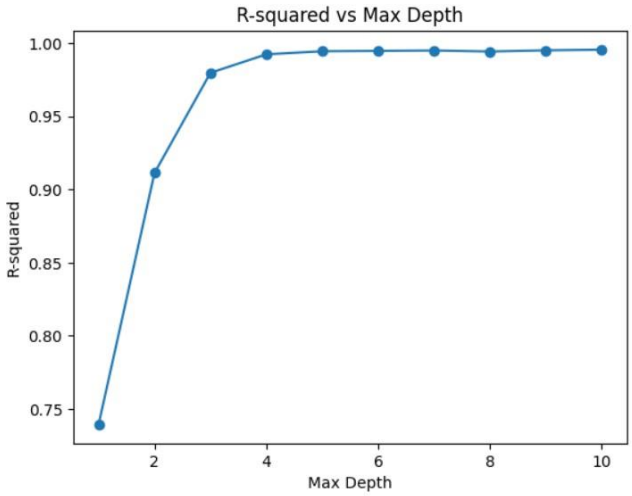

Fig R-Squared vs Max Depth

REFERENCES

[1] Phanikrishna, V.B. and Chinara, S., 2020. Time domain parameters as a feature for single-channel EEG-based drowsiness detection method. In 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-5).

[2] Shah, M., Degadwala, S. and Vyas, D., 2022, February. Diet recommendation system based on different machine learners: A review. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 290-295). IEEE.

[3] Bhat, S.S. and Ansari, G.A., 2021, May. Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

[4] Khan, A.S. and Hoffmann, A., 2003. Building a case-based diet recommendation system without a knowledge engineer. Artificial Intelligence in Medicine, 27(2), pp.155-179.

[5] Kim, J.H., Lee, J.H., Park, J.S., Lee, Y.H. and Rim, K.W., 2009, November. Design of diet recommendation system for healthcare service based on user information. In 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology (pp. 516-518). IEEE.

[6] Acton, R.B., Vanderlee, L., Hobin, E.P. and Hammond, D., 2017. Added sugar in the packaged foods and beverages available at a major Canadian retailer in 2015: a descriptive analysis. Canadian Medical Association Open Access Journal, 5(1), pp.E1-E6.

[7] Abdus Salam Khan, Achim Hoffmann, Building a case-based diet recommendation system without a knowledge engineer,Artificial Intelligence in Medicine, Volume 27, Issue 2,2003,

[8] Manoharan, D.S. and Sathesh, A., 2020. Patient diet recommendation system using K clique and deep learning classifiers. Journal of Artificial Intelligence and Capsule Networks, 2(2), pp.121-130.

[9] Husain, W., Wei, L.J., Cheng, S.L. and Zakaria, N., 2011, December. Application of data mining techniques in a personalized diet recommendation system for cancer patients. In 2011 IEEE Colloquium on Humanities, Science and Engineering (pp. 239-244). IEEE.