Accepted in BioMaterials and Polymers Horizons on 8th Feb 2023
https://mc04.manuscriptcentral.com/bph

Data mining to map nutrition value of industrial cheese produced in France

Dinh T Nguyen[a], Swati Singh[b] and Saurav Goel[a,b,c*]
[a] School of Engineering, London South Bank University, SE10AA, UK
[b] Department of Mechanical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, India
[c] Department of Mechanical Engineering, University of Petroleum and Energy Studies, Dehradun, 248007, India

Corresponding author: GoeLs@Lsbu.ac.uk*

Abstract

Cheese is a dairy product with a long history in the human diet. In countries such as France, cheese is one of the strongest attractions for visitors from the globe. Cheese was thought of only a source of energy alone however, with advances made in nutritional science, cheese has become a rich source of essential nutrients such as proteins, bioactive peptides, amino acids, vitamins and minerals. This perspective offers new insights into the nutrition analysis of various types of cheese sold in France while using advanced data visualization and analysis techniques. The objective of this article is to raise awareness about making an informed selection of the type of cheese people should consume as a long-term measure of taking a health-conscious diet. Overall, the study provides a testbed for **turophiles** (cheese lovers) in selecting the right kind of cheese to relish themselves while taking care of their health condition depending on their physical condition, in particular lactose intolerance.

1. Introduction

The word ‘cheese’ originated from the latin word ‘caseus’ (Johnson, 2017). It is believed that the art of making cheese came into being about 8000 years ago (Beresford, Fitzsimons, Brennan, & Cogan, 2001) and by now more than 2000 cheese varieties are reported, mostly prepared by coagulation of milk by chymosin, and matured for between 2 weeks and 2 years (Feeney, Lamichhane, & Sheehan, 2021). Among “ricotta”, “gouda”, “parmesan”, “Roquefort”, “brie”, “stilton”, “cheddar”, “camembert” and others, “mozzarella” continues to be the most popular cheese across the world, as its daily consumption in most types of pizza, or pasta and others is growing (Salque et al., 2013). The global cheese market is estimated to be about $100 bn and global cheese consumption is expected to increase by ~13.8% between 2019 and 2029 (Feeney et al., 2021). Cheeses are rich in nutrients and serve as a major source of high-quality proteins, lipids, vitamins (e.g. vitamins B2, A and B12) and minerals such as calcium and phosphorus (Ermolaev, Yashalova, & Ruban, 2019). However, cheese also contains high levels of saturated fatty acids (SFAs), which are commonly perceived as negatively impacting the healthfulness of the diet, and have been associated with increased blood low-density lipoprotein cholesterol (LDL) levels which presents the risk of cardiovascular

disease (CVD) and much work in this area is currently under intensive investigation (O'Brien & O'Connor, 2017).

Cheese is produced using a complex milk processing pathway. The process begins with the coagulation of milk through enzymes or using an acid treatment leading to obtaining semi-solid curds (a combination of the major milk nutrients—protein, mainly casein, and milk fat) (Feeney et al., 2021). It then requires removing water-soluble lactose by straining off the liquid whey. The straining process is commonly achieved using a coarse textile, 'cheesecloth', or plastic or metal sieves (Salque et al., 2013).

Extensive need for functionality, taste, application in baking, and nutritional aspects (low fat and low sodium) of cheese presents a strong necessity to understand the fundamental principles of cheese making which in turn requires advanced scientific investigations into the chemical, microbiological, and enzymatic changes involved in the cheese making process (Johnson, 2017).

There are two types of cheese, namely, Natural cheese and Process cheese. Natural cheese is made from four main ingredients namely, milk, rennet, microorganisms and salt which are processed through several common steps such as gel formation, whey expulsion, acid production and salt addition, followed by a period of ripening. While all acid coagulated cheeses are consumed fresh, most rennet coagulated cheese undergo a period of ripening which can range from about three weeks for Mozzarella to two years or more for Parmesan and extra-mature Cheddar. On the other hand, Process cheese requires natural cheese as the raw material. Process cheese is produced by blending natural cheese of different ages and degrees of maturity in the presence of emulsifying salts and other dairy and nondairy ingredients followed by heating and continuous mixing to form a homogeneous product with an extended shelf life (Kapoor, Metzger, & Safety, 2008). The three major types of processed cheese described by the Code of Federal Regulations (CFR) are (a) pasteurized process cheese (PC), (b) pasteurized process cheese food (PCF) and (c) pasteurized process cheese spread (PCS).

Cheese is now *internationally* known to be a tourism attraction (gastronomic, culinary) and promotes tourism as people from all over the world, travel 1000s of miles to taste a variety of cheese products. Thus, the exploitation of cheese by the tourism industry contributes to sustainability, supporting rural lifestyles and facilitating the integration of rural traditions, heritage, and natural landscapes (Ermolaev et al., 2019).

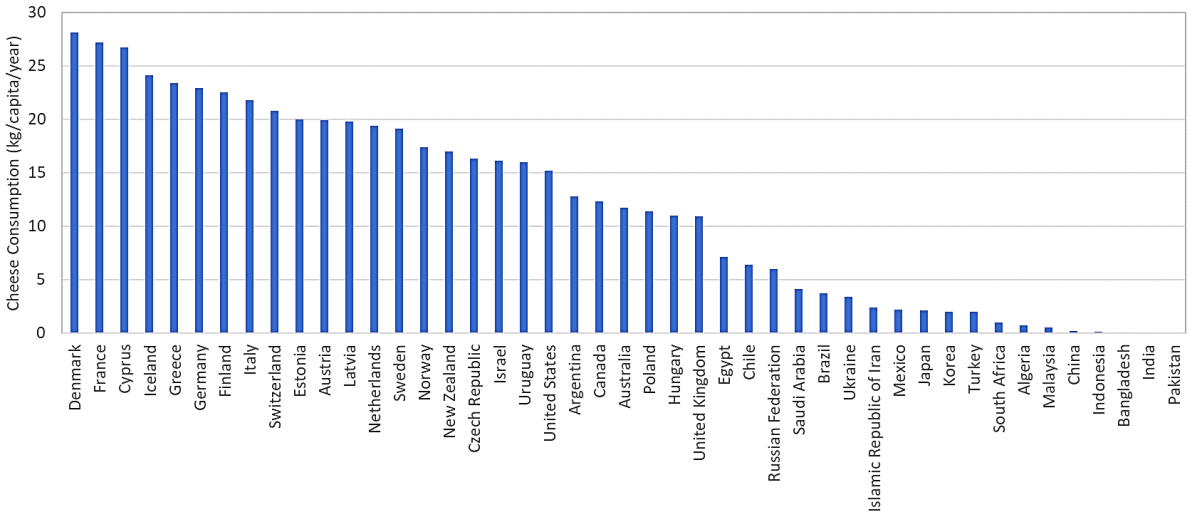


Figure 1: Global cheese consumption with respect to individual countries (authors original plot)

Fig. 1 highlights the global consumption of cheese in every country. It may be seen that European countries such as Denmark, France, Cyprus, Iceland, Greece, Germany, Finland and Italy are among the top cheese consuming countries, while countries such as India, Pakistan, China, Bangladesh and Indonesia are on the other side of the scale where cheese consumption is very low. A plausible reason for this has been pointed out by (Feeney et al., 2021) in their recent review that there are perceived health risks with excessive cheese consumption. The review by (Feeney et al., 2021) highlighted that although the preventing measure for cardiovascular disease risk is to limit the intake of saturated fat consumption due to adverse associations with low-density lipoprotein cholesterol (LDL), this advice does not account for the diversity of fatty acids present in dairy foods and cheese, whereby the combinatorial presence of various components present within the food could rather lead to health benefits (Johnson, 2017). Cholesterol and saturated fat are potential risk factors for atherosclerosis and besides fat, the calcium–magnesium ratio, lactose and milk fat globule membrane antigens may also have specific coronary atherogenic effects (Ropars, Cruaud, Lacoste, & Dupont, 2012). However, other components may reduce risks, for example, conjugated linoleic acid (CLA) which have antioxidant and anticancer properties as well as calcium which can protect against hypertension and osteoporosis, and the presence of folic acid, vitamin B6 and vitamin B12 can provide beneficial effects on plasma homocysteine level (an independent risk factor for atherosclerosis) (O'Brien & O'Connor, 2017).

These vital ingredients within cheese can adversely affect health while some ingredients are supportive of health. This contradiction opens the possibility of reexamination and analysis of varieties of cheese, which was the major motivation of this paper.

Moreover, lactose intolerant individuals (who have inadequate production of the lactase enzyme to digest lactose) may have difficulty digesting fresh milk, but they can eat certain

dairy products such as cheese or yogurt without having any problem. This is due to the fermentation processes involved in cheese preparation, which breaks down a high percentage of lactose compared to fresh milk (Beresford et al., 2001). Therefore, cheese can be an excellent candidate to replace dairy milk for those who are not fully lactose intolerant (Beresford et al., 2001). Furthermore, these days lactose-free milk can be used to make cheese for fully lactose-intolerant individuals. A list of lactose content (g) per 100 g of milk and derivatives is tabulated in the study from (Facioni, Raspini, Pivari, Dogliotti, & Cena, 2020). Still, a detailed description of all nutrients in worldwide popular cheese is essential for consumers to make the right choice.

Through this paper, our ambition was to provide clarity to those individuals, who perceive cheese negatively and avoid consuming it. This is especially crucial for the Asia region given the imminent food security challenge in the wake of climate change. We performed intensive data analysis to understand the hidden nutritive value in the cheese using large data collection of market cheese products with the objective of helping consumers to choose the right cheese to support their need.

2. Research methodology

To examine the nutrition value of various cheese products concerning their nutrition score, energy content, fat/saturated fat content, cholesterol, carbohydrates, sugar/fiber, protein, and salts content per 100 g of cheese, a data-mining approach was adopted with the support of the Python 3.9 platform. This study used the dataset collected from an open-source database (See Data statement). This dataset contained different categories of cheese which were grouped into five different grades (a, b, c, d and e) based on their nutritional value. Also, the dataset contains eleven qualitative variables (additives, nutriscore_score, energy-kcal_100g (energy in kcal in per 100g of cheese), fat_100g, saturated-fat_100g, cholesterol_100g, carbohydrates_100g, sugars_100g, fiber_100g, proteins_100g, salt_100g).

Table 1 describes all these variables present in the dataset and their corresponding number of instances. It can be noted that data is missing for some of the variables.

Table 1. List of variables and their corresponding number of instances that are analyzed in this study.

Variables	Description of used variables	Available data
nutriscore_grade	Nutrition grade	11943
additives_n	Number of added additives	10368
nutriscore_score	Nutrition score	11557
energy-kcal_100g	Energy content (kcal) in per 100g of cheese	12169
fat_100g	Fat content (kcal) in per 100g of cheese	8461
saturated-fat_100g	Saturated fat content (kcal) in per 100g of cheese	2275

cholesterol_100g	Cholesterol content (kcal) in per 100g of cheese	1435
carbohydrates_100g	Carbohydrates content (kcal) in per 100g of cheese	12218
sugars_100g	Sugar content (kcal) in per 100g of cheese	12030
fiber_100g	Fiber content (kcal) in per 100g of cheese	11077
proteins_100g	Protein content (kcal) in per 100g of cheese	4454
salt_100g	Salt content (kcal) in per 100g of cheese	9578

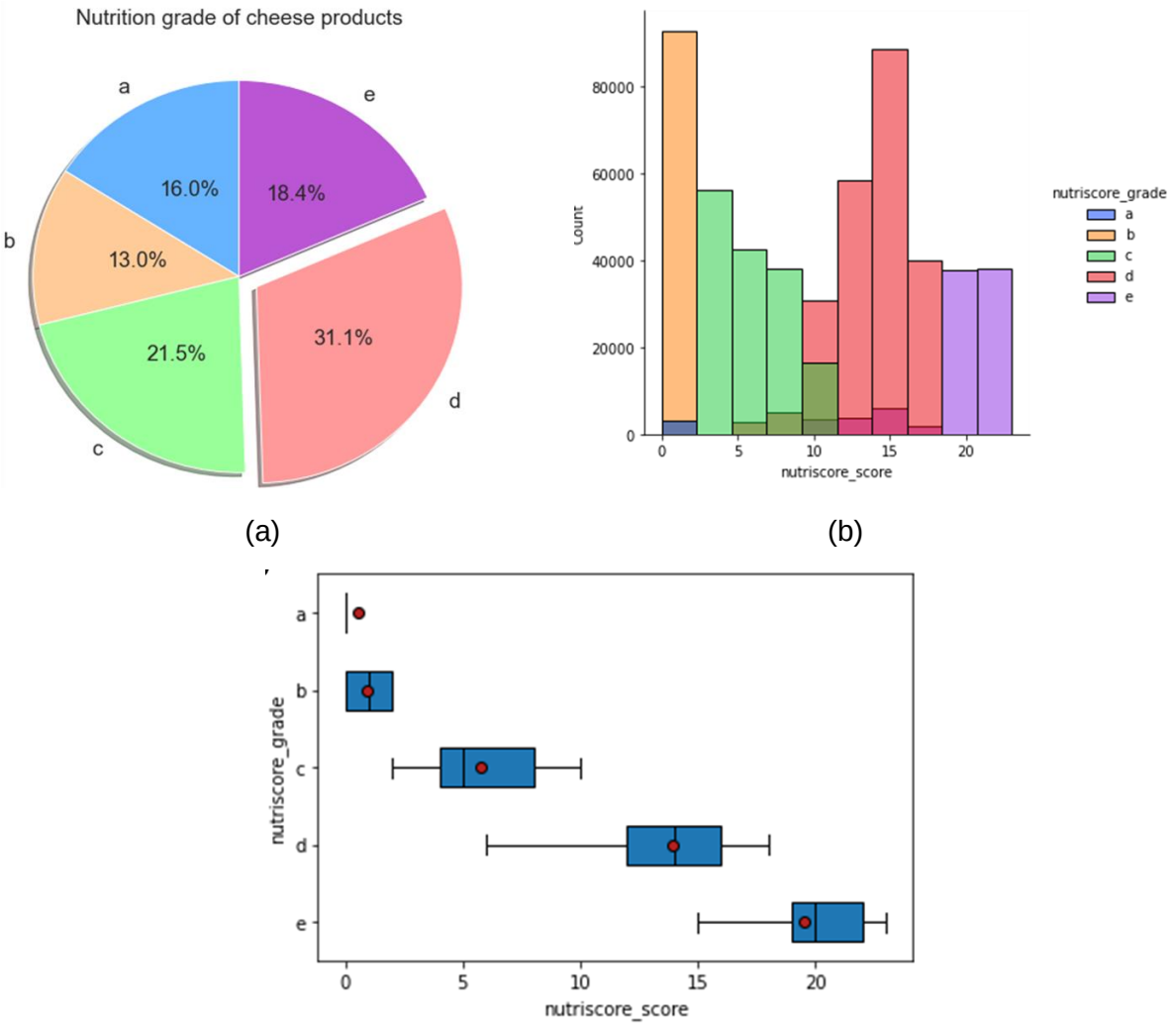


Figure 2. Complete data description **(a)** pie chart showing the percentage of data belonging to each nutrition grade of cheese products, **(b)** bar chart showing the correlation of nutrition grade to the nutrition score, **(c)** boxplot between nutrition score and nutrition grade in different cheese products.

Furthermore, to get better insights into complete data, matplotlib and seaborn libraries in python were used for visualization and analysis of data. A pie chart shown in Fig. 2(a), represents the percentage of data corresponding to five different grades of cheese. The highest instances of cheese correspond to nutrition grade 'd' (31.1%) while grade b (13.0%) contained the least instance of cheese. Fig. 2(b) demonstrates a bar chart representing the nutrition score as per different nutrition grades of cheese. A box plot between nutrition score and nutrition grade for a variety of cheese is shown in Fig. 2(c). One can see that the better the nutrition grade (towards a), the smaller the nutrition score (in the range of 1-2 for grade a cheese), while the nutrition score is quite high for grades 'd' and 'e' cheese types.

3. Result and discussions

To observe the correlation between different qualitative variables, a correlation matrix is presented in Fig. 3. It can be observed that some groups of nutrients are well correlated, for example, nutrition score vs salt, nutrition score vs energy, nutrition score vs saturated fat, nutrition score vs fat and carbohydrates vs. sugar. The number of additives was seen to correlate with sugar and carbohydrate content per 100g of cheese.

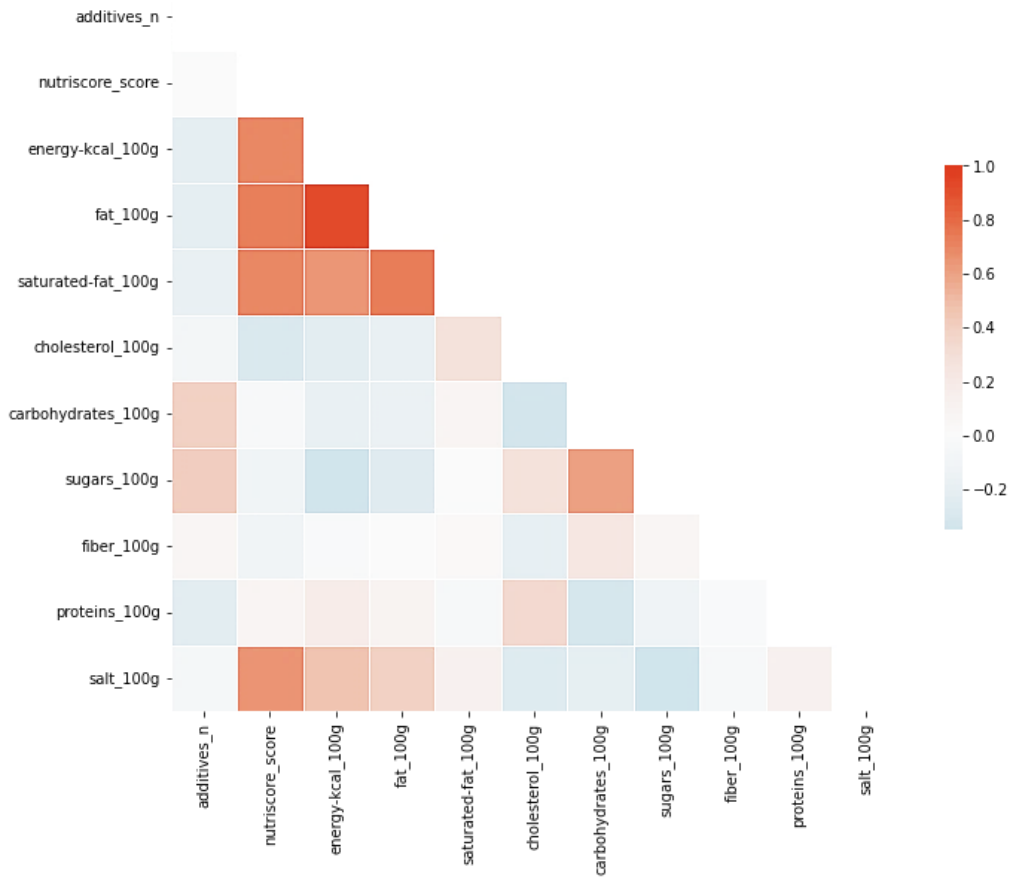
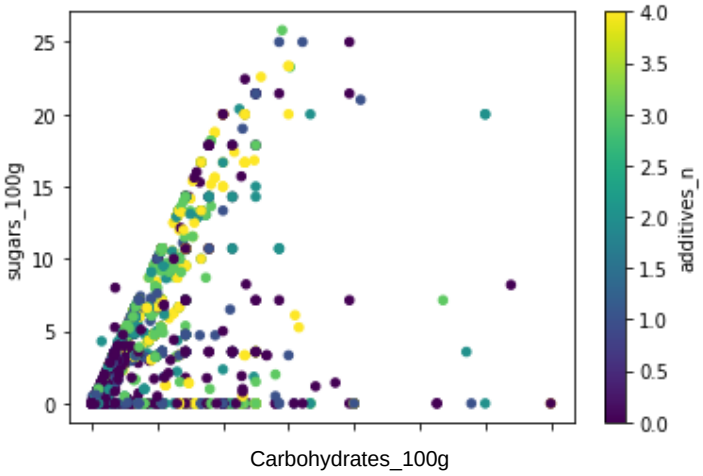
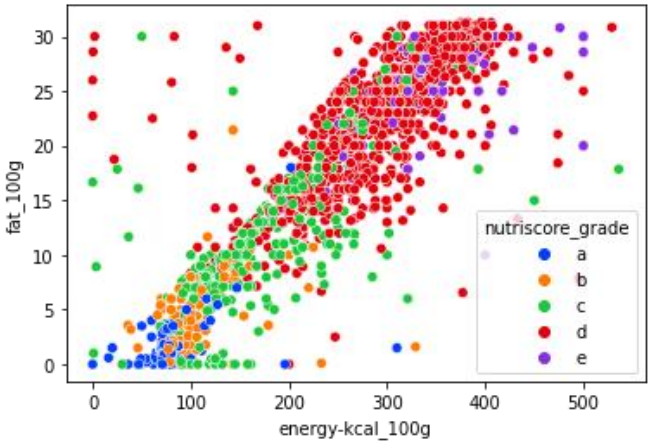


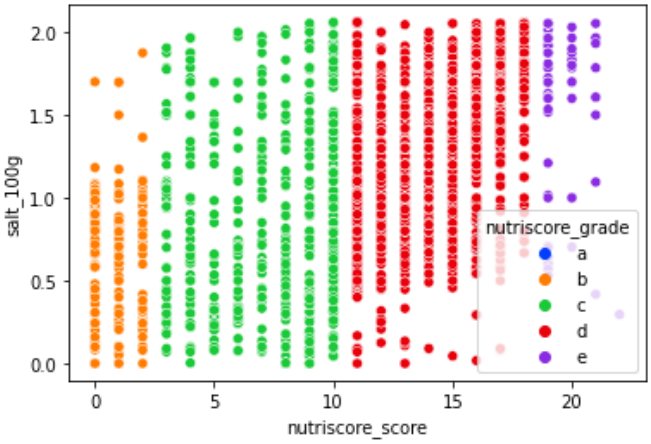
Figure 3. Correlation matrix of the nutrients in cheese products per 100g


sugars_100g
Carbohydrates_100g
additives_n
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

(a)


fat_100g
energy-kcal_100g
nutriscore_grade
a
b
c
d
e
0
5
10
15
20
25
30
0
100
200
300
400
500

(b)


salt_100g
nutriscore_score
nutriscore_grade
a
b
c
d
e
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20

(c)

Figure 4. Scatter plot showing **(a)** sugars vs. carbohydrates in relation to number of additives, **(b)** fat vs. energy content per 100g of cheese for different grades of cheese, **(c)** Salt vs nutrition score for different grades of cheese.

Furthermore, the relation between sugars and carbohydrates concerning the number of additives is highlighted in Fig. 4(a). It is well known that carbohydrates are linearly correlated to sugars, but surprisingly these were also seen related to the number of additives such that more sugars and carbohydrates showed a positive correlation with the additives. A linear correlation between the amount of fat and energy in cheese products for every 100g of cheese was also noticed which is demonstrated in Fig. 4(b). The higher grades cheese types (a and b) tend to have lower fat and energy content, contrary to lower grades cheese types (d and e).

Fig. 4(c) presents the nutrition score concerning salt for all five grades of cheese. It shows a linear relationship between the amount of salt and the nutrition score; a product with a higher amount of salt tends to have a higher nutrition score.

The distribution of different nutrients in cheese products is presented in Fig. 5. The interesting fact here is that most of the cheese products in this dataset have energy content ranging from 0 to 550 kcal, fat in the range of 0 to 32 g, saturated fat in the range of 0 to 12 g, cholesterol in the range of 0 to 0.035 g, carbohydrates in the range of 0 to 75 g, sugar in the range of 0 to 27 g, fiber in the range of 0 to 8 g, protein in the range of 0 to 18 g, and salt in the range of 0 to 2g per 100g of cheese for five different grades of cheese.

Although the fat content of cheese varies considerably depending on the milk used and the method of manufacture, it is recommendable by experts worldwide to reduce the intake of both total and saturated fat (Johnson, 2017). In this study, higher grade cheese was found to contain lower content of total fat and saturated fat per 100g of cheese as shown in Fig. 5(b) and Fig. 5(c) respectively. The cholesterol content of cheese is a function of its fat content which can be observed from Fig. 5(d). It shows that the lower grade cheese *d* and *e* showed higher levels of cholesterol per 100g of cheese, as the fat content of these grades of cheese is much higher compared to the cheese of grade *a* and grade *b*. Furthermore, as it is well known that approximately 70% of the global adult population are lactose intolerant and therefore lower carbohydrate content cheeses are in demand (Fox et al., 1996). Although during cheese manufacturing, only trace amounts of carbohydrates remain, still the lower grade cheese (grade d and grade e) can be observed to contain higher carbohydrates per 100g of cheese, as shown in Fig. 5(e). Thus, grade a and b cheese types can be consumed without ill effects by lactose-intolerant individuals who are deficient in the intestinal enzyme, β-galactosidase. Sugar content per 100g of cheese was found higher for grade ‘d’ cheese as shown in Fig. 5(f).

Cheese contains a high level of biologically valuable protein and fiber. The protein/fiber content of different cheeses tends to vary inversely with fat content (O’Brien & O’Connor, 2017) which is shown in Fig. 5(b), Fig. 5(g) and Fig. 5(h). It can be observed that higher-grade cheese (type a and b) contains much higher protein and fiber per 100g of cheese compared to lower-grade cheese (type d and e). Both sugar and salt are observed to be on the higher side in the low-grade types of cheese (grade d and grade e). This study

alluded to the fact that the higher-grade cheese type a and type b are healthier compared to other cheese types and they must be preferred as a healthier diet option considering the growing worldwide health issues.

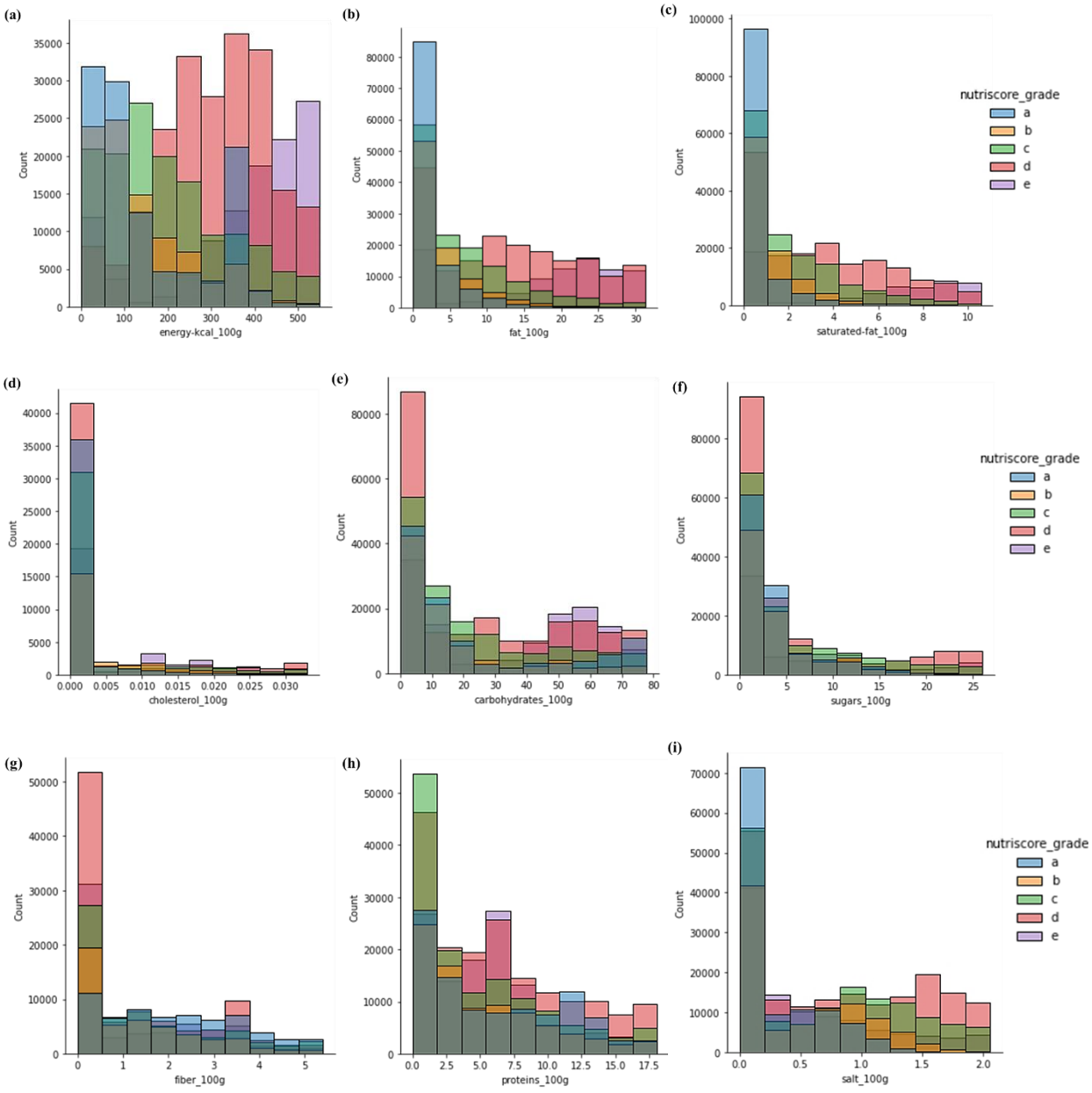


Figure 5. Distribution of different nutrients in cheese products **(a)** energy content, **(b)** fat content, **(c)** saturated fat content, **(d)** cholesterol content, **(e)** carbohydrates content, **(f)** sugars content, **(g)** fiber content, **(h)** proteins content, **(i)** salt content per 100 g of cheese.

This study demonstrates the application of data mining technique in examining the nutrition value of various cheese products using the python platform. From the available dataset it can be seen that the information is missing from various qualitative features,

which limit its reliability. The well-known fact that these techniques provide more robust and reliable predictions with large and complete dataset that contains no missing values. Thus, this study can be extended in future for big data with a complete set of information on each variable for improving the current status quo, robustness, and the degree of certainty in making reliable predictions.

It may be noticed that the information in the available datasets was missing from various qualitative features. Data mining techniques are robust for large and complete dataset that contains no missing values. Thus, this study can be extended in future for more reliable results after obtaining complete information on each variable and more amount of data.

4. Conclusion

The essential nutrients such as proteins, minerals, and vitamins in cheese are known to provide health benefits via the production of certain peptides and free amino acids. However, due to the presence of saturated fatty acid, cholesterol, sugar and salt content, it suffers from adverse nutritional benefits. Thus, it is necessary to differentiate between good and bad cheese depending on the different nutrients it contains. This study used data mining techniques using Python software to map the crucial nutrition value of industrial cheese produced in France as a testbed study. A large dataset of Industrial cheese produced in France was accessed to study nutrient values such as energy, fat, carbohydrates, protein, fiber, salt and cholesterol content.It was observed that high-grade cheese (type a and type b) possesses higher fiber and protein content with lower salt, sugar, cholesterol, energy, carbohydrates, total fat, and saturated fat contents per 100g of cheese. Thus, this study suggests that one should consider minimizing consumption or switching of cheese if their favorite cheese contains sugar of more than 25g, salt of more than 2g, or energy above 500 kcal to remain healthier.

This study advocates the use of data mining techniques for illustrative visualization of complex nutritional information supplied with the food products that can help consumers to have informed idea of taking nutritive diet by using simple tools like python. This study is especially helpful to ‘Turophiles’ (cheese lovers) as it can aid in selecting the right type of cheese to maximise the nutritive benefits. Furthermore, this study can be extended to analyse other food products for food labelling to enhance consumer’s awareness of a safer choice.

Data availability

The data accessed and used for analysis in this paper can be downloaded from the

open database: https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv

Acknowledgments

SG acknowledge the financial support provided by the UKRI via Grants No. EP/S036180/1and EP/T024607/1, feasibility study awards to LSBU from the UKRI National Interdisciplinary Circular Economy Hub (EP/V029746/1) and Transforming the Foundation Industries: a Network+ (EP/V026402/1), the Hubert Curien Partnership award 2022 from the British Council and Transforming the Partnership award from the Royal Academy of Engineering (TSP1332).

References:

Beresford, T. P., Fitzsimons, N. A., Brennan, N. L., & Cogan, T. M. J. I. d. j. (2001). Recent advances in cheese microbiology. *11*(4-7), 259-274.

Ermolaev, V. A., Yashalova, N. N., & Ruban, D. A. J. S. (2019). Cheese as a tourism resource in Russia: The first report and relevance to sustainability. *11*(19), 5520.

Facioni, M. S., Raspini, B., Pivari, F., Dogliotti, E., & Cena, H. (2020). Nutritional management of lactose intolerance: the importance of diet and food labelling. *Journal of translational medicine, 18*(1), 1-9.

Feeney, E. L., Lamichhane, P., & Sheehan, J. J. J. I. J. o. D. T. (2021). The cheese matrix: understanding the impact of cheese structure on aspects of cardiovascular health–a food science and a human nutrition perspective. *74*(4), 656-670.

Fox, P., O'connor, T., McSweeney, P., Guinee, T., O'brien, N. J. A. i. f., & research, n. (1996). Cheese: physical, biochemical, and nutritional aspects. *39*, 163-328.

Johnson, M. J. J. o. D. S. (2017). A 100-year review: cheese production and quality. *100*(12), 9952-9965.

Kapoor, R., Metzger, L. E. J. C. R. i. F. S., & Safety, F. (2008). Process cheese: Scientific and technological aspects—A review. *7*(2), 194-214.

O'Brien, N. M., & O'Connor, T. P. (2017). Nutritional aspects of cheese. In *Cheese* (pp. 603-611): Elsevier.

Ropars, J., Cruaud, C., Lacoste, S., & Dupont, J. J. I. j. o. f. m. (2012). A taxonomic and ecological overview of cheese fungi. *155*(3), 199-210.

Salque, M., Bogucki, P. I., Pyzel, J., Sobkowiak-Tabaka, I., Grygiel, R., Szmyt, M., & Evershed, R. P. J. N. (2013). Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *493*(7433), 522-525.