Glycaemic Index Prediction: a Pilot Study of Data Linkage Challenges and the Application of Machine Learning

Jingyuan Li and Ognjen Arandjelović
University of St Andrews, United Kingdom

Abstract—The glycaemic index (GI) is widely used to characterize the effect that a food has on blood glucose which is of major importance to diabetic individuals as well as the general population at large. At present, its applicability is severely limited by the labour involved in its measurement and the lack of understanding about how different foods interact to produce the GI of the meal comprising them. In this pilot study we examine if readily available biochemical properties of foods can be used to predict their GI, thus opening possibilities for practicable use of the GI in the management of blood glucose in everyday life. We also examine practical challenges in the cross-linking of food information sources collected by different organizations, and highlight the need for the development of a universal standard which would facilitate automatic and error free data integration.

I. INTRODUCTION

With an increasing prevalence of metabolic disorders and the cost they incur (both in terms of so-called human cost, and the associated health care financial costs), various aspects of the modern diet have come under scrutiny. One of the important considerations in this context pertains to the control of blood glucose [2], with the glycaemic index (GI) being most widely used to characterize the effect of a specific food [1]. However, its everyday applicability is severely limited by the fact that the GI is for practical reasons measured for individual foods – a laborious process in its own right – while real meals include combinations of foods which interact in terms of their glycaemic effects. The goal of the present pilot study is to investigate if the GI of an arbitrary food can be predicted from its readily available biochemical properties, and to examine various practical and methodological issues which need to be overcome to facilitate the use of such an approach in the real world.

II. RELEVANT BACKGROUND

Carbohydrates, together with proteins and fats, comprise one of the three types of macronutrients i.e. food components consumed in large quantities (in contrast to micronutrients, such as vitamins and minerals, which are consumed in far lower doses). Consisting of carbon, hydrogen, and oxygen atoms only, by their broad chemical structure carbohydrates can be classified into monosaccharides and more complex carbohydrates (polysaccharides) formed by the polymerization of monosaccharides, as illustrated in Fig 1. The former group includes molecules such as glucose, fructose, and galactose, as well as e.g. glyceraldehyde, ribose, and deoxyribose. Some well known polysaccharides are starches, amylose and amylopectin, cellulose, glycogen, and numerous others.

Under ‘normal’ circumstances, glucose is the body’s preferred source of energy. A constant supply of glucose to different tissues is facilitated by the blood stream – the blood glucose concentration is tightly regulated by hormones, being approximately in the range from 4.4 to 6.1 mmol/L except in postprandial periods.

The consumption of food affects blood glucose. It is generally considered that by far the greatest effect is caused by the intake of carbohydrates, a small effect by proteins, and little to no effect by fats [7]. As noted earlier, in healthy individuals the blood glucose levels is tightly regulated by means of negative feedback. Thus a rise in blood glucose levels effects the secretion of insulin which triggers glycogenesis (the conversion of glucose into glycogen), and the take-up of glucose by myocytes and adipocytes through the GLUT4 transporter. Similarly, a drop in blood glucose is counteracted by means of a number of catabolic hormones – mainly glucagon, cortisol and catecholamines – which enhance the release of glucose from glycogen, its synthesis from amino acids or fatty acids, etc. A notable malfunctioning of this feedback system is exhibited by individuals suffering from different types of diabetes mellitus [10].

A. Glycaemic index

Blood glucose changes following the intake of food can be of importance to a variety of subpopulations. Individuals who suffer from different metabolic disorders, such as diabetes mellitus mentioned earlier [10], and athletes are amongst the most notable examples [6], but for general health reasons, blood glucose control is useful and is advised universally [2]. The glycaemic index (GI) is a widely used measure used to characterize the effects of a food on the blood glucose concentration. In recent years the GI has permeated the mainstream and become a well-known term used in the marketing of foods.

The GI is often incorrectly understood as a measure of how quickly the carbohydrates present in a food are digested. For example, a web site dedicated to the GI and run by University of Sydney\(^1\) states:

\(^1\)http://www.glycemicindex.com/
As stated earlier, this is incorrect. The glycaemic index of a food is defined as the area under the incremental blood glucose concentration curve following the ingestion of the food, normalized to contain 50g of carbohydrates, relative to the corresponding area of response to the intake of 50g of glucose. Thus, contrary to the claims above, high GI does not imply a greater spike in blood glucose, nor faster carbohydrate digestion.

“...the ingestion of a food (x-axis shows time in minutes following the ingestion, and y-axis the blood glucose concentration in mg/dL). The glycaemic index (GI) of a food is defined as the area under the incremental blood glucose concentration curve following the ingestion of the food, normalized to contain 50g of carbohydrates, relative to the corresponding area of response to the intake of 50g of glucose. Thus, contrary to common claims (present in many academic articles, health guidelines issued by government organization, food manufacturers' claims, and popular media, to name just a few), high GI does not imply a greater spike in blood glucose, nor faster carbohydrate digestion.

"The Glycemic Index (GI) is a relative ranking of carbohydrate in foods according to how they affect blood glucose levels. Carbohydrates with a low GI value (55 or less) are more slowly digested, absorbed and metabolised..."

Various health advisory bodies, including government boards, such as Diabetes UK\(^2\) echo similar statements:

"...Slowly absorbed foods have a low GI rating, while foods that are more quickly absorbed have a higher rating..."

As stated earlier, this is incorrect. The glycaemic index of a food is defined as the area under the incremental blood glucose concentration curve following the ingestion of the food, normalized to contain 50g of carbohydrates, relative to the corresponding area of response to the intake of 50g of glucose. Thus, contrary to the claims above, high GI does not imply a greater spike in blood glucose, nor faster carbohydrate digestion. Indeed, even a cursory appreciation of the control loop used to control blood glucose should make it readily clear that the claims such as those quoted above cannot be correct for they ignore the effects of food on insulin secretion and the effected glucose clearance. This explains what is sometimes seen as a confusing lack of correlation between the GI and the less well-known insulin index (II) [9].

III. TECHNICAL CONTRIBUTION

In this section we explain our methodology in detail, summarize the results of our experiments, and discuss our findings.

A. Data linkage

Nutritional composition analysis and the testing of the glycaemic response of different foods are generally performed by different organizations. The former is usually done by government bodies in charge of ensuring food safety and quality such as the United States Department of Agriculture (USDA) in the USA\(^3\) or the Food Standards Agency (FSA) in Australia\(^4\). In principle this does not pose problems – all that is needed is to cross-link different data sources. However, numerous challenges emerge in practice.

Firstly, it should be noted that there is a major imbalance between the volume of reliable nutrition composition and glycaemic index data available. For example, the USDA food composition list includes 77,055 different individual foods at the time of this writing, whereas the most comprehensive glycaemic index list is sized only 2,480 [1] – a difference of one and a half orders of magnitude. More importantly, there is a major discrepancy in the manner in which the two data sources label different foods. For example, a search for entries containing the words “muffin” and “apple” returns the following list of USDA entries:

- MARKETS OF MEIJER, MUFFINS, APPLE SPICE, UPC: 04125000228
- KODIAK CAKES, MINUTE MUFFINS, APPLE CINNAMON OAT, UPC: 70559011566
- KRUSTEAZ, SUPREME MUFFIN MIX, APPLE CINNAMON, UPC: 041449301051
- EUROPEAN GOURMET BAKERY, ORGANIC MUFFIN MIX, APPLE CINNAMON, UPC: 058336175220
- JIFFY, MUFFIN MIX, APPLE CINNAMON, UPC: 072486002502
- English muffins, raisin-cinnamon (includes apple-cinnamon)
- HODGSON MILL, APPLE CINNAMON MUFFIN MIX, UPC: 071518007102
- APPLE WALNUT MUFFINS, UPC: 014821044037
- APPLE OAT MUFFIN MIX, UPC: 094522077848
- English muffins, raisin-cinnamon, toasted (includes apple-cinnamon)
- DUNCAN HINES, SIMPLE MORNINGS, PREMIUM MUFFIN MIX WITH OATMEAL GRANOLA TOPPING MADE WITH REAL APPLES, APPLE CINNAMON, UPC: 644209420285

and the following from the glycaemic index corpus:

- Apple muffin, made with rolled oats and sugar
- Apple muffin, made rolled oats and without sugar
- Apple Blueberry muffin (Sara Lee Bakery, Australia)
- Apple, oat, sultana muffin (Australia)

Clearly even for a human it is difficult to see which (if any) pairs of entries from the two information sources correspond to one another.

Considering the size of the data corpus, we sought to investigate if the matching process can be automated. We evaluated several approaches (described next) all of which adopt the same broad strategy. Each entry in the glycaemic index list is compared with all entries in the USDA corpus for similarity. Then, USDA entry candidates are ranked by similarity and the one with the highest similarity chosen as the correct match if (i) it exceeds a certain threshold, and (ii) it is sufficiently higher then the second ranked candidate. Both criteria are aimed at ensuring that the matching confidence is high. The former threshold can be set simply, in a data driven approach, by considering the histogram of pair-wise similarities and setting the threshold at the level which splits the data in the ratio of approximately 1 : 2.480/77.050. The latter parameter is a free parameter of the method which we experimented with in an effort to find an acceptable threshold.

\(^2\)https://www.diabetes.org.uk/Guide-to-diabetes/
\(^3\)Managing-your-diabetes/Glycaemic-Index-GI/
recall/precision operating point. We next describe different entry similarity measures we evaluated.

1) Unordered word matching mean based similarity: We first attempted what is arguably the simplest way of comparing two textual food entries: by looking at the proportion of common words. Let the two strings be $s_1$ and $s_2$, e.g. “English muffins, raisin-cinnamon, toasted (includes apple-cinnamon)” and “Apple muffin, made with rolled oats and sugar”. The two are tokenized into words and normalized for the grammatical number (plural form is converted into singular), and so-called stopwords removed. Then the number of mutual words is counted. In this example, this count would be $n_c = 2$ as the words in common are “apple” and “muffin”. Next, the count is normalized by unique token counts in each string which in this case are 7 and 6 (1: “English”, 2: “muffin”, 3: “raisin”, 4: “cinnamon”, 5: “toasted”, 6: “includes”, 7: “apple”, and 1: “apple”, 2: “muffin”, 3: “made”, 4: “rolled”, 5: “oats”, 6: “sugar”). If the two proportions are $p_1$ and $p_2$, then the final similarity score is computed as their average:

$$\rho_u = \frac{p_1 + p_2}{2} \tag{1}$$

In the given example $\rho_u = \frac{2/7 + 2/6}{2} \approx 0.31$.

2) Unordered word matching minimum based similarity: In our experiments we quickly found that the simple matching method described previously does not work satisfactorily. In particular we observed that in cases when a short string is compared with a long one, the averaging step in (1) would often result in an artificially overestimated similarity due to an often high proportion of matched words in the short string. As a means of overcoming this problem, we next examined a score based on the minimum rather than mean proportion:

$$\rho_n = \min(p_1, p_2) \tag{2}$$

For the example given in the previous section this gives $\rho_n = \min(2/7, 2/6) \approx 0.29$. Clearly in this case the strings are of approximately the same length so the difference is not large but in many instances the discrepancy in the length of entries is much greater as is the difference between $\rho_u$ and $\rho_n$.

3) Order weighted word matching based similarity: While the alteration in the way asymmetrical string matching scores are combined into a single similarity estimate did successfully address the problem of imbalanced lengths of textual descriptions of foods, we found that the use of unordered word matching carries an inherent limitation exhibited most prominently in comparisons involving detailed entries. In particular, we found that often the less relevant, minor qualifiers appended at the end of descriptions (e.g. regarding the use of a specific spice or flavouring) would mislead our matching procedure. Based on the insight that the descriptions tend to be hierarchical, with the most dominant ingredients being listed first, we formulated a matching procedure that takes this into account. Drawing from the abundance of empirical evidence in the literature on the so-called power law structure exhibited in similar problems [5], [4] we estimate the matching fitness of a string $s_1$ with respect to string $s_2$ by weighting each word in $s_1$ which is also present in $s_2$ by $1/r$ where $r$ is the order of the corresponding word in $s_1$. To continue using our previous example, the description “English muffins, raisin-cinnamon, toasted (includes apple-cinnamon)” matched against “Apple muffin, made with rolled oats and sugar” yields the following score:

$$\phi_1 = \frac{1}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \approx 0.24 \tag{3}$$

Here the power law weights are 1, 1/2, …, 1/8 for the 8 words in the string, and the common words present are in places 2 (“muffin”) and 7 (“apple”). The denominator normalizes the score to be in the range $[0, 1]$. As before, the two asymmetric and unidirectional similarities are fused as per (2):

$$\rho_\sigma = \min(\phi_1, \phi_2) \tag{4}$$

4) Structural block order weighted word matching based similarity: As an alternative to the preceding approach but following the same idea, we also explored the possibility of applying weighting not on a word level but to groups (blocks) of words. This approach is motivated by the observation that the specific word order may exhibit some variability but that saliency (in the context of food descriptions in the semi-structured corpora considered here) is better captured by the order of a segment of the description, separated by commas. For example, the description “English muffins, raisin-cinnamon, toasted (includes apple-cinnamon)” contains three segments: (1) “English muffins”, (2) “raisin cinnamon”, and (3) “toasted includes apple cinnamon”. A power law weight is then applied uniformly to all words in a segment. Thus comparing the aforementioned description with “Apple muffin, made with rolled oats and sugar” yields the following score:

$$\hat{\phi}_1 = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} \approx 0.36 \tag{5}$$

Here the power law weights are 1, 1/2, 1/3 for the 3 word blocks in the string, and the common words present are in blocks 1 (“muffin”) and 3 (“apple”). As before the denominator normalizes the score to be in the range $[0, 1]$, and the two asymmetric and unidirectional similarities are fused as per (2):

$$\rho_\sigma = \min(\hat{\phi}_1, \hat{\phi}_2) \tag{6}$$

5) Data linkage: summary: We have already highlighted the limitations of the first two methods for cross-linking data sources on the nutrition composition and glycaemic index measurements on different foods, which use unordered word based matching. The two alternatives, which exploit the textual structure of data entries and employ the power law to weight different word matches do a good job at overcoming the aforementioned limitations but our experiments demonstrate that they nevertheless fail to attain the necessary performance level in terms of precision and recall, to yield a practical, robust solution. Our examination of failures strongly suggests that rather than being a weakness of the matching methods themselves, this is explained by the ambiguity in the way data is entered, and the lack of a uniform standard across the two information sources. Hence, for the remainder of this paper we rely on a subset of data (100 food entries) which we cross-linked manually and for which ambiguity was minimal. Generally these were raw foods such as vegetables and fruits.
B. Glycaemic index prediction

Recall that our goal herein is to examine if it is possible to predict the glycaemic index of a food from its readily measurable physical and chemical properties, such as macronutrient composition, water content etc. Specifically in this experiment we used seven independent variables: the content of (1) water (% of mass), (2) energy (kJ per 100g), (3) protein (% of mass), (4) total carbohydrates (% of mass), (5) sugars (% of mass), (6) fibre (% of mass), and (7) lipids (% of mass).

Considering that the present work is the first attempt at addressing the problem at hand, we sought to adopt a well understood and readily interpretable prediction model. Our goal was both to examine the potential of the idea itself as well as to facilitate an increased understanding of the underlying challenges, thus providing guidance for future work efforts. Hence we opted for the widely used multiple regression model which bases its prediction on a weighted linear combination of the independent input variables:

$$y_n = \sum_{i=1}^{n} \omega_i x_i$$

Here $y_n$ is the glycaemic index, $x_i$ the $n = 7$ food characteristics detailed earlier, and $\omega_i$ the corresponding weights. Fitting is performed by seeking such weights $\omega_i$ (free parameters of the model) which maximize the goodness of fit on a training set in terms of the mean squared error. We used the standard five-fold cross-validation training and testing methodology.

1) Results and discussion: The result of our experiments is summarized by the scatter plot shown in Fig 3 (a randomly selected subset of 100 foods is shown for clarity). The corresponding Pearson’s correlation coefficient is 0.62.

For these findings to be interpreted correctly, that is, in order to assess how good or not the prediction is, it is imperative to contextualize the problem. Specifically, firstly it is important to understand that the glycaemic response of a person to the intake of a food can be highly variable across different individuals. Depending on the metabolic health of the individual, the insulin response itself exhibits variability. Similarly, glucose clearance from the bloodstream in response to insulin is dependent on the insulin sensitivity of the person’s cells (primarily myocytes and adipocytes). Indeed, this is captured by the variance across studies reported in the glycaemic index tables [1]. Herein we used merely the average value which introduces a significant loss of information.

Moreover, it is well known that the glycaemic index of foods with identical compositions can vary substantially. For example, the GI is increased with the increase in the surface area of a food exposed to digestive enzymes, which explains why the GI of rice greatly depends on its grain size/length. In this pilot study we conducted our experiments aware of these limitations, choosing to limit the set of independent variables used to those most comprehensively available at this time.

In the context of the aforementioned caveats, the prediction results in Fig 3 look remarkably promising. That such good prediction is possible using a small and inherently limited set of independent variables suggests that the use of a richer set of features would result in improved performance.

IV. Summary

In this paper we presented experiments demonstrating that the glycaemic index of a food can be predicted from simple and readily available biochemical properties of the food. We also highlighted the need for the adoption of a common standard for recording different types of information on foods so that this information can be cross-linked automatically and without ambiguity. In addition to the development of such standards, future work should concentrate on the development of models which capture stochastic and inter-personal variability in blood glucose regulation, and the use of a larger number of food properties, such as density, viscosity etc. It is our hope that the present work will pave the way towards the development of a novel set of innovative applications of computer science in health care, thereby contributing to the promising recent contributions in this sphere [3], [11], [12], [13].

REFERENCES