Spoon Guru®

Can AI perform as accurately as healthcare professionals in recommending foods suitable for those with nut allergy?

Authors:

Dr Danielle McCarthy[1]
[1]*Live It Up Ventures, Belfast, United Kingdom*

Oliver Platt
Statistical Analysis Lead

Spoon Guru Whitepaper, 2018

Contents

Abstract

People are using mobile applications to track their health, tailor their food intake and reach wellness goals. The integration of machine learning (ML) capabilities, a branch of artificial intelligence (AI), beyond mobile health devices into, for example, smart home solutions and food retail offerings, stands to deliver significant progress in the field of personalised nutrition and health behaviour change. It is of paramount importance that these innovations are based on robust nutrition and health science, that their outputs are appropriately validated and deliver meaningful impact.

The aim of this study was to compare if a ML technology can identify foods that are suitable or unsuitable for those with a nut allergy as accurately as qualified health professionals.

A selection of 2000 products were randomly sampled from a database of 96,141 products. Three Registered Dietitians regularly consulting patients with Food Allergy independently assessed the product information and then reached full consensus on each product's suitability. This formed the benchmark against which the Spoon Guru Machine Learning Model (SGML) was compared. Further product suitability assessments of the same 2000 products were conducted independently by five additional Registered Dietitians belonging to the British Dietetic Association (BDA), regularly consulting patients with Food Allergy, and three Clinical Allergists. The performance of each was compared to the benchmark.

The SGML was 99.3% accurate (CI: 0.75 +/- 0.38), which is on par with the highest level of accuracy achieved when a healthcare professional performed the product suitability assessments in this study (83.1; CI 16.90 +/- 1.64). The SGML model had the highest precision scores and made the lowest number of errors compared with the health professionals (SGML 99.8% precision, 15 errors vs average healthcare professional precision 90.5%, 183.6 errors).

The SGML tested can offer a robust way to screen thousands of food products and accurately determine those suitable and unsuitable for people with a nut allergy. Integration of such systems within clinical practice could enable health professionals to discuss a significant range of suitable products during their patient consultations. It could also offer individuals with nut allergy a robust, supportive tool when choosing suitable foods.

Introduction

There is an abundance of health-based apps available in the market today, with a significant number directed at the food and nutrition space (Hingle & Patrick, 2016). These technological tools span food intake trackers, food choice filters, wearables, self diagnostics, and mobile health management (Kao C-K & Liebovitz DM, 2017; Paglialonga et al., 2018). The incorporation of artificial intelligence (AI) within their design offers the opportunity for personalised, tailored health and lifestyle recommendations based on an individual's preferences and behavior.

As these applications move to AI integration within other platforms such as e-commerce and virtual home assistance, the potential to truly deliver personalised nutrition has never been greater. However, a critical consideration in this fast-paced innovative environment, that is beyond the essentials of user experience and personal data protection, is the paramount importance that these innovations are evidence based, appropriately validated and deliver proven efficacy (Paglialonga et al., 2018).

Currently, the majority of mHealth technologies are not designed with nutrition professionals input (Chen et al. 2017a). Considering the plethora of digital health offerings in the marketplace, there is a paucity of published, systematic assessments of their accuracy and health impact. A recent study has shown dietitians in New Zealand, Australia and the United Kingdom (UK) are using nutrition apps in practice but they are not currently an integral part of the nutrition care process (Chen et al. 2017b). This is unsurprising given the evidence-based approach to healthcare provision.

One area where there is an uncompromising need for accuracy, is in tailored food and health solutions for those with a food allergy. Most recent epidemiological data suggest food allergies are common, with up to 1 in 10 people affected. It appears that people in industrialised regions are disproportionately affected, with food allergies more common in children compared to adults. The foods that account for the most serious burden are peanut, tree nut, fish, shellfish, egg, milk, wheat, soy and seeds (Sicherer & Sampson 2018).

Fatal food anaphylaxis is rare, however the fear associated with such an event leads to some people with food allergy and their families living restricted lives (Umasunthar et al., 2013). The impact of food allergy on quality of life (QoL) goes far beyond simple avoidance of a couple of food items. Parents of children with food allergy had significantly lower overall quality of life than healthy non-food allergy comparison groups (Valentine 2011). In families with food allergies and intolerances, parents may limit many out of home activities, food shopping is often laborious, food choices limited, and social anxiety levels are often increased.
In December 2014 food allergen labelling laws in the European Union (EU) were updated. The law now states that if a pre-packaged food product contains one of the 14 major allergens outlined by the EU regulatory list, the product label must clearly embolden the allergen ingredient. Although the legislation has improved the level of information available to those with allergies, the time taken and stress associated with food selections has not been eased for individuals. Explicit food allergen labelling legislation has increased the level of allergen data available, this has enabled the development of a number of technologies that offer promise in this space. A brief survey by Venter C (2017) found that using the Spoon Guru mobile app made shopping easier for over 90% of those with food allergies or intolerances. However, although people are increasingly using apps and the number of apps in the allergy space are increasing, their quality of information has often been deemed to be poor (Cuervo-Pardo et al., 2015).

The purpose of this study was to validate whether machine learning (ML), a branch of AI, could accurately assess the suitability of foods for those with a nut allergy compared to health professionals. To note, this was not an assessment of whether the Spoon Guru ML model (SGML) could replicate the role of a Registered Dietitian or Clinician as these roles do not routinely involve the suitability assessment of such large volumes of product information at one time. Instead, the purpose of including these healthcare professionals in the present study was to provide a validated benchmark and comparator on which to assess the performance of the SGML.

Objective

The aim of this study was to assess whether the SGML can identify foods that are suitable or unsuitable for those with a nut allergy as accurately as qualified health professionals. The primary outcome was accuracy, secondary outcomes included performance precision and error classification.

Hypotheses

The hypothesis is that ML can perform equally well as humans with allergy expertise in this task of accurately identifying foods that are suitable or unsuitable for those with a nut allergy.

There will be no significant difference between healthcare professionals with allergy expertise and the SGML performance on this task.

Method

Product Database
A randomised selection of 2,000 products were sampled from a database of 96,141 products.

Product Suitability Benchmark
Three BDA Registered Dietitians, regularly consulting patients with Food Allergy, were provided with a spreadsheet that contained product information. This included the product name, the ingredients list and all of the on pack statements for each of the 2,000 products. Physical food labels were not used in this study. Based on the information on the spreadsheet each dietitian provided an independent classification of the products suitability for consumption by people with nut allergy on a dichotomous (yes/no) outcome. Following completion of the analysis by each of the three Dietitians, all disparities were discussed collectively amongst the Dietitians and full agreement was reached on whether the product in question was suitable or not suitable for those with a nut allergy. All 2,000 products were classified as either suitable or not suitable. The results from this were deemed to be the benchmark, the established ground truth, from which the performance of the product suitability assessment by health professionals and the machine model could be assessed.

Product Suitability Assessments by Health Professionals and the SGML

The SGML and five additional Registered Dietitians, regularly consulting patients with Food Allergy and members of the British Dietetic Association (BDA), and three Allergists individually completed the 2000 product suitability assessment spreadsheet. The purpose of this was to enable a comparison of suitability assessment performance between the SGML and health professionals.

Performance Measures

The accuracy, the proportion of responses that correctly identified a product as suitable or not suitable for those with a nut allergy, and precision, which reflects the correct detection of the important class (products not suitable for nut allergy) of the Dietitians, the Allergists and the SGML in determining the suitability of the products as defined by the product suitability benchmark were individually assessed using the Python Programming Language (version 3.6.3).

Error classification was also evaluated. This refers to whether the product was put in the right category, as defined by the Product Suitability Benchmark. There are four potential categories in this test;

1. Product correctly defined as not suitable
2. Product correctly defined as suitable
3. Products incorrectly defined as suitable (i.e. the product was not suitable, also known as a false positive)
4. Products incorrectly defined as unsuitable (i.e. the product was suitable, also known as false negative)

Statistical Analysis

The analysis was conducted by a Spoon Guru Data Scientist. The total number of errors made by each tester, compared with the ground truth, were calculated. The nature of each error detected was assessed to establish the level of false negatives (products incorrectly defined as unsuitable) and false positives (products incorrectly defined as suitable) for each tester. Differences between the health professionals and the SGML were assessed for significance using McNemar's test (McNemar, 1947).

Conflicts of interest

This study was funded by Spoon Guru.

Results

There were 59 discrepancies noted between the three Registered Dietitians after their individual assessments of the 2000 products. These were subsequently discussed, and full consensus was reached amongst the three dietitians on the suitability of each product. This established the accurate dataset/ benchmark from which the performance of the SGML was validated.

The accuracy, precision and error classification results for the eight product assessors (five Registered Dietitians and three Allergists) and the SGML compared with the validated benchmark, are shown in Table 1. 95% confidence intervals around the error for accuracy were calculated using Wilson Score Interval (1927).

- **Table 1:** Average results for the Health Professional product suitability assessments were calculated and are included in Table 1
- **Figure 1:** The distribution of accuracy and precision scores
- **Figure 2:** Illustrates the nature of the errors made by the SGML and the health professional with highest accuracy

Table 1. Accuracy and Error Classifications from Product Suitability Assessments performed by Health Professionals and Spoon Guru Machine Learning Model (SGML).

	Dietitian 1	Dietitian 2	Dietitian 3	Dietitian 4	Dietitian 5	Allergist 6	Allergist 7	Allergist 8	Healthcare Professional Average	SGML
Accuracy (%) [Error with 95% CI]	84.0 [16.00 +/- 1.61]	98.4 [1.60 +/- 0.55]	85.4 [14.65 +/- 1.55]	99.0 [1.05 +/- 0.45]	98.0 [2.00 +/- 0.61]	94.0 [6.05 +/- 1.04]	84.8 [15.20 +/- 1.57]	83.1 [16.90 +/- 1.64]	90.8 -	99.3 [0.75 +/- 0.38]
Precision (%)	82.9	98	84.2	98.8	97.8	96	83.6	83	90.5	99.8
Number of Errors made	320	32	293	21	40	121	304	338	183.6	15
False Negatives (% of 2000 products) [number of false negatives]	0.05% [1]	0.05% [1]	0.25% [5]	0.15% [3]	0.25% [5]	2.95% [59]	0.1% [2]	1.3% [26]	0.64% [12.75]	0.6% [12]
False Positives (% of 2000 products) [number of false positives]	15.9% [319]	1.55% [31]	14.4% [288]	0.9% [18]	1.75% [35]	3.1% [62]	15.1% [302]	15.6% [312]	8.54% [170.9]	0.15% [3]

Accuracy, precision and averages are calculated to 1dp

Figure 1. Box-plot comparing accuracy and precision of Health Professionals and the Spoon Guru Machine Learning Model (SGML) when classifying 2000 products as suitable or unsuitable for those with a nut allergy.

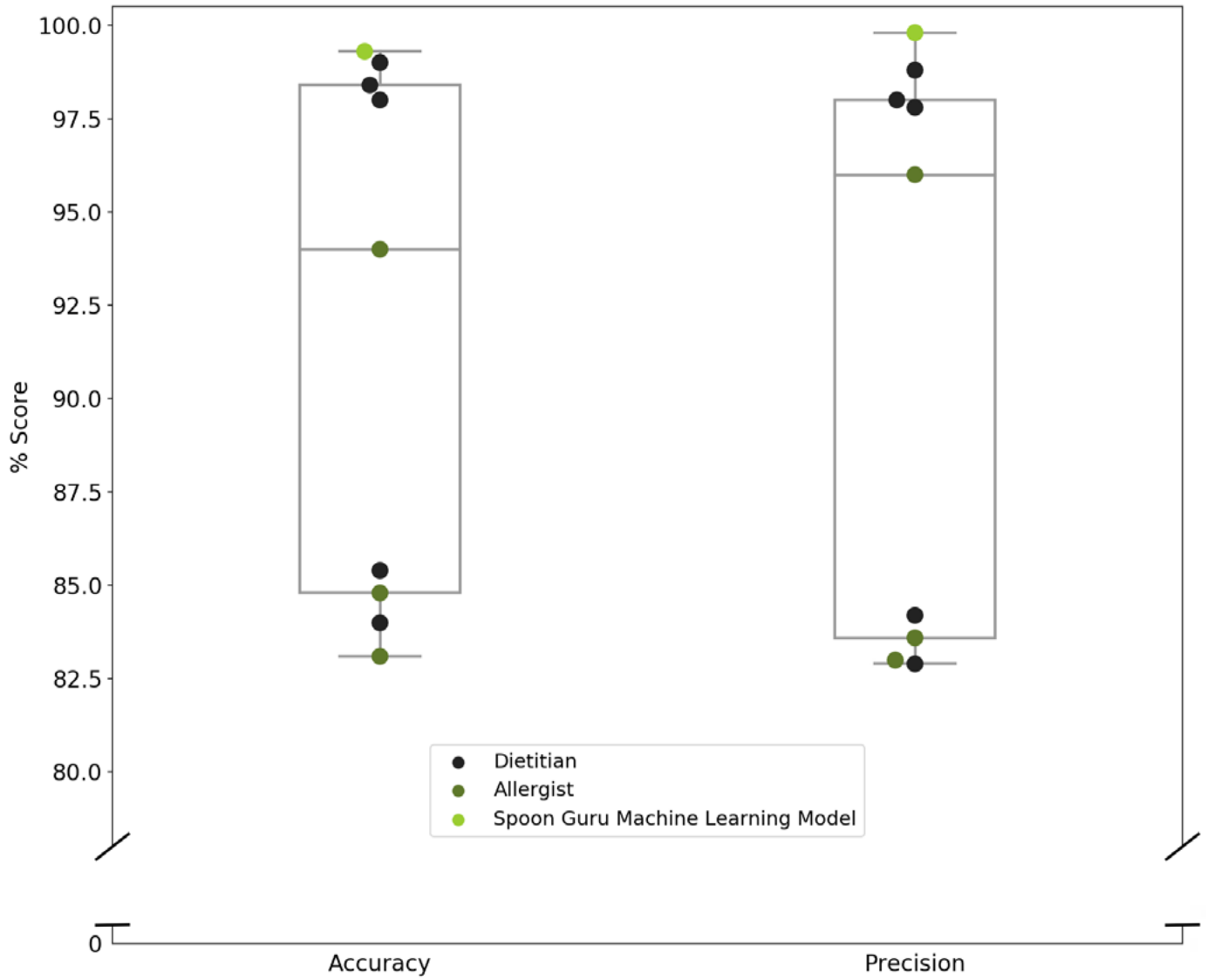
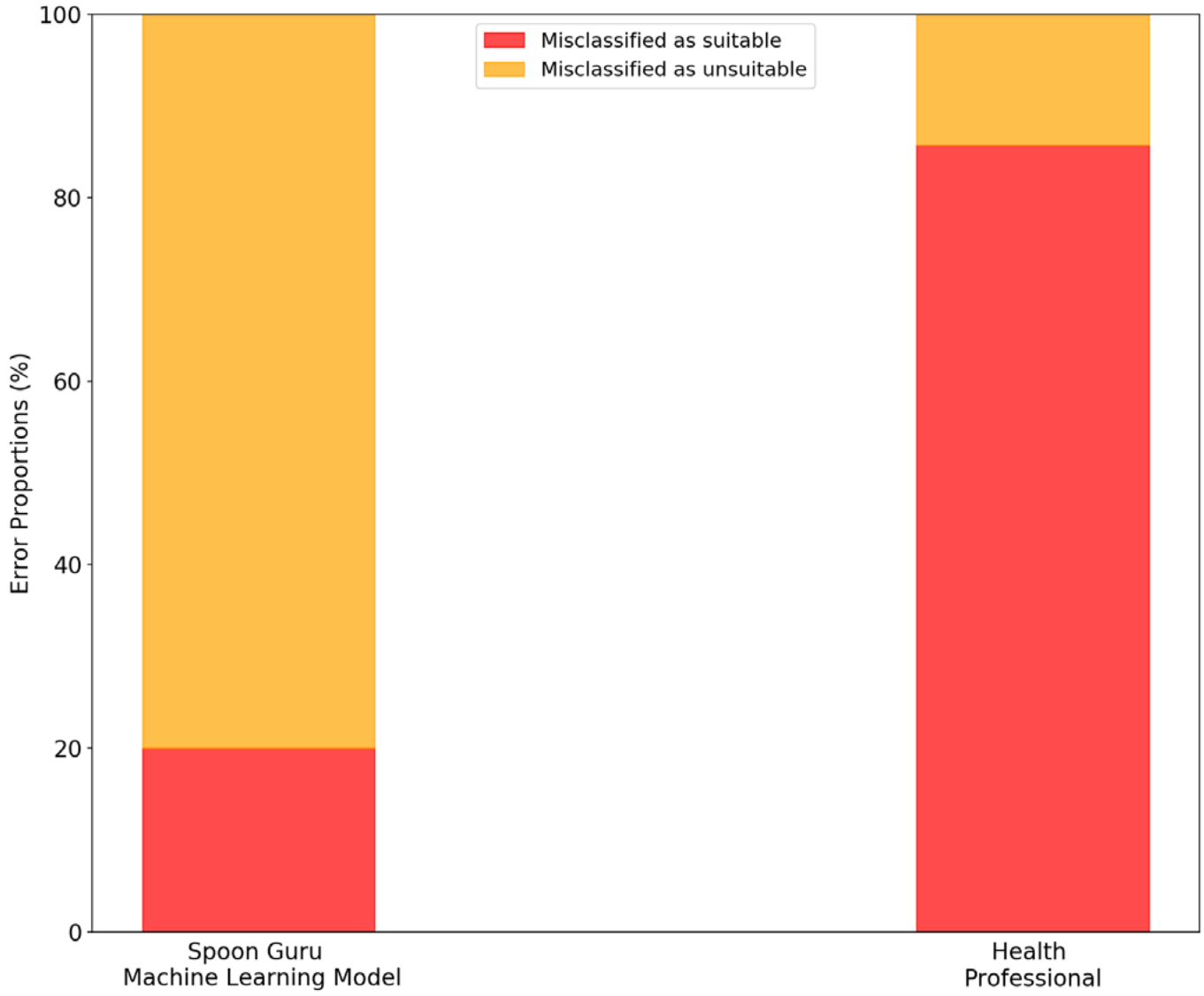


Figure 2. Nature of errors made by the most accurate Health Professional and the Spoon Guru Machine Learning Model (SGML) when classifying 2000 products as suitable or unsuitable for those with a nut allergy



The three products the SGML incorrectly defined as suitable for those with a nut allergy were as follows:

1. Indian Black Pepper Banana Chips
2. Almond Essence
3. Bakery Counter Comte

The eighteen products the most accurate dietitian incorrectly defined as suitable for those with a nut allergy were as follows;

1. Cadbury Twirl Bites
2. Cadbury Highlights Stick Pack Hazelnut
3. Cadbury Twirl multipack
4. Cadbury Twirl Bar
5. Cadbury Dairy Milk Ritz Sweet Biscuit
6. Cadbury Snow Bites
7. Indian Black Pepper Banana Chips
8. Milkybar dessert
9. Kit Kat Cookies and Cream biscuit
10. Smarties bar
11. Dipped waffle cones
12. Milk chocolate digestives
13. Cadbury Flakes
14. KitKat chunky milk chocolate bar
15. Aero chocolate bar
16. Munchies chocolate bar
17. Tunnocks milk chocolate wafers
18. Cadbury hot chocolate

McNemar's test (McNemar, 1947) was used to evaluate the null hypothesis that there were significant differences in performance between healthcare professionals and SGML. One of the health professionals was found to not significantly differ in performance from SGML. The SGML was the most precise and the most accurate assessor when compared to this benchmark method, and was found to be significantly different than six out of eight of the health professionals ($p<0.01$), and significantly different to the second most accurate health professional ($p<0.05$).

Discussion

The SGML was 99.3% accurate, which is on par with the highest level of accuracy achieved by a healthcare professional tested in this study (99.0%). The SGML was the most precise assessor of product suitability (99.8%), compared to 98.8% achieved by the most precise healthcare professional. It is not known how representative the performance of these health professionals is and whether they performed better, worse or average, in comparison with their peers.

Figure 1 highlights the inter-individual variation between the healthcare professionals in their product suitability assessments. They indicate a varying degree of human error when these professionals were tasked with an extensive, spreadsheet-based, product assessment that required a high level of manual data review. This error and variation is a likely consequence of the monotonous, repetitive nature of the product suitability assessment task. This error was also present in the panel of three dietitians that agreed the product suitability benchmark after their independent assessments. The 59 discrepancies identified amongst these three health professionals were resolved largely by the identification of a human error. Food labels under EU regulation must embolden the allergens within the ingredient list. The amalgamated spreadsheet did not embolden allergens in this way and this may also explain some of the errors made as usual practice would be to seek out the emboldened ingredients. It is not known if this level of variation is representative of the wider healthcare community. These findings are not a reflection on dietetic and clinical practice as this task is not one that is part of the day to day role of these healthcare professionals. Instead they reflect the challenges we as humans face when tasked to manually interpret a considerable level of product information data, even with such high level subject expertise. The results demonstrate the capacity and high accuracy and precision with which machines such as the SGML can perform such tasks. The SGML tested in the current study has the capacity to assess the suitability of 3000 products per hour, with the only limitation to the technology's speed being the processing power of the computer. This suggests ML technology could offer a robust and complementary tool in dietetic and clinical practice, increasing the food choices that health professionals can confidently discuss within their dietary counselling sessions and patient consultations.

A methodological difficulty is that there is no established reference standard of product suitability on the scale required to assess ML capabilities. Thus, inherent in this method of performance assessment is human error given the role of three health care professionals in setting the product suitability benchmark. Although consensus was reached in the definition of the benchmark, it may be their peers agree or disagree with their agreed suitability verdicts.

Figure 2 illustrates the nature of errors reported by the SGML, compared with the most accurate Health Professional product assessment. No product suitability assessment was 100% accurate. Given the potential health consequences of inaccuracies in this field, the nature of these inaccuracies was investigated further. The majority of the errors made by the health professionals were false positives, whereby products were selected as suitable for those with nut allergies when the ground truth stated they were not suitable. Conversely, the majority of errors the SGML made were false negatives, whereby the system determined products that would have been suitable for those with nut allergies as unsuitable. Although the latter limits the selection of foods determined as suitable for those with a nut allergy, the nature of this error has significantly less potential to result in an adverse health event compared with a false positive error. It is indicative of a risk averse approach within the design of the SGML.

Verbatim notes taken during the establishment of the ground truth by three Registered Dietitians highlight that the suitability decision reached for the three products the SGML incorrectly selected as suitable were based on information held outside of the product information provided in the dataset. [Indian Black Pepper Banana Chips - "In the warning it says can choke on nuts. I looked online and couldn't see anything online about them. I put not suitable from a just in case perspective. It may be their standard phrase so factory may contain nuts"; Almond Essence - "Anaphylaxis campaign website said to avoid essence/extract"; Compte Product - "Product information said for full allergen detail refer to store and can't do that from the information available, very vague of manufacturer and can't guarantee suitable."] These examples highlight that the three false positives detected by the SGML were classified as unsuitable by human logic and information available to the dietitians that was above and beyond the product information provided. This suggests, for maximal accuracy, it is a continual combination of ML and professional nutrition expertise in the development and delivery of such technological solutions that will drive the greatest accuracy and quality of output.

The eighteen false positive errors made by the most accurate dietitian included seventeen chocolate based products which included shea butter in the ingredients information. During the consensus discussion in the establishment of the ground truth it was agreed by the three dietitians that products containing shea butter should be classified as unsuitable, although during the discussion it was recognized that this decision erred very much on the side of caution, as the level of protein likely to remain in such a refined product was minimal. The other false positive was the Indian Black Pepper Banana Chips which were also misclassified by the SGML.

The data presented in this paper demonstrates accurate and validated technology exists that can enable extensive product suitability assessments at a volume that is perhaps beyond our usual human capacity. The opportunity exists for such technology to support the Dietitian or Clinician's work in the field, as they can offer a support tool to ensure a varied and simple approach to safe food selection that is available to patients. Dietitians want access to credible apps, recognizing the ability for these to streamline processes and enable them to spend more time on dietary counseling and negotiating patient goals for dietary and lifestyle behavior change, as well as functionality that offers tailored solutions to specific patient needs (Chen et al. 2017a). These results suggest that such needs can be robustly met by systems such as this SGML. The findings in the present study are similar to those in other healthcare fields and demonstrate that artificial intelligence systems can achieve performance on par with experts with demonstrable competence that is comparable to health professionals (Esteva et al., 2017; Burlina et al., 2017).

This study was not designed to compare dietetic or clinical performance to that of a machine, nor was it a test of knowledge. The key focus of the present study was to understand if AI can accurately assess an extensive set of product information, apply complex search terminology correctly and provide lists of suitable foods for both healthcare professionals and individuals, to assist in their management of food allergy. These results demonstrate that the SGML was able to successful perform such a task. The results highlight the difficulty of the manual application of such an extensive amount of product suitability assessments even if trained at a specialist level.

Conclusion

These results demonstrate that the AI based SGML can accurately assess the suitability of thousands of products for those with a nut allergy.

The performance of the SGML was on par with health professionals when assessing the suitability of foods for those with a nut allergy. None of the health professionals or the SGML assessments were 100% accurate. Fewer false positives were generated by the SGML, compared with the health professional assessments, indicative of a high level of risk management within its design.

Although a small study, the results demonstrate that AI can offer a robust way to screen thousands of food products and accurately determine those suitable and unsuitable for people with a nut allergy. Integration of such systems within dietetic practice could enable health professionals to discuss a significant range of suitable products during their dietary coaching sessions. It stands to offer great potential to strengthen and support the management of food hypersensitivity and could help facilitate Dietitians and other healthcare professionals in their employment of best practice.

No study to date has explored whether AI can accurately select products from large datasets that are suitable and unsuitable for those with nut allergy. This study is the first of its kind to validate the approach, finding it to be as accurate and more precise at this activity than a number of allergy specialists. The study design offers an example of a methodological approach to ensure the development of high quality technological innovations. It is important that for such accuracy to remain the SGML model needs to be continually updated with most recent product information.

A critical underpinning to the integration of AI within dietetic or clinical practice and personalised nutrition must be evidence based, expert driven inputs and validated outputs. This is essential to ensure robust, positive, meaningful impacts on health. Failure to evaluate the accuracy underpinning such innovations could compromise user health and safety (Stoyanov et al.,2015). This will also ensure user trust, be that individual or healthcare professional, is maintained and the full potential of this artificial intelligence in scientific innovation is realised.

Future work should comparatively assess and validate the various AI based solutions that are increasingly available in this space. This is required to ensure the identification of robust offerings amongst the plethora of those available and to generate an independent approach to their quality assessment.

References

Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM (2017) Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. Computers in Biology and Medicine, volume 82, pages 80-86.

Chen J, Lieffers J, Bauman A, Hanning R, Allman-Farinelli M (2017a) Designing health apps to support dietetic professional practice and their patients: qualitative results from an international survey. JMIR UHealth 5(3): e40.

Chen J, Lieffers J, Bauman A, Hanning R, Alllman-Farinelli (2017b) The use of smartphone health apps and other mobile (mHealth) technologies in dietetic practice: a three country study. Journal of Human Nutrition and Dietetic Practice Vol 30, Issue 4 pages 439 – 452.

Cuervo-Pardo L, Barcena-Blanch MA, Gonzalez-Estraa A, Schroer B (2015) Apps for food allergy: A critical assessment The Journal of Allergy and Clinical Immunology: IN Practices, Volume 3, Issue 6, Nov-Dec 2015, Pages 980-981.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 524, pages 115-118.

Hingle M & Patrick H (2016) There are thousands of Apps for that: Navigating mobile technology for Nutrition Education and Behaviour. Journal of Nutrition Education and Behaviour, Volume 48, Issue 3, pages 213-218e1.

Kao C-K & Liebovitz DM (2017) Consumer mobile health apps: Current state, barriers and Future Directions. PM & R, Volume 9, Issue 5, S106-S115.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153-157.

Paglialonga A, Luga A, Santoro E (2018) An overview on the emerging area of identification, characterization and assessment of health apps. Journal of Biomedical Informatics, Volume 83, pages 97-102.

Sicherer SH, Sampson HA (2018) Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention and management. Journal of Allergy and Clinical Immunology, Vol 141 Issue 1 pages 41-58.

Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegroro D, Mani M (2015) Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth Volume 3 (1), e27.

Umasunthar T, Leonardi-Bee J, Hodes M, Turner PJ, Gore C, Habibi P, Warner JO, Boyle RJ (2013) Incidence of fatal food anaphylaxis in people with food allergy: a systematic review and meta-analysis. Clinical and Experimental Allergy, Volume 43, Issue 12 pages 1333- 1341.

Venter C (2017) Smartphone applications: are they a help or hindrance? Network Health Digest-EXTRA: Research and Resources. November 2017, issue 129, pages 58-60.

Machine Learning Software References:
Analysis carried out in Python:
Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org

Analysis supported by following packages:
Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

Visualisations:
John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

Spoon Guru®

www.spoon.guru

hello@spoon.guru