PERSON—Personalized Expert Recommendation System for Optimized Nutrition

Chih-Han Chen, *Student Member, IEEE*, Maria Karvela, Mohammadreza Sohbati, *Member, IEEE*,
Thaksin Shinawatra, and Christofer Toumazou, *Fellow, IEEE*

***Abstract*—The rise of personalized diets is due to the emergence of nutrigenetics and genetic tests services. However, the recommendation system is far from mature to provide personalized food suggestion to consumers for daily usage. The main barrier of connecting genetic information to personalized diets is the complexity of data and the scalability of the applied systems. Aiming to cross such barriers and provide direct applications, a personalized expert recommendation system for optimized nutrition is introduced in this paper, which performs direct to consumer personalized grocery product filtering and recommendation. Deep learning neural network model is applied to achieve automatic product categorization. The ability of scaling with unknown new data is achieved through the generalized representation of word embedding. Furthermore, the categorized products are filtered with a model based on individual genetic data with associated phenotypic information and a case study with databases from three different sources is carried out to confirm the system.**

***Index Terms*—Expert system, recommendation system, personalized diets, deep learning, grocery decisions, neural networks, genetic algorithm.**

I. Introduction

EXPERT systems have been expected as a solution for better production quality and diet choices in the food industry since 1998 [1]. Together with the advancement of Big data technology, systems that can cope with high variety, volume and velocity bring success to business [2], science and health-care [3]. A recommendation system aiming to scale with big data through Map-Reduce has led to the birth of many applications, such as product recommendation [4], music recommendation and health recommendation [5]. In this paper, we have proposed a system combining existing algorithms to construct a practical application for grocery product recommendation. As shown in Fig. 1, the included algorithms cover the task of data categorisation, data analysis and decision recommendation. Deep Neural Network (DNN) model is applied to deal with data categorisation due to its outstanding performance on the complexity of data with noise. Furthermore, A novel model is invented to cover data analysis, while decision recommendation is optimised with Genetic Algorithm.

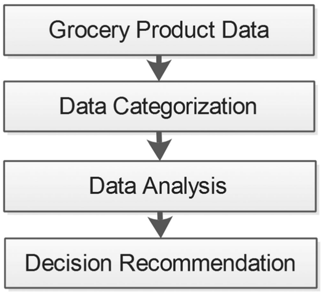


Fig. 1. Expert system overview.

The scalability of the models within all systems requires the ability to deal with unknown data [6], which is typically measured through model generalisation. In the area of health-care, the introduction of deep learning analysis, or particularly DNNs is believed to be the reason why recommendation expert systems improved over recent years [5]. Neural networks aiming to simulate neuron behavior, started from utilizing existing devices with the development of programming frameworks, evolving into circuit implementation. Technology companies such as Microsoft and Google have published their open source research tools with the ability to commendably satisfy the simulation of neural networks on CPUs or GPUs. They allow researchers to focus more on designing network structures and applications [7]. Due to the needs of higher accuracy and better performances while solving problems occurring during scaling with neuron sizes, the research on neural network systems has expended to FPGA programming and expected to be fully implemented in circuits down to device level [8]. Some recent research publications have shown such interest in expansion, including memristor-based neural networks solving finite time synchronisation problems or neural-type neural networks finite time stability problems [9], [10].

From the application perspective, personalised health-care and diets have gained great attention with the rise of nutrigenetics studies. It is believed that all molecules involved in human metabolism are controlled directly or indirectly by genes and therefore the health of individuals can be optimised through personalized dietary advice [11], [12]. Direct-to-consumer

Manuscript received March 31, 2017; revised August 3, 2017; accepted September 20, 2017. Date of publication October 30, 2017; date of current version January 26, 2018. This work was supported by the DNAnudge Ltd. This paper was recommended by Associate Editor J. Lü. *(Corresponding author: Chih-Han Chen.)*

C.-H. Chen, M. Sohbati, and C. Toumazou are with the Centre for Bio-Inspired Technology and the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: chih-han.chen13@imperial.ac.uk; m.sohbati@imperial.ac.uk; c.toumazou@imperial.ac.uk).

M. Karvela and T. Shinawatra are with the DNAnudge, London SW3 1JJ, U.K. (e-mail: maria@dnanudge.com; dr.thaksin@dnanudge.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBCAS.2017.2760504

(DTC) genetic services have emerged [13] with the invention of cost effective genetic detection tests [14], which led to the high expectation of relevant applications. Some surveys show up to 50% of subjects willing to receive genetic tests and follow the personalised suggestions [15]. However, most of the services provide complicated information to the consumers and lead to the gap between the personal genetic information and adherence to a personalised diet. In order to provide meaningful and straight-forward solutions, a novel expert system is proposed in this paper, which aims to personalize grocery shopping through product suggestions based on genetic phenotype.

Other research work on systems for dietary decisions and personalisation can be found in some recent publications. For example, a system that also targets on providing decision support for food [16]. However, instead of using genetics as input, they focus on the temporal decision for diets through optimizing reasoning. Another similar topic is the so called medical education system, which provides personalised suggestions through individual characteristics and health objectives [17]. In our work, we focus on the temporal decision of comparison between a grocery product and all other products within the same food category. Furthermore, we aim to perform personalisation through correlating the nutrition facts of grocery products and generating recommendations. The system is built based with the aim of processing two datasets: the genetic data and the grocery data. The genetic data relating to phenotypes influencing the consumption ability of five main nutrition factors are energy, fat, protein, sugar and salt. They are used to perform simulations of all possible combinations. The genetic data collection of any customer is static and assumed to remain the same over time. On the other hand, the grocery data is designed to be dynamic and varied with continuous data collection. In order to combine and correlate the two datasets, a framework covering the process of data collection, data categorisation with deep neural network and the grocery product recommendation generation is proposed.

The proposed expert recommendation system framework consists of four main components:

1) Word Embedding & Padding model is introduced for converting text data into generalised vectors to deal with data variety, unknown grocery product names and mitigate the out-of-vocabulary problem.
2) A DNN model for product categorisation is covered to cope with the complex features and sequence logics of the grocery product vectors. The selected model can perform categorisation task with good accuracy.
3) Decision recommendation model is designed to take the categorisation results,analyze nutritional data and optimize the suggestion provided to the consumer with the designed fitness score and Genetic Algorithm (GA).
4) An operational state machine is included with the aim to control the state and retrain models of each components during updates of training data.

In this paper, we will first describe the word embedding method that is applied in Section II. A DNN architecture, called Deep Long Short Term Memory Recurrent Neural Network (DLSTM-RNN) will be covered in Section III. In Section IV, the recommendation model is described to provide grocery shopping decisions considering behavior, personal phenotype data and nutrient value based on different food category. The state machine with an interface for human input is presented in Section V, and finally a case study is introduced with applications of the proposed expert system on 3 different databases in Section VI. The contribution of this paper includes, the novel architecture of the personalised expert recommendation system for optimised nutrition (PERSON) based on genetic phenotype, the utilisation of a deep neural network for grocery product categorisation, and implementation of a recommendation system using Genetic Algorithm.

II. Word Embedding & Padding

Word Embedding is a distributed representation of words in numerical vectors and it has demonstrated outstanding performance on word similarity tasks and word analogy/analysis in deep learning studies [18]. General method of bag-of-word (BoW) model has proven to be insufficient for the use of classifying short texts [19], since BoW ignores the order and semantic relations between words. More advanced models such as neural network language word embedding, are shown to preserve semantic type and syntactic relationships. Furthermore, embedding words into vectors with real numbers from one dimension per word to a continuous vector space ensures the low dimensionality of vectors [20]. The vectors of words in the embedding space have the property of similar semantic relationships [18], which allow the advancement of clustering tasks in Natural Language Processing (NLP). In this paper, the continuous Skip-gram model is used to generate word embedding, which consists of an input, projection and an output layer. The input layer observes each current word and predicts a certain range of words located before and after. The increase of range improves the quality of word embedding but causes a trade off at the expense of computation load.

The skip-gram word embedding model is trained with 100 billion words from Google News. The word vector has 300 dimensions and covers three million unique words [20]. With dimensionality reduction techniques such as t-distributed stochastic neighbor embedding (t-SNE), the word vector can be projected down to two dimensions for the purpose of visualisation [21], as shown in Fig. 2. The figure shows words of the case study data of this project. Name entities with similar semantic type can potentially cluster and create different regions for classifications.

III. Deep Neural Network for Product Categorisation

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have emerged as two widely used architectures in deep neural networks (DNNs). CNNs have excellent performance in extracting features of different positions within the data and they can learn the relationships between positions and features through an operation called max-pooling [22]. However, they are not designed to handle sequences of the data or to capture long-term dependencies. On the other hand, RNNs are specialised for sequential modeling, which are ideal for learning

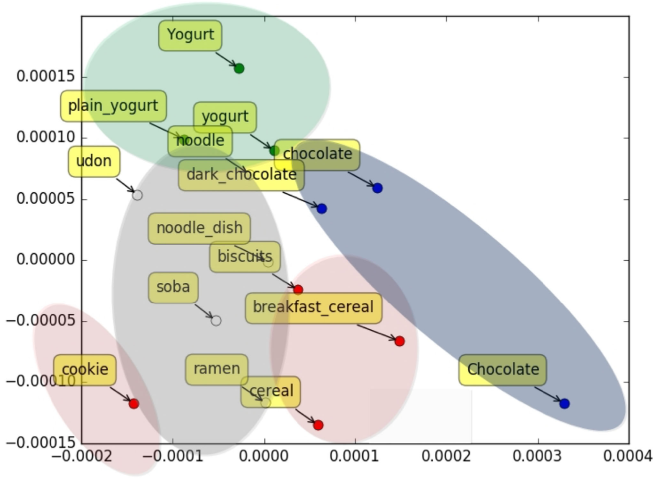
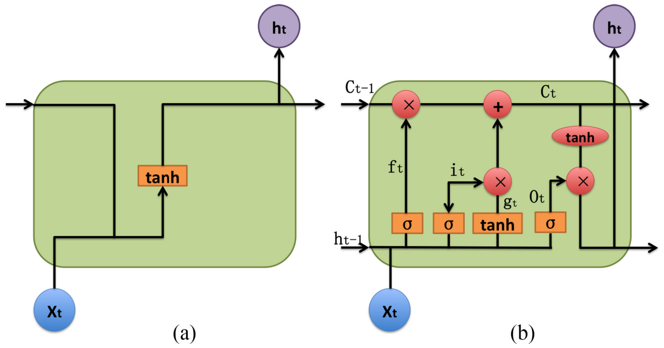


Fig. 2. The 2 dimensional plot of product name vector.



(a)

(b)

Fig. 3. The internal structure of (a) RNN and (b) LSTM.

the word dependencies [23]. In this project, one type of RNNs is applied due to the significant importance of dependency between words in product categorisation.

All traditional Recurrent Neural Networks have simple structures such as a single hyperbolic tangent layer shown in Fig. 3(a). In the standard RNN model with a long chain of repeating components, vanishing gradients for the parameter updates becomes a serious issue for training, which leads to the introduction of Long short-term memory recurrent neural network (LSTM-RNN) [24]. The center of LSTM cell is the memory sub-cell, the information can be removed or added to the cell state through the designed gates, as shown in Fig. 3(b). The operation of a single LSTM cell is described in (1)–(6).

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{1}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{2}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{3}$$

In the above (1)–(3), $\mathbf{x}_t$ is the input vector of current time step t to the LSTM-RNN cell, while $\mathbf{h}_{t-1}$ is the vector of the previous hidden layer state. These two vectors are put together and weighted by the forget gate weight matrix $\mathbf{W}_f$, input gate weight matrix $\mathbf{W}_i$ and output gate weight matrix $\mathbf{W}_o$.The weighted vectors are summed up with the biases of the same gate type, therefore $\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t]$ is summed with the forget gate bias $\mathbf{b}_f$, $\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t]$ is summed with the input gate bias $\mathbf{b}_i$ and $\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t]$ is summed with the output gate bias vector $\mathbf{b}_o$. The summed results are taken as inputs to sigmoid activation functions and form the outputs of the three gates: forget gate output vector $\mathbf{f}_t$, input gate output vector $\mathbf{i}_t$ and output gate output vector $\mathbf{o}_t$. In another words, each gate is formed with a sigmoid layer and a point-wise multiplication operator that can output a value from 0 to 1. The decision of passing the information is decided based on the output value.

$$\mathbf{g}_t = \tan h(\mathbf{W}_g[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_g) \tag{4}$$

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tan h(\mathbf{C}_t) \tag{6}$$

The actual input information is described in (4). Similar to the (1)–(3), $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$ are put together and weighted by the input weight matrix $\mathbf{W}_g$. Hyperbolic tangent activation function takes the sum of weighted result and input bias $\mathbf{b}_g$ and generates input state $\mathbf{g}_t$. The Internal current state coefficient $\mathbf{C}_t$ is evaluated through the sum of forget gate output vector $\mathbf{f}_t$ multiplied by the previous state $\mathbf{C}_{t-1}$ and the input gate output vector $\mathbf{i}_t$ multiplied by the input state $\mathbf{g}_t$, see (5). Finally, the hidden layer output state $\mathbf{h}_t$ is produced by the activation of $\mathbf{C}_t$ multiplied by the output gate vector $\mathbf{o}_t$ in (6). Note that $\mathbf{h}_t$ is treated as the output of LSTM-RNN cell as well as the recurrent input of the same cell in the next time step.

In this paper, LSTM is used to capture the semantic dependency of product names for classification of categories. The nature of product names is complex in terms of the various importance of different sequences. For instance, dark chocolate is a type of chocolate with supporting information dark, while chocolate biscuit is a type of biscuit with supporting information chocolate. The word chocolate represents a very different type of information in different scenarios, indicating that a different sequence causes a word to play a very different role in semantic classification.

Single layer of LSTM RNN already has deep architectures, since it is treated as a feed-forward neural network with the recurrent connection that loops multiple layers in time, with each layer sharing the same parameters. However, the single layer has limited features due to the fact that the information is processed by a single nonlinear layer through multiple given time instants before contributing the output. Deep LSTM RNNs are formed with many layers of LSTM. The Deep LSTM RNNs have both properties of processing through time and layers, where the parameters more optimally used via the distribution of the meaning over space of the multiple layers [25], [26].

Fig. 4 shows the architecture of Deep LSTM (DLSTM) model of this project, the input data are converted into word vectors and padded into the maximum length of all training data. At each time instance one word vector is fed in the input $\mathbf{x}_i$ of the DLSTM, where i is the sequence of time instance. LSTM cells output values to hidden layer $\mathbf{h}_j$ at each layer of DLSTM, where j is the number of layers of LSTM RNN. Finally, the last hidden layer $\mathbf{h}_j$ is connected to a soft-max layer which is expressed as (8), where $\mathbf{u}_g$ is the output of the fully connected layer before soft-max function, G is the number of prediction

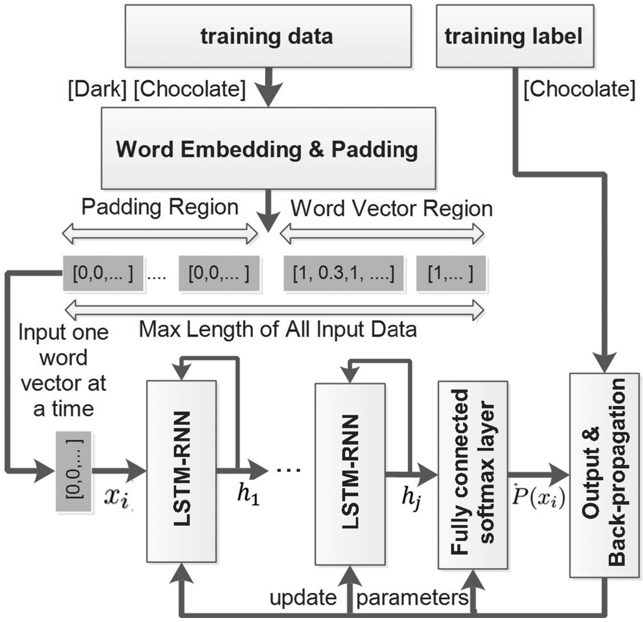


Fig. 4. The architecture of DLSTM model.

of the classification label.

$$\mathbf{h}_j \odot \mathbf{W}_s = \mathbf{u}_g \tag{7}$$

$$P(\mathbf{x}_i) = \frac{e^{\mathbf{ug}}}{\sum_{g=1}^{G} e^{\mathbf{ug}}} \tag{8}$$

$$\mathbf{l} = -\frac{1}{n}\sum_{i}^{N}\sum_{k}^{C} P_k^d(\mathbf{x}_i)\log(P_k(\mathbf{x}_i)) \tag{9}$$

The LSTM model is trained with cross-entropy over the training data. In (9), $\mathbf{l}$ represents the loss vector, $P_k(\mathbf{x}_i)$ is used to denote the distribution of the model prediction, and $P_k^d(\mathbf{x}_i)$ is the distribution of the inputs. N is used to describe the total number of training data and C is the number of classification labels. The Adam method is used to achieve the stochastic optimisation and update all parameters in each layer of DLSTM-RNN and the soft-max layer [27].

IV. Decision Recommendation Model With Genetic Algorithm

Having categorised food products and stored their nutritional information, the second part of the project is on what to recommend based on a user's genotype from the products that are now in the same standardised category. To do so, a threshold matrix is considered which sets nutritional value thresholds. These thresholds are set based on phenotypic information that each corresponds to genotype; basically, the user's DNA. In this project, the personal data is expressed as a nutrition threshold matrix $\mathbf{T}_{l,n}$, that maps to different genetic phenotypes for metabolism. The representation of the value in the matrix can be treated as nutritional value boundaries, where n represents the different nutrients, that can vary from 1 to M. In this project $M = 5$, representing energy, fat, salt, sugar and protein. l is the level of the boundary of n nutrients. l can vary from 1 to L which in the case of this study is 3 representing High, Medium and Low threshold. The mapping of Phenotype data to the threshold matrix $\mathbf{T_{1,n}}$ is described in Section IV-A. The $\mathbf{T_{1,n}}$ is treated as the input variables to the genetic algorithm for optimizing recommendations based on a designed fitness score. The recommendation varies with a process of product filtering described in Section IV-B. The fitness score is described in Sections IV-C and IV-D, which is designed to observe the behavior differences between food choices that follows given recommendations and choices by random decisions. The aim of applying the genetic algorithm is to optimize the grocery shopping decisions by maximizing the differences between consumption of nutrition before and after receiving recommendation. The detail of the overall procedure and fitness score evaluation are described in Section IV-E.

TABLE I
An Example of the Genotypes of rs5400

Genotype	Magnitude	Summary
(C; C)	0.1	Normal sugar consumption
(C; T)	1.7	Significantly higher sugar consumption
(T; T)	1.8	Significantly higher sugar consumption

A. Introduction to Phenotype Data

A phenotype is the observable characteristic of an organism in biology that is caused by the expression of the genetic code, or genotype, and environmental factors [28]. The distinction of genotype-phenotype is first proposed by Wohelm in 1911 [29], and later discussed by many researches [30], [31]. In general the relationship can be expressed as (10):

$$\mathbf{G} + \mathbf{E} + \mathbf{GE} \approx \mathbf{P}, \tag{10}$$

where $\mathbf{G}$ is the genotype, $\mathbf{P}$ is the phenotype, $\mathbf{E}$ represents environmental factor, $\mathbf{GE}$ represents the interaction between the genotype and the environment. Genotype data can be found in SNPedia with the reference to papers sourced from Pubmed, where new research results and publications are continuously updated [32]. Table I shows an example of the genotypes of rs5400 from SNPedia. The genotype rs5400 is a gene location and three variation with different expressions that are associated with sugar consumption.

Environmental factors are defined differently in various practical situations but usually directly related to the genotype. For example if our genotype is focusing on sugar, we can consider taking individual sugar intake as the environmental factor. In most practical cases, genetic phenotype is defined as the observable characteristic of performance of a persons ability. Therefore in this paper we consider the ability to consume the five selected nutrition factors: energy, fat, salt, sugar and protein. Inspired by the three expressions of genotypes shown in Table I, we design the phenotypes as having three ability levels: good, medium and bad, hence we simplify our equation to a direct mapping $\mathbf{G} \approx \mathbf{P}$. While aiming to actually link this relationship to the grocery data, we design the threshold table $\mathbf{T_{1,n}}$ to express the strength of ability to consume each nutrient. In the real-world environment, a persons $\mathbf{G}$ is defined by the outcome pattern of a genetic test, and $\mathbf{P}$ is mapped to $\mathbf{G}$, while $\mathbf{P}$ has a direct one

TABLE II
DIRECT MAPPING OF P TO $\mathbf{T}_{(\mathbf{1},\mathbf{n})}$ OF A GROCERY CATEGORY

	$\mathbf{n}$ = Energy	$\mathbf{n}$ = Fat	$\mathbf{n}$ = Salt	$\mathbf{n}$ = Sugar	$\mathbf{n}$ = Protein
l = Good	$\mathbf{P}_{\mathbf{Good,Energy}} \to \mathbf{T}_{\mathbf{High,Energy}}$	$\mathbf{P}_{\mathbf{Good,Fat}} \to \mathbf{T}_{\mathbf{High,Fat}}$	$\mathbf{P}_{\mathbf{Good,Salt}} \to \mathbf{T}_{\mathbf{High,Salt}}$	$\mathbf{P}_{\mathbf{Good,Sugar}} \to \mathbf{T}_{\mathbf{High,Sugar}}$	$\mathbf{P}_{\mathbf{Good,Protein}} \to \mathbf{T}_{\mathbf{High,Protein}}$
l = Medium	$\mathbf{P}_{\mathbf{Medium,Energy}} \to \mathbf{T}_{\mathbf{Medium,Energy}}$	$\mathbf{P}_{\mathbf{Medium,Fat}} \to \mathbf{T}_{\mathbf{Medium,Fat}}$	$\mathbf{P}_{\mathbf{Medium,Salt}} \to \mathbf{T}_{\mathbf{Medium,Salt}}$	$\mathbf{P}_{\mathbf{Medium,Sugar}} \to \mathbf{T}_{\mathbf{Medium,Sugar}}$	$\mathbf{P}_{\mathbf{Medium,Protein}} \to \mathbf{T}_{\mathbf{Medium,Protein}}$
l = Bad	$\mathbf{P}_{\mathbf{Bad,Energy}} \to \mathbf{T}_{\mathbf{Low,Energy}}$	$\mathbf{P}_{\mathbf{Bad,Fat}} \to \mathbf{T}_{\mathbf{Low,Fat}}$	$\mathbf{P}_{\mathbf{Bad,Salt}} \to \mathbf{T}_{\mathbf{Low,Salt}}$	$\mathbf{P}_{\mathbf{Bad,Sugar}} \to \mathbf{T}_{\mathbf{Low,Sugar}}$	$\mathbf{P}_{\mathbf{Bad,Protein}} \to \mathbf{T}_{\mathbf{Low,Protein}}$

to one mapping to $\mathbf{T}_{\mathbf{l},\mathbf{n}}$. Such relation is shown in Table II. The extension of adding E and GE is simply creating more threshold types, extending Table II and will be covered in future work.

Table II presents the mapping from $\mathbf{P}$ to $\mathbf{T}_{\mathbf{l},\mathbf{n}}$. In this paper we treat $\mathbf{T}_{\mathbf{l},\mathbf{n}}$ as a matrix of variables. These variables get their values through optimizing the fitness score of the model with the genetic algorithm described in Sections IV-B–IV-E. $\mathbf{T}_{\mathbf{l},\mathbf{n}}$ is treated as the target variables that changes our recommendations. Therefore the action of filtering different a number of products, increases or decreases the probability of choosing a product within a category. With such a relationship, we have already linked the genetic test result $\mathbf{G}$ to the product recommendations through the variation of filtering thresholds.

B. Product Filtering & Suggestions

The input of the filter $\mathbf{p}_n^{i_c}$ is the Product nutrient value, where i_c is the product identity number of the grocery category c, which varies from 1 to the number of the total product I_c of category c.

$$H_{l,n}^c(P_n^{i_c}) = \begin{cases} 1 & \text{if } \mathbf{p}_n^{i_c} \leq \mathbf{T}_{l,n}^c & (11a) \\ 0 & \text{if } \mathbf{p}_n^{i_c} > \mathbf{T}_{l,n}^c & (11b) \end{cases}$$

The filter function is low pass, meaning the filter outputs 1 when the threshold value $\mathbf{T}_{l,n}^c$ is larger than the Product Nutritional Value $\mathbf{p}_n^{i_c}$ and outputs 0 when the opposite condition occurs.

The Decision of suggesting a product is based on the function D, see (12). The output of the decision is either one or zero, which is the multiplication in series of the outputs of filters. This procedure can be imagined as a product passing through all layers of filters and receiving a decision of one when satisfying the condition that all the nutritional factors are lower than the assigned thresholds to a person.

$$D_l(\mathbf{p}_n^{i_c}) = \prod_{n=1}^{\mathrm{M}} H_{l,n}^C(\mathbf{p}_n^{i_c}) \tag{12}$$

C. Nutrients Consumption

Scanning through all the products, the number of the suggested products within a category is expressed in (13), which is the sum of all the $D_l(\mathbf{p}_n^{i_c})$ of products within category c.

$$s_l^c = \sum_{i_c=1}^{I_c} D_l(\mathbf{p}_n^{i_c}) \tag{13}$$

Now the average of the nutritional value of the suggested products $a_1^{n,l}$, can be obtained with (14) and the average of all nutritional value of all products a_2^n is obtained with (15)

$$a_1^{n,l} = \sum_{i_c=1}^{I_c} \frac{\mathbf{p}_n^{i_c} \odot D(\mathbf{p}_n^{i_c})}{s_l^c} \tag{14}$$

$$a_2^n = \sum_{i_c=1}^{I_c} \frac{\mathbf{p}_n^{i_c}}{I_c} \tag{15}$$

D. Behavior of Suggestion Follower

In order to achieve more realistic experiments, a simple behavior model with a Gaussian function is introduced. This model is used to describe the probability of a person following the suggestion if he/she is given the suggestion from the system, see (16). It is assumed that the probability of the person following suggestion is at its maximum value when half of the product is recommended. When the number of the product recommendation is above 50% of the total products, the probability decreases, due to the psychological doubt that the system does not filter any products for the person. On the other hand, the probability also decreases when the number of the products recommendation is below 50%, because of the frustration of having fewer options. Note that this assumption of behavior can be changed to fit the real data if a survey is done on the topic.

$$Q_l^c = \mathrm{f}\left(\frac{s_l^c}{I_c}\right) = 1 - 0.5 \times e^{\frac{\left(\frac{s_l^c}{I_c} - 0.5\right)^2}{4.6^2}} \tag{16}$$

E. Fitness Score & Genetic Algorithm

Finally, there are two scenarios that are compared. The first scenario is the sum of the consumption of the nutrients with the product recommendation and the consumption of the nutrients without the recommendation from the system. This can be obtained through the sum of the probability of following the suggestion multiplied by the average nutrition value of the suggested products and the remaining probability multiplied by the average nutrition value of all products. The second scenario is simply the average of nutrition values of all products, see (17) and (18).

$$v_1^{n,l} = a_1^{n,l} \times Q_l^c + a_2^n \times (1 - Q_l^c) \tag{17}$$

$$v_2^n = a_2^n \tag{18}$$

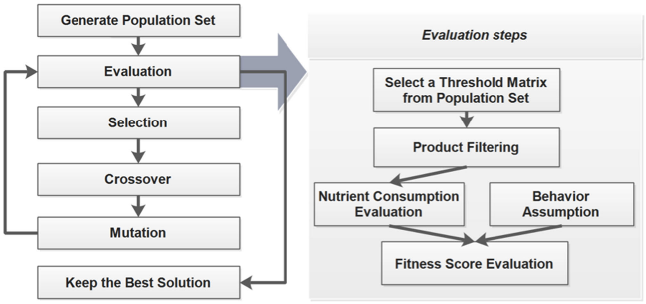


Fig. 5. The procedure of recommendation optimisation with GA.

The difference of the two scenario is the reduced consumption $r_{n,l}$ of the nutrition n with the nutrition threshold boundary l, see (19) and (20).

$$r_{n,l} = v_2^n - v_1^{n,l} \tag{19}$$

$$r_n = \sum_{l=1}^{L} \frac{r_{n,l}}{L} \tag{20}$$

With the aim of considering all nutrients, for $n = 1, 2, \ldots, 5$, or in other words, energy, fat, salt, sugar and protein, the fitness score f is measured based on the sum of the normalised form of all the reduced consumption of the nutrition z_n, see (21) and (22).

$$z_n = \frac{r_n}{a_2^n} \tag{21}$$

$$f = \sum_{n=1}^{M} z_n \tag{22}$$

The fitness score f is designed as a measurement of the difference in nutritional intake average after receiving recommendation, with the aim of maximizing this score. Genetic Algorithm (GA) is applied due to its outstanding performance for stochastic optimisation. The GA is inspired by the process of natural selection, which is commonly applied to generate high quality optimisation solutions. There are three main bio-inspired operations: mutation, crossover and selection. The procedure of recommendation optimisation with the genetic algorithm is shown on the left side of Fig. 5. In the step of population generation, random sets of threshold matrices are initialised. The generated threshold matrices are fed one by one to the evaluation step, where the recommendation model proposed in Sections IV-A–IV-D is applied as shown on the right side of Fig. 5. The evaluation steps include selecting one of the sample threshold matrix from the population sets, product filtering with (11)–(12), nutrient consumption evaluation following (13)–(15) and fitness score evaluation with the behavior assumption described in (16)–(22). Once the fitness scores are evaluated for all threshold matrices in the population sets, these scores are compared and the best few sets of the threshold matrices that generate higher fitness scores are selected. The non-selected threshold matrices are deleted from the population sets. In the step of crossover, random pairs of the threshold matrices from the selected threshold matrices are formed and some parameters of randomly selected locations are swapped between pairs. Furthermore, some locations of the selected sets of threshold matrices are modified with a low probability in the step of mutation. Finally, the new modified population sets of threshold matrices are sent back to the evaluation step. In this project package GA is applied to achieve optimisation tasks [33], [34]. GA is set up as the algorithm to maximize the fitness score f of the model with the input variable $\mathbf{T}_{l,n}^{C}$ of grocery category C under constrain that $\mathbf{T}_{1,n}^{C} > \mathbf{T}_{2,n}^{C} > \mathbf{T}_{3,n}^{C}$ or in other words $\mathbf{T}_{\text{High},n}^{C} > \mathbf{T}_{\text{Medium},n}^{C} > \mathbf{T}_{\text{Low},n}^{C}$.

V. Expert Recommendation System Architecture and Application

In this section, the architecture is first introduced and the operation during practical application is then described. The top level block diagram of the system is shown in Fig. 6. The whole system operates in two states; training state and recommendation provider state. Each block operates individually as:

1) Input New Product Data Buffer: The top of the figure shows the input data buffer, which temporally stores the new product information.
2) State Machine & Interface for training data updates: The block receives new product information including the nutrition fact tables and the product names. The training data is selected and updated with an operation of human experts through an interface. The state machine switches to training state during the training updates and sends the state signal to all other blocks.
3) Training Database: The block receives updates from the state machine block and stores the updated data based on expert-decided categories.
4) Word Embedding & DNN model block: The block is controlled by the state machine. It receives data from the training database during the state of training and performs category classification to the new product names with DLSTM model when the state is switched to the recommendation provider.
5) Decision recommendation Model: This block also operates based on the state signal, which receives training data and sets up the nutrition threshold matrix using genetic algorithm introduced in Section IV. In the state of recommendation provider, the decision recommendation model provides a personalised suggestion of products based on the phenotype of the personal data, the category classification outcome of DNN model and the corresponding nutrition facts information. The output is a list of filtered products based on a set of threshold $T_{l,n}$ that maps to the phenotype $P_{l,n}$.

During operations for application, lists of grocery products are constantly updated to the system through input new product data buffer. Then the products will be categorised into groups by the Word Embedding & DNN model block. Product filtering and recommendation processes are carried out in the decision recommendation model block. Each consumer can receive their personalised product recommendation list through uploading their Phenotype/Genotype dataset from the decision

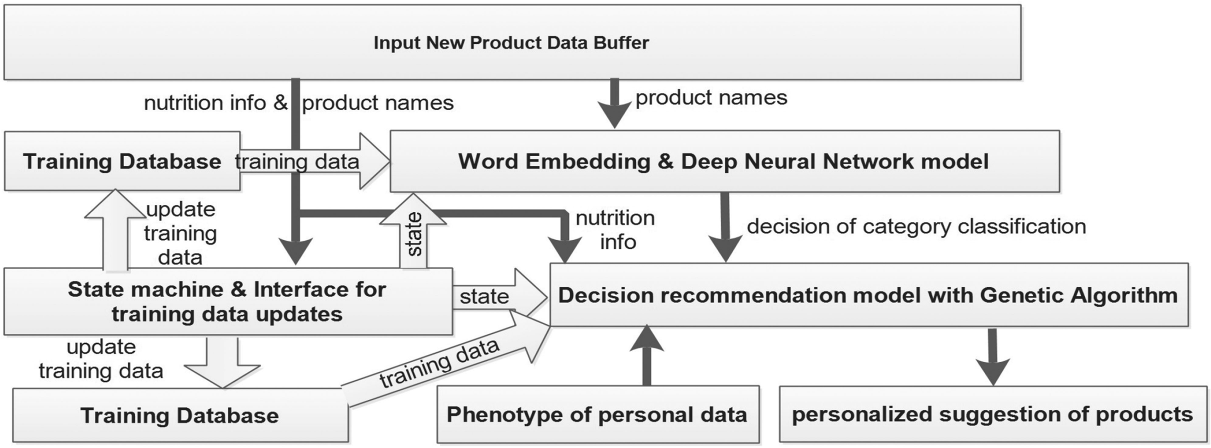

Input New Product Data Buffer
nutrition info & product names
product names
Training Database
training data
Word Embedding & Deep Neural Network model
update training data
state
nutrition info
decision of category classification
State machine & Interface for training data updates
state
Decision recommendation model with Genetic Algorithm
update training data
training data
Training Database
Phenotype of personal data
personalized suggestion of products

Fig. 6. The architecture of the personalised expert recommendation system for optimised nutrition (PERSON).

recommendation model. Note that Word Embedding & DNN model block and the decision recommendation model can be retrained to fit the new grocery data with higher accuracy of categorisation through switching the state and manually giving labels to the new product data.

VI. Case Study

In order to generalize the experiments and test the system, three databases are incorporated into this project: a consumer contributed, a government organised and a corporation created. For the consumer contributed database, Open Food Facts is chosen, which is a global packaged food database with contribution from consumers. In February 2017, the Open Food Facts covered 135,891 world-wide products [35]. For the government organised database, the United States Department of Agriculture (USDA) database is selected. The USDA database is a Public and Private Partnered organised database, which aims to provide support to the development of consumer applications. In February 2017, USDA database covered 184,022 U.S. products' data [36]. For the corporation created database, Tesco is chosen. Tesco is a British grocery retailer, which has branches in different countries. For the research purpose, data from Tesco supermarket is gathered from both on-product labels and public Tesco supermarket websites. 23,518 food products of interests were collected in October 2016. The gathered information includes product names and nutrition facts tables [37]. Once we have compared the results of DLSTM with the three databases, a different structure of DNN is experimented to demonstrate the reason of selecting the structure of LSTM model. Finally with the expert recommendation system introduced in Section V, the potential changes after receiving a grocery recommendation is shown in Section VI-C.

A. Performance of the DNN Categorisation Model

All the product names from the three datasets are used as the input data of the DLSTM-RNN model. In order to compare the difference between the databases and categories, only some of the common categories are chosen to participate in the experiments. The selected categories are Cereals, Biscuits, Cookies, Yogurts, Chocolates, Rice, Noodle and Pasta. Due to the complexity of common subgroups of Tesco and Open Food Facts, the datasets are further grouped into Cereals & Biscuits & Cookies, Yogurts, Chocolates, Rice & Noodle & Pasta.

TABLE III
The Performance of DLSTM Models

	Open Food Facts	USDA	Tesco
1-layer LSTM	41.8%	40.3%	81.2%
2-layer LSTM	41.9%	45.2%	83.8%
3-layer LSTM	42.1%	46.2%	84.0%

After the removal of some corrupted data, the actual number of products participated in the experiments are 942 from Open Food Facts, 9,126 from USDA data and 2,012 from Tesco. Each data consists of a category tag, a name and a nutrition facts table. The data is separated randomly into 60% training data and 40% testing data. With random permutations cross-validation, the accuracy is shown in Table III. The results show a good trace of improvement with more layers of LSTM-RNNs. The Tesco datasets obtain high accuracy due to the organised order of product name, which sequences as brand name, flavor and food name. While some USDA and all Open Food Facts data follow random orders. Since the Open Food Facts data are created manually, some data do not cover the full name of the product. The cross-validation method is time-consuming for deep learning models, therefore training beyond three layers of DLSTM is not covered in this project at the current stage. Note that the incorrect classifications of products are adjusted manually through an interface during updates and the updated data is fed to the stage for personalised grocery decision during the training stage to avoid errors in product recommendation due to mis-classification.

B. Comparison of DNN Structures

CNNs and LSTM-RNNs are popular architectures in DNNs. The comparison of the two architectures based on the Tesco

TABLE IV
ACCURACY OF DNN STRUCTURE COMPARISON

	LSTM	CNN
1-layer	81.2%	72.2%
2-layer	83.8%	72.8%
3-layer	84.0%	72.0%

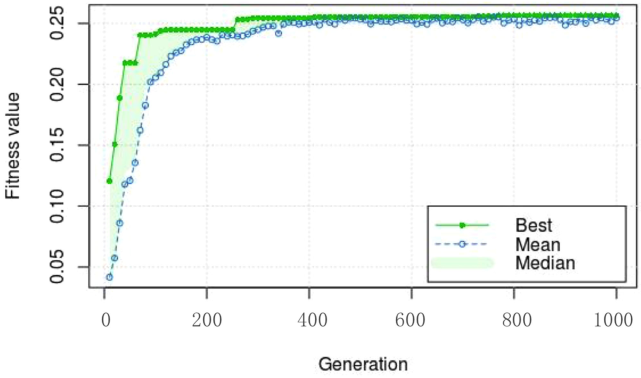


Fig. 7. The optimisation of fitness score using Genetic Algorithm.

TABLE V
AVERAGE REDUCTION OF CONSUMTION IN RESULT OF FOLLOWING THE EXPERT SYSTEM RECOMMENDATION

Per 100 g of the product	Biscuit & Cereal & Cookie	Chocolate	Yogurt	Rice & Noodle & Pasta
Energy	61.2 kJ	68 kJ	63 kJ	200 kJ
Fat	3.5 g	1.64 g	1.58 g	0.38 g
Sugar	5 g	5.6 g	5.82 g	0.3 g
Salt	0.101 g	0.29 g	0.07 g	0.1 g
Protein	1.12 g	1.16 g	1.18 g	0.5 g

dataset is with the same experiment set up of the previous section. The results are shown in Table IV. The experiment results show that LSTM-RNN outperform CNNs. This is due to the fact that CNN is not able to capture dependencies of words within product names as LSTM-RNN. Comparing between different number of layers of CNN does not show traces of improvement, either. The 3-layer LSTM is selected to be implemented in our system due to better performance.

C. Potential Changes to the Personalised Grocery Decisions

The categorised data has fed to the Decision recommendation models introduced in Section IV, the genetic algorithm has been set up to maximize the fitness score with 1,000 iterations of the three databases. Fig. 7 shows the optimisation of fitness score where the value saturated at iteration 410 and the maximum fitness score stay at 0.2568. It is assumed that the decision of product consumption is with the same weight: 100 g. The improvement is measured through how much less of a nutrient that is consumed, comparing the scenario of a decision made with the personalised suggestion and without. It is found that the reductions of nutrition intake do not show significant difference between databases, however the values between different categories are comparable. In Table V, the improvement for each nutrient from different categories is the average of the experiment outcome based on the three databases.

The category rice, noodle and pasta have a much higher reduction on energy intake, while fat, sugar and protein show less reduction, since most of the products are either with low value or zero value of these nutrient. The category chocolate produces greater improvement on salt, while the category biscuits, cereal and cookie provides a greater improvement on fat. The possible key nutrition that causes difference in diets can be presented through relatively greater improvement described above.

VII. CONCLUSIONS

In this paper, we proposed a potential architecture of a personalised expert recommendation system for optimised nutrition (PERSON) with direct recommendation of products based on individual genes. DLSTM is applied as our grocery product categorisation model, while its performance is compared with CNNs'. The categorised grocery products are compared to their own group and recommended to the consumer with different filters based on individual genetic phenotype. It is demonstrated that the interpretation of genetic data can be simplified through the application of nudging daily decisions such as grocery shopping. This is the optimisation using GA to evaluate a fitness score and thresholds for recommendation based on nutrition value. The contribution of this paper includes a novel architecture of the expert system, utilisation of DLSTM for the first time on grocery product categorisation and the implementation of a new recommendation system with genetic algorithm. Further work such as improving the categorisation accuracy can be done through the generalised word embedding on grocery related articles and modification of the introduced DNN model. Furthermore, a big data framework, the consideration of ingredients and the correlation of genetics to nutrients data are planned to be covered in future papers.

ACKNOWLEDGMENT

This project is under supervision of Prof. C. Toumazou at Imperial College London. The authors would like to thank Dr. M. Karvela, Dr. M. Sohbati, and Dr. T. Shinawatra for the help and inspiration.

REFERENCES

[1] S. Linko, "Expert systems what can they do for the food industry?" *Trends Food Sci. Technol.*, vol. 9, no. 1, pp. 3–12, 1998.

[2] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data," *Manage. Revolution Harvard Bus. Rev.*, vol. 90, no. 10, pp. 61–67, 2012.

[3] M. Hansen, T. Miron-Shatz, A. Lau, and C. Paton, "Big data in science and healthcare: A review of recent literature and perspectives," *Yearbook Med. Informat.*, vol. 9, pp. 21–26, 2014.

[4] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, Jan./Feb. 2003.

[5] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Res.*, vol. 2, no. 1, pp. 2–11, 2015.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[7] J. Seide, “Keynote: The computer science behind the Microsoft cognitive toolkit: An open source large-scale deep learning toolkit for windows and linux,” in *Proc. IEEE/ACM Int. Symp. Code Gener. Optim.*, 2017, p. xi.

[8] D. Schneider, “Deeper and cheaper machine learning [top tech 2017],” *IEEE Spectr.*, vol. 54, no. 1, pp. 42–43, Jan. 2017.

[9] C. Chen, L. Li, H. Peng, Y. Yang, and T. Li, “Finite-time synchronization of memristor-based neural networks with mixed delays,” *Neurocomputing*, vol. 235, pp. 83–89, 2017.

[10] M. Zheng, L. Li, H. Peng, J. Xiao, Y. Yang, and H. Zhao, “Finite-time stability analysis for neutral-type neural networks with hybrid time-varying delays without using Lyapunov method,” *Neurocomputing*, vol. 238, pp. 67–75, 2017.

[11] M. Hyman, *Ultrametabolism: The Simple Plan for Automatic Weight Loss*. New York, NY, USA: Simon and Schuster, 2006.

[12] C. M. Phillips, “Nutrigenetics and metabolic disease: Current status and implications for personalised nutrition,” *Nutrients*, vol. 5, no. 1, pp. 32–57, 2013.

[13] C. S. Bloss, N. J. Schork, and E. J. Topol, “Effect of direct-to-consumer genomewide profiling to assess disease risk,” *New Eng. J. Med.*, vol. 2011, no. 364, pp. 524–534, 2011.

[14] C. Toumazou *et al.*, “Simultaneous DNA amplification and detection using a pH-sensing semiconductor system,” *Nature Methods*, vol. 10, no. 7, pp. 641–646, 2013.

[15] L. F. Cherkas, J. M. Harris, E. Levinson, T. D. Spector, and B. Prainsack, “A survey of UK public interest in internet-based personal genome testing,” *PloS One*, vol. 5, no. 10, 2010, Art. no. e13473.

[16] L. Anselma, A. Mazzei, and F. De Michieli, “An artificial intelligence framework for compensating transgressions and its application to diet management,” *J. Biomed. Informat.*, vol. 68, pp. 58–70, 2017.

[17] S. Quinn, R. Bond, and C. Nugent, “Ontological modelling and rule-based reasoning for the provision of personalized patient education,” *Expert Syst.*, vol. 34, no. 2, pp. 1–27, 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[19] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 841–842.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[21] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. Nov., pp. 2579–2605, 2008.

[22] K. S. Tai, R. Socher, and C. D. Manning, “Improved semantic representations from tree-structured long short-term memory networks,” in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Language Process., ACL*, 2015, pp. 1556–1566.

[23] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4580–4584.

[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in *Proc. Interspeech*, 2014, pp. 338–342.

[26] M. Huang, Y. Cao, and C. Dong, “Modeling rich contexts for sentiment classification with LSTM,” arXiv preprint arXiv:1605.01478, 2016.

[27] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Repre-sentations*, 2014, pp. 1–15.

[28] F. B. Churchill, “William Johannsen and the genotype concept,” *J. Hist. Biol.*, vol. 7, no. 1, pp. 5–30, 1974.

[29] W. Johannsen, “The genotype conception of heredity,” *Int. J. Epidemiol.*, vol. 43, no. 4, pp. 989–1000, 2014.

[30] R. C. Lewontin, “The units of selection,” *Annu. Rev. Ecol. Systematics*, vol. 1, no. 1, pp. 1–18, 1970.

[31] B. Angers, E. Castonguay, and R. Massicotte, “Environmentally induced phenotypes and DNA methylation: How to deal with unpredictable conditions until the next generation and after,” *Mol. Ecol.*, vol. 19, no. 7, pp. 1283–1295, 2010.

[32] M. Cariaso and G. Lennon, “SNPedia: A wiki supporting personal genome annotation, interpretation and analysis,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1308–D1312, 2011.

[33] L. Scrucca *et al.*, “GA: A package for genetic algorithms in R,” *J. Statist. Softw.*, vol. 53, no. 4, pp. 1–37, 2013.

[34] L. Scrucca, “On some extensions to ga package: Hybrid optimisation, parallelisation and islands evolution,” *The R J.*, pp. 187–206, 2016.

[35] O. F. Facts, Open food facts website, Oct. 10, 2016. [Online]. Available: http://world.openfoodfacts.org/who-we-are

[36] U. database, US department of agriculture, agricultural research service, nutrient data laboratory, USDA branded food products database, Jan. 14, 2017. [Online]. Available: http://ndb.nal.usda.gov

[37] T. Grocery, Tesco groceries website, Oct. 15, 2016. [Online]. Available: http://www.tesco.com/groceries/

Chih-Han Chen (S’17) received the B.Sc. degree in electrical and electronic engineering from the Tatung University, Taipei, Taiwan, in 2012, the M.Sc. degree in microelectronics from the University of Newcastle Upon Tyne, Newcastle, U.K., in 2013, the M.Sc. degree in electronic engineering with business management from King’s College London, London, U.K., in 2014, and the M.Sc. degree in analog and digital integrated circuit design from Imperial College London, London, U.K., in 2015, where he is currently working toward the Ph.D. degree in the Centre for Bio-Inspired Technology, Imperial College London, with a focus on expert systems for personalized decision based on genetics.

Maria Karvela received the B.Sc. degree in biology from the University of Athens, Athens, Greece, the M.Sc. degree in medical genetics from the University of Glasgow, Glasgow, U.K., and the Ph.D. degree in oncology/hematology from the University of Glasgow, in 2008, 2009, and 2012, respectively. In 2014, She went for postdoctoral studies with Stephen Tait in the Beatson Institute for Cancer Research, Mitochondria, and Cell Death. She was the Principal Retail Scientist in GENEU, in 2015. She is currently the CEO and cofounder of DNAnudge.

Mohammadreza Sohbati (S’09–M’15) received the B.Sc. degree in electrical engineering—telecommunications from the University of Tehran, Tehran, Iran, in 2009, the M.Sc. degree in analog and digital integrated circuit design, the D.I.C. degree from Imperial College London, London, U.K., in 2010, and the Ph.D. degree under supervision of Regius Professor Chris Toumazou, on “Circuits and systems for DNA detection by ion-sensitive field effect transistor” from Imperial College London, in 2015. He is currently a Research Associate in Winston Wong Centre for Bio-inspired Technology, Imperial College London. His current research focuses on genetic technology.

Thaksin Shinawatra received the Master's degree from Eastern Kentucky University, Richmond, KY, USA, and the Ph.D. degree from Sam Houston State University in Huntsville, Huntsville, TX, USA, in 1975 and 1978, respectively. He is a Thai Businessman and Technology Entrepreneur. He has founded many ventures. To name a few, in 1982, he established ICSI; in 1986, he founded the mobile phone operator Advanced Info Service; and in 1987, the IT and telecommunications conglomerate Shin Corporation, the largest mobile phone operator in Thailand. He introduced mobile telephony into Thailand and set up many of the first satellite and 3G mobile infrastructures. He joined politics in 1994 and founded the Thai Rak Thai Party in 1998. In 2001, after a landslide electoral victory, he became the Prime Minister and served on this position until 2006. Since then, he has been an active investor, entrepreneur, and founder in communication and biotechnologies. He is a founder of technology investment company Medtekwiz, a founding investor in Owlstone Medical, a breathing technology for early detection of cancer. He is also a cofounder of DNAnudge, a consumer genetics technology company.

Christofer Toumazou (M'87–SM'99–F'01) is the London's first Regius Professor of engineering and founder of Imperial Colleges Institute of Biomedical Engineering. He is currently the Chairman of DNA Electronics, London, U.K. He is a multiaward winning inventor and serial entrepreneur. His invention of semiconductor DNA sequencing revolutionized genetic testing. In recognition, he received the prestigious 2014 European Inventor Award. His newest venture is startup company DNAnudge, which is developing the first saliva-based, user-operated genetic self-test to personalize consumers shopping experience. He is also a founder of medtech companies DNA Electronics Ltd., and Toumaz Holdings Ltd., He is currently a Trustee at the Royal Institution.