Expert Systems With Applications 115 (2019) 245–255

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes

Jake A. Carter[a], Christina S. Long[b], Beth P. Smith[b], Thomas L. Smith[b], George L. Donati[a,*]

[a] Department of Chemistry, Wake Forest University, Salem Hall, Box 7486, Winston-Salem, NC 27109, USA
[b] Department of Orthopaedic Surgery, Wake Forest School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, USA

ARTICLE INFO

Article history:
Received 18 May 2018
Revised 31 July 2018
Accepted 1 August 2018
Available online 2 August 2018

Keywords:
Diabetes diagnosis
Machine learning
Trace Elemental analysis
Chemometrics
ICP-MS
MIP OES

ABSTRACT

Described for the first time is the use of elemental analysis of diabetic toenails and machine learning techniques for the robust classification of type-2 diabetes. Aluminum, Cs, Ni, V and Zn concentrations in toenails were found to be significantly ($p < 0.05$) different between healthy volunteers and type-2 diabetes patients. Seven different machine learning algorithms were then studied to develop a non-invasive diagnostic method using concentrations of twenty-two elements in toenails, and personal information such as age, gender and smoking history as features. Models were enhanced through feature selection and two different ensembling strategies. The performance of forty-six distinct machine learning models were compared on resampled training data and testing data. A random forest model, trained with concentrations of Al, Ba, Ca, Cr, Cs, Cu, Fe, Mg, Mn, Ni, P, Pb, Rb, S, Sb, Se, Sn, Sr, V and Zn (µg g^{-1}), as well as information on age, gender and smoking history, had an area under the receiver operating characteristic curve (AUC) of 0.73 on the training data, and correctly predicted seven out of nine test samples (including control and disease), with an AUC of 0.90. The results at this stage of the research prove the concept of combining elemental analysis of toenails and machine learning techniques for non-invasively diagnosing type-2 diabetes. With proper sample collection and shipping, mobility-limited patients may be able to mail toenail samples for analysis and monitor their type-2 diabetes over time. A health clinic equipped with common instrumentation, software and trained algorithms similar to those used in the present study may be able to serve a large number of patients from across the world.

1. Introduction

Diabetes is a worldwide epidemic and burden with significant economic impact (World Health Organization [WHO], 2016). In 2012, this disease was the eighth leading cause of death in the world, with 1.5 million fatalities (WHO, 2016). In 2015, it was estimated there were 415 million people with diabetes and 5 million deaths related to the disease (Ogurtsova et al., 2017). This epidemic tends only to grow, with projections showing approximately 642 million people having the disease by 2040 (Ogurtsova et al., 2017). Previous estimates (2010) on diabetes' economic impact predicted a global health expenditure of US$ 490 billion in 2030 (Zhang et al., 2010). However, more recent economic studies estimate the annual cost of the disease ranged from US$ 673 billion in 2015 to more than US$ 827 billion in 2016, already surpassing the prediction for 2030 (NCD Risk Factor Collaboration [NCD-RisC], 2016; Ogurtsova et al., 2017; Seuring, Archangelidi, & Suhrcke, 2015).

Currently, diabetes is diagnosed by measuring the amount of glycated haemoglobin (HbA1C or simply A1C) in the blood, or by measuring glucose in a sample of blood two hours after a patient has taken a 75-g oral dose of glucose (American Diabetes Association [ADA], 2012; WHO, 2016). The former gives results corresponding to an average blood glucose level over two to three months, with an A1C value ≥ 6.5% signifying a positive test for disease, and values between 5.7 and 6.4% indicating a pre-diabetes condition. Both methods, however, are invasive and require blood to be drawn from the patient. Although non-invasive glucose-monitoring technology exists and is a popular field of research (Malik, Khadgawat, Anand, & Gupta, 2016; Witkowska Nery, Kundys, Jeleń, & Jönsson-Niedziółka, 2016), and other human serums such as saliva, sweat and urine also contain levels of glucose and have been studied for glucose monitoring

[*] Corresponding author.
E-mail addresses: cartja16@wfu.edu (J.A. Carter), csigur@wakehealth.edu (C.S. Long), bpsmith@wakehealth.edu (B.P. Smith), tsmith@wakehealth.edu (T.L. Smith), donatigl@wfu.edu (G.L. Donati).

https://doi.org/10.1016/j.eswa.2018.08.002
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

(Makaram, Owens, & Aceros, 2014), blood still is the gold standard for determining glucose levels in diabetes diagnostics.

Some studies have shown differences in the elemental profile (*e.g.* Cu, Mn, Mg, Fe, Se, V, Cr and Zn) of samples taken from urine, hair and blood of diabetic patients (Kazi et al., 2008; Meyer & Spence, 2009). Zinc, for example, has been shown to be significantly ($p < 0.05$) higher in urine, and lower in hair and serum samples when compared with those from healthy individuals. These results suggest an efficiency of Zn in urine is correlated with a deficiency within the body (Badran, Morsy, Soliman, & Elnimr, 2016; Chen, Tan, Lin, & Wu, 2014; Kazi et al., 2008). In this context, elemental analysis of nails may provide an interesting, non-invasive alternative to diabetes screening and diagnosis. Some studies have shown that the concentration of certain elements in fingernails and toenails can be significantly different in healthy and sick individuals, and that chemical imbalances typically present at the onset of a disease may be used for early diagnosis (Ahmed & Santosh, 2010; Fawcett, Linford, & Stulberg, 2004; Hozumi et al., 2011; Mehra & Juneja, 2005). Nails can be collected non-invasively and are relatively inert, so they are less prone to contamination and can be stored for long periods without degradation. In addition, fingernails and toenails grow at an average rate of 3.47 and 1.62 mm/month, respectively, which may be used to collect historical information at a more frequent rate than the current A1C method (Yaemsiri, Hou, Slining, & He, 2010). Thus, with a diabetes diagnostic method based on elemental analysis of nails, patients may be able to mail their samples out for analysis from anywhere in the world, with no need for travel, blood drawing or pain.

Although research suggests metals can be descriptors of diabetes, there still are questions regarding the mechanistic properties, location, and participation of specific elements in biological functions (Meyer & Spence, 2009). Given the complex nature of biological functions and relative lack of knowledge on the exact role certain elements play in the pathogenesis of diabetes, there is a need for more computationally stringent techniques to further aid the application of elemental analysis in diabetes diagnosis. A powerful strategy to such application is statistical learning, or machine learning, which is associated to the process of getting insight on and improving the understanding of complex problems by applying statistical models to a large amount of data (Friedman, Hastie, & Tibshirani, 2001; James, Witten, Hastie, & Tibshirani, 2013). Machine learning algorithms fall into two different types of models: supervised or unsupervised. Supervised learning involves labeled data, whereas unsupervised learning lets the algorithm find patterns without considering to which class the data originally belongs (Friedman et al., 2001; James et al., 2013). Elemental analysis and machine learning have been applied to diverse areas such as food analysis (Batista et al., 2012; Canizo, Escudero, Pérez, Pellerano, & Wuilloud, 2018; Maione et al., 2016), and the analysis of other solid samples (Boucher et al., 2015; Neiva, Chagas Jacinto, Mello de Alencar, Esteves, & Pereira-Filho, 2016). However, there is little work describing the combination of elemental analysis and machine learning techniques for diabetes diagnostics. Machine learning is currently being applied to a wide variety of fields: from self-driving cars and face recognition software, to algorithms used in successful search engines and social networks (Bartlett et al., 2005; Bello-Orgaz, Jung, & Camacho, 2016; Chi & Mu, 2017; Joachims, Granka, Pan, Hembrooke, & Gay, 2005). Considering the flexibility and efficiency of machine learning techniques at identifying patterns in a large and complex set of data, and the high sensitivity, precision and accuracy of elemental analysis techniques such as inductively coupled plasma mass spectrometry (ICP-MS) (Montaser, 1998), a non-invasive method for diabetes diagnosis based on these tools may represent a significant advancement to the field. Disease control and aspects related to patient access to test facilities would be significantly improved if significant differences in the mineral constitution of such stable samples as fingernails and toenails could be detected and used for disease screening, monitoring and diagnosis.

In the present work, we use microwave-induced plasma optical emission spectrometry (MIP OES) and ICP-MS to determine twenty-two elements in toenail samples taken from people with and without type-2 diabetes. Univariate, multivariate, and machine learning analyses are performed to determine the importance and significance of each element in discriminating disease from control. Seven different machine learning models are studied for the robust classification of diabetes using the concentrations of the twenty-two elements evaluated, as well as information on age, gender and smoking history as features. Feature selection and two different ensemble strategies are employed to further enhance predictions. Forty-six distinct models are compared. The best model is then chosen based on its performance during a training phase, which involves tuning hyperparameters across a resampling process, and testing data external to the entire training process.

2. Materials and methods

2.1. Instruments

A MIP OES (4200 MP-AES, Agilent Technologies, Santa Clara, CA, USA) and a tandem ICP-MS (8800 ICP-MS/MS, Agilent, Tokyo, Japan) were used to determine elemental concentrations in toenails. A closed-vessel microwave-assisted digestion system (ETHOS UP, Milestone, Italy) was used for sample digestion. Instrumental parameters and optimized conditions for elemental analysis are listed in Table 1.

2.2. Reagents and standard reference solutions

All samples and analytical solutions were prepared using trace-metal-grade nitric acid (Fisher, Pittsburgh, PA, USA) and distilled-deionized water (18 MΩ cm, Purelab Option-Q, Elga, Woodridge IL, USA). Low trace metals hydrogen peroxide (Veritas, Columbus, OH, USA) was also used for sample digestion. Single-element stock solutions of Al, B, Ba, Ca, Cr, Cs, Cu, Fe, Mg, Mn, Mo, Ni, P, Pb, Rb, S, Sb, Se, Sn, Sr, V and Zn (1000 mg L^{-1}, SPEX CertPrep, Metuchen, NJ, USA) were used to prepare the standard solutions used for calibration. The external standard calibration method was employed in all determinations.

2.3. Samples and sample preparation

Toenail samples from type-2 diabetes patients (disease, $n = 21$) and healthy volunteers (control, $n = 19$) were collected in accordance with procedure approved by the Wake Forest School of Medicine (IRB 00033754). All samples were taken using a sharp sterile nail nipper and stored in sterile clear cups, under cool, dry conditions until analysis. A1C values were recorded at the time the toenail sample was collected. For diabetic participants tested between October 7th, 2016 and August 25th, 2017, A1C values were in the 5.6 - 11.9% range. One of these patients, although presenting an A1C value of 5.6, is a confirmed diabetic individual, who has shown an A1C value of 6.0 in a latter visit to the clinic. Age-matched healthy volunteers were not submitted to the A1C test. Individual information, including smoking history (never a smoker, former smoker, or current smoker), gender (male, female), and age were also collected and used in the study.

Sample masses in the 0.1–0.2 g range were accurately measured and transferred to a polytetrafluoroethylene (PTFE) digestion vessel. Aliquots of 1.0 mL of concentrated HNO_3 and 2.0 mL of 30% v/v H_2O_2 were added to the samples, and the solution was diluted with distilled-deionized water to a final volume of 10.0 mL. The

Table 1
Instrumentation and operating conditions used for elemental analysis of toenail samples.

Instrument	Instrumental parameter	Operating condition
MIP OES	Microwave applied power (kW)	1.0
	Nebulizer gas flow rates (L min^{-1})	0.95 (Al), 0.60 (Ca), 0.65 (Fe), 0.90 (Mg), 0.75 (N_2^+) (Lowery et al., 2016)
	Peristaltic pump speed (rpm)	15
	Integration time (s)	3
	Plasma observation position	0 (all analytes)
	Nebulizer	Inert OneNeb
	Spray chamber	Cyclonic, double pass
ICP-MS/MS	Radio frequency (RF) applied power (W)	1550
	Sampling depth (mm)	10.0
	Carrier gas flow rate (L min^{-1})	1.05
	Nebulizer pump rate (rps)	0.10
	Nebulizer	Micromist, concentric
	Spray chamber	Scott-type, double pass, operated at 2 °C
	Replicates	3
	Sweeps per replicate	100

heating program used for microwave-assisted digestion included a 15 min ramp to 200 °C, a 15 min hold at 200 °C, and a 15 min cool down period for a total run time of 45 min. Digested samples and blanks were diluted with distilled-deionized water to 25.0 mL for a final acid concentration of 4% v/v HNO_3.

2.4. Sample analysis

Concentrations of the macro elements Al, Ca, Fe, Mg and Zn were determined by MIP OES using the 396.152, 393.366, 371.993, 285.213 and 481.053 nm emission lines, respectively. The N_2^+emission peak at 391.470 nm was used to correct for any potential signal bias in MIP OES determinations (Lowery, McSweeney, Adhikari, Lachgar, & Donati, 2016). All samples were diluted 5-fold and maintained in 4% v/v HNO_3 before analysis.

Microelements Cs, Sb, Sn, Sr and Rb were determined by ICP-MS at the mass-to-charge ratios (*m/z*) 133, 121, 118, 88 and 85, respectively. Single quadrupole mode (Q2), with no gas in the octopole collision / reaction cell (ORC), was adopted in this case. For B, Ba, Cr, Cu, Mn, Ni, Pb, Se and V (*m/z* = 11, 137, 52, 63, 55, 60, 208, 78 and 51, respectively), on-mass ICP-MS/MS was used, with H_2 gas flowing at 4 mL min^{-1} (for Ba, Cr, Mn and Pb), or He gas flowing at 3.5 mL min^{-1} (for B, Cu, Ni, Se and V) in the ORC (Fernández, Sugishama, Encinar, & Sanz-Medel, 2012). Finally, P and S were determined using mass-shift mode ICP-MS/MS, with O_2 gas flowing at 20 mL min^{-1} in the ORC (Balcaen, Bolea-Fernandez, Resano, & Vanhaecke, 2015). In this case, the first quadrupole (Q1) was set to *m/z* 31 and 32, and the second quadrupole (Q2) was set to *m/z* 47 and 48, with samples diluted 50-fold and 1000-fold for P and S determination, respectively.

2.5. Data analysis

The concentrations of 22 elements in the digested toenail solutions ranged from ng L^{-1} to mg L^{-1}. All concentrations were normalized by sample mass and converted to µg g^{-1} in the original solid sample. Values found to be below the limits of detection (LOD) were considered as zero for statistical analysis and modeling. The categorical individual information values, *i.e.* gender and smoking history, were represented with discrete values ranging from 0 to 2 for statistical analysis. Gender was coded as 1 and 2 for male and female, respectively. For smoking history, 0, 1 and 2 were adopted for the categories of *never a smoker, former smoker*, and *currently a smoker*, respectively. A final $n \times q$ data matrix was organized and used in all further analyses, with rows representing samples, and columns representing variables. In this case, the variables were the concentrations of the 22 elements evaluated (in µg g^{-1}) plus the values for age, gender and smoking history. In the data mining community, different terms are used when referring to samples and variables. Therefore, for clarity, samples are referred to as observations and variables are referred to as features in the present work.

The statistical language R and the Caret package were used to develop machine learning models (Kuhn & Johnson, 2013; R Core Team, 2017). All code is available upon request.

2.5.1. Feature importance

The importance of each feature (*i.e.* elements and other individual information) was measured using three different types of data analyses: univariate, multivariate, and machine learning. Univariate analysis was performed according to the two-sample Student's *t*-test, *F*-score determination, and the *chi*-statistic. For multivariate and machine learning analysis, PCA and random forest were used, respectively. Considering machine learning as an alternative measure of statistical importance, results from the random forest feature importance test were weighted against those from the univariate analyses during data interpretation (Craig-Schapiro et al., 2011).

2.5.2. Univariate analysis

A two-sample Student's *t*-test was used to determine whether or not to reject the null hypothesis that the data from control and disease observations came from independent, random samples with normal distributions, with equal means, and equal but unknown variances (Warren, Denley, & Atchley, 2014).

The *F*-score is a simple technique to measure the discrimination of two sets of real numbers. Given training vectors, $\boldsymbol{x}_k$ (with $k = 1, \ldots, m$), let the number of positive and negative responses be n_+ and n_-, respectively. The *F*-score of the *i*th feature is defined by Eq. (1), where $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the average of the *i*th feature of the whole, positive, and negative data sets, respectively (Chen & Lin, 2006). In the present work, observations of the disease and control groups were represented as positive and negative, respectively. Hence, $x_{k,\,i}^{(+)}$ was the *i*th feature of the *k*th disease observation, and $x_{k,\,i}^{(-)}$ was the *i*th feature of the *k*th control observation. The larger the F-score, the more likely the feature was discriminative of disease and control.

$$F\,(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,\,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,\,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \tag{1}$$

The *chi*-statistic measures the lack of independence between a term, *t*, and a category, *c*. The term-goodness measure is defined by Eq (2), where *A* is the number of times *t* and *c* co-occur, *B* is the number of times *t* occurs without *c*, *C* is the number of times

c occurs without t, and D is the number of times neither c nor t occur; N is the total number of events (Yang & Pedersen, 1997). For this work, c was adopted as referring to disease.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{2}$$

From the χ^2 value, the Cramer's V coefficient was determined (Cramér, 1946). This is a post-test to *chi*-square where the value of V ranges from 0 (no dependence between variables) to 1 (completely dependent variables). Eq. (3) defines V, where k is the smaller of the number of rows minus one, or the number of columns minus one.

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{3}$$

2.5.3. Multivariate analysis

PCA was used as an exploratory and visual tool to uncover mutual relationships and correlations in the data. Patterns are described in terms of complimentary scores and loadings plots (Wold, Esbensen, & Geladi, 1987). This is an unsupervised technique which seeks to explain the majority of variance within the first few principal components (PCs) without taking into consideration the response of each sample (*i.e.* disease or control). Therefore, the results are an unbiased description of the underlying patterns within the data. PCA is sensitive to the scale of the data across the features in the data matrix (Van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006). To ensure a uniform scale, features were normalized to a mean of 0 and a standard deviation of 1.

2.5.4. Random forest importance

In addition to univariate and multivariate approaches, feature selection methods may use machine learning algorithms to produce measurements of feature importance for classification (Chandrashekar & Sahin, 2014; Chen & Lin, 2006; Guyon & Elisseeff, 2003; Guyon, Weston, Barnhill, & Vapnik, 2002; Saeys, Inza, & Larranaga, 2007). Some models perform inherent feature selection and are considered robust, or relatively robust, to noise. Models represented in this study, which fall under this category, are penalized logistic regression and random forest (Breiman, 2001; Friedman, Hastie, & Tibshirani, 2010). This work used random forest to represent the application of machine learning for determining feature importance.

Random forest classification measures feature importance in two ways: permutation importance and *gini* importance (Breiman, 2001). The results in this study were determined from measurements based on the *gini* index (GI). The GI is a criterion used when growing data trees in random forest classification. The *gini* importance measures the significance of a feature in relation to a tree and a split in the random forest ensemble of trees. By indexing the node for a given tree by n, Eq. (4) is obtained, where gVI_j measures the importance of feature j from summing over the nodes containing feature j in tree k (Goldstein, Polley, & Briggs, 2011). The higher the value for gVI_j, the better the feature was in splitting the data, and the greater the significance of that feature. The list of importance for each feature in the study was determined and the results were scaled to the highest-ranking feature.

$$gVI_j = \frac{1}{ntree} \sum_{k=1}^{ntree} gVI_{jk} \tag{4}$$

2.5.5. Feature selection

The two-sample Student's t-test, F-score, and *chi*-statistic are examples of filter methods when applied to feature selection in machine learning applications (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2013; Saeys et al., 2007). A subjective threshold value would be set, and all features not meeting a specific criterion would be discarded. The level of significance in a two-sample Student's t-test, and the values of F-score and V are examples of potential threshold values to set when choosing representative features. Disadvantages of univariate filter methods include the loss of mutual information between features, and the fact that the performance of a given feature with regard to a specific machine learning algorithm is not considered (Kuhn & Johnson, 2013; Saeys et al., 2007).

Recursive feature elimination (RFE) was used for feature selection in the present work. This strategy is an example of a backward feature elimination method in which the importance of a feature is measured by the performance of a machine learning algorithm on a set of data (Guyon & Elisseeff, 2003; Guyon et al., 2002). The strategy is based on three steps: (*i*) train the classifier, (*ii*) compute the ranking criterion for all features, and (*iii*) remove features with the smallest-ranking criterion. The process repeats a number of times equal to the number of initial features included in the loop. Thus, the optimal subset of features is determined by the performance of a model across the different subsets of features from the RFE process (Guyon et al., 2002).

All models were trained using the entire set of features first. Models which do not perform implicit feature selection, including random forest, were trained again according to a feature selection loop using the RFE strategy. The random forest feature importance scores (based on the initial first model) determined the order of importance for random forest. Logistic regression used the absolute value of the Z-statistic for each model parameter, and all other models ranked predictors with the area under the receiver operating characteristics curve (ROC) for each individual feature. To avoid feature selection bias, the RFE method included an outer resampling loop that encompassed the entire process (Ambroise & McLachlan, 2002; Castaldi, Dahabreh, & Ioannidis, 2011; Friedman et al., 2010; Kuhn & Johnson, 2013). Ten-fold cross validation repeated five times was used to tune hyperparameters and test the out-of-fold predictions for each model. In addition, each model was trained on the same set of resampling data to allow for equal comparisons.

Although proper resampling was employed in the present work, given the small set of samples, overfitting may still occur (Kuhn & Johnson, 2013; Varma & Simon, 2006). An initial 75 / 25 split of the data was adopted as a means of testing the performance of models with data external to the entire training process. The data was split into a training data set consisting of thirty-one observations and a testing data set consisting of nine observations. The two sets were split to have equal disease / control ratios. Models were developed and optimized using the training data set. Overfitting, and subsequently how well the models generalized across different subsets of data, were evaluated by measuring the performance of models on the testing data set.

2.5.6. Machine learning algorithms

A diverse set of popular machine learning algorithms were chosen for the present study (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). The set included examples of linear, non-linear, tree-based, and ensemble statistical models. The following models were developed and tested: random forest, support vector machine with a radial basis function kernel, k-nearest neighbors, naïve Bayes, logistic regression, penalized logistic regression, and linear discriminant analysis. Table 2 lists each model used with the respective descriptions and abbreviations.

2.5.7. Machine learning ensembles

An ensemble is a collection of statistical models in which the predictions of each model within the ensemble are combined via

Table 2
Statistical model, ensemble, and feature set abbreviations and respective definitions.

Abbreviation	Meaning
rf	Random forest
svm	Support vector machine with a radial basis function kernel
lr	Logistic regression
lr_net	Penalized logistic regression
nb	Naïve Bayes
knn	k-nearest neighbor
lda	Linear discriminant analysis
full	Full set of twenty-five features
RFE	Reduced set of features after RFE
_ENS	Stacked ensemble where the model name before "_ENS" is the model outside of the first layer
k	An ensemble consisting of averaging the predictions of the k number of models within the ensemble

methods such as weighting averages, voting, and stacking (Caruana, Niculescu-Mizil, Crew, & Ksikes, 2004; Ren, Zhang, & Suganthan, 2016). By consulting different models within the ensemble, individual model biases may be accounted for and a more accurate prediction may be obtained (Tan & Gilbert, 2003). Two ensembling approaches were featured in the present work. The first approach was to compute simple averages for each unique group of trained models. All unique groups ranging from 2 to 13 members were considered. The best models were submitted to testing on the testing data set. The second approach involved a simple one-layer-stacked ensemble in which the predictions of each individual model trained in the study were used to train a model outside of the first layer. The training set featured rows representing observations and columns representing individual models. The value in each cell was the average out-of-fold prediction from the resampling process for both the observation and the model corresponding to that cell.

2.5.8. Machine learning performance metrics

Hyperparameters are tuning parameters specific to a model which are set to maximize a given performance metric during training (Kuhn & Johnson, 2013). Hyperparameters in this study were optimized to maximize AUC. Model comparisons after training were made based on AUC values alone. The accuracy, sensitivity, and specificity were reported for models with the highest AUC values on the testing data.

In a two-dimension ROC graph, the true positive rate of a classifier is plotted on the *y*-axis and the false positive rate is plotted on the *x*-axis (Fawcett, 2006). The AUC value is the probability that a randomly chosen observation is correctly classified using a given classifier. The diagonal line, *i.e.* $y = x$, represents the probability of random guessing (Fawcett, 2006; Hanley & McNeil, 1982). Therefore, an AUC value > 0.50 implies predictions using a classifier are better than random guesses.

Accuracy is the ratio of the number of correctly predicted observations to the total number of observations. Sensitivity (true positive rate) and specificity are defined by Eqs. (5) and (6), respectively, where *TP, TN, FP*, and *FN* are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specifcty = \frac{TN}{TN + FP} \tag{6}$$

3. Results and discussion

3.1. Univariate descriptions of disease

Table 3 shows the mean values (element concentration and categorical values) for control and disease observations, along with respective minimum, maximum and standard deviation. Age, Al, Cs, Ni, V and Zn were significantly ($p < 0.05$) different between control and disease. Al, Ni, V and Zn were higher, whereas age and Cs were lower in observations from diabetic patients. *F*-score and *chi*-square further confirmed the results from the two-sample Student's *t*-test (Figs. 1 and 2). The highest-ranking features according to *F*-Score were Ni, age, V, Al, Zn and Cs, respectively. There was a drop off in importance after Cs, which suggests the majority of the biasing information was gathered from the top six features. Similarly, *chi*-square listed Zn, Ni, Al and Cs as the most important features, respectively. Contrasting *t*-test and *F*-score, however, age was the tenth most important feature in *chi*-square, and unlike *F*-score, there was no significant cutoff point for the features evaluated.

The LODs for each element evaluated in the present study were determined according to the IUPAC definition (IUPAC, 1997). The values calculated for Al, B, Ba, Ca, Cr, Cs, Cu, Fe, Mg, Mn, Mo, Ni, P, Pb, Rb, S, Sb, Se, Sn, Sr, V and Zn were 10, 4, 0.08, 10, 0.2, 0.001, 0.04, 30, 3, 0.007, 0.009, 0.04, 5, 0.004, 0.003, 3, 0.001, 0.5, 0.03, 0.003, 0.004 and 100 ng g^{-1}, respectively. There was a high variability of results for most of the elements investigated in this study (Table 3 and Fig. 3). As such, it may be inaccurate to rely on one element and its mutual relationship with a response as a descriptor of disease. A more suitable approach for determining predictors of disease in this case should involve multivariate and flexible strategies.

3.2. Multivariate and machine learning descriptions of disease

Fig. 4 shows the biplot of the scores and loadings from the first two principal components of a PCA performed with all features evaluated in this study. One of the control observations was separated from the other observations due to its large concentration of Rb and Pb (top left-hand side in Fig. 4). It was responsible for the maximum values for these two elements listed in the control group in Table 3. This observation may represent an outlier within the data, specifically in the control group. With the limited data set, however, no outlier tests were performed and each observation was included in the analysis. Seven disease observations may be considered separated from the rest of the data due to their relatively high score values in the first principal component. A positive correlation between the elements Al, Ni, V and Zn, and a negative correlation between these elements and both age and Cs was responsible for the separation (Fig. 4). Al, Ni, V and Zn had relatively high values in the loadings from the first principal component, whereas age and Cs had relatively low values. This agrees with the negative correlation between age and Cs, and Al, Ni, V and Zn considering the means of disease and control from Table 3, as well as the results from the univariate analyses.

In addition to the statistically significant elements from Table 3, Fe, Mn, Sn and Se also played a role in separating disease from control (Fig. 4). This suggests further information was to be gained

Table 3
Summary statistics and *p*-values after two-sample Student's *t*-tests for the concentrations of twenty-two elements (μg g^{-1}) and the age of individuals participating in this study at the time of sample collection.

Feature	Control				Diabetes				
	Min	*Max*	*Mean*	*S.D.*	*Min*	*Max*	*Mean*	*S.D.*	*p-value*
Age	36	87	74	14	48	82	65	10	0.02
Al	0.63	48.1	30	10	8.66	77.2	40	20	0.03
B	0.17	9.62	1	2	0.06	7.17	1	2	0.74
Ba	< LOD	1.51	0.5	0.4	< LOD	3.94	1	1	0.96
Ca	487	1490	800	300	448	1800	1000	400	0.15
Cr	< LOD	5.26	0.4	0.1	< LOD	1.42	0.2	0.3	0.57
Cs	< LOD	0.04	0.007	0.01	< LOD	0.01	0.002	0.004	0.03
Cu	1.22	4.57	2.4	0.8	1.51	20.7	4	4	0.20
Fe	< LOD	96.0	40	30	< LOD	81.6	40	20	1.00
Mg	103	418	180	70	77.3	1720	200	400	0.62
Mn	0.06	0.75	0.3	0.2	0.12	0.82	0.4	0.2	0.08
Mo	< LOD	0.01	0.003	0.004	< LOD	0.08	0.01	0.02	0.17
Ni	< LOD	0.36	0.09	0.09	0.02	0.80	0.2	0.2	0.02
P	196	1000	500	200	205	1230	600	300	0.18
Pb	0.01	4.44	0.3	1	0.01	0.32	0.06	0.08	0.30
Rb	0.13	4.85	0.7	1	0.12	1.66	0.6	0.4	0.63
S	10,500	30,300	20,000	5000	11,200	27,300	20,000	5000	0.18
Sb	< LOD	0.49	0.05	0.1	< LOD	0.31	0.05	0.08	0.99
Se	0.29	1.32	0.8	0.3	0.55	4.97	1	1	0.11
Sn	< LOD	0.34	0.09	0.1	< LOD	169	8	40	0.34
Sr	0.29	3.52	1.1	0.9	0.30	2.76	1	0.6	0.49
V	0.00	0.06	0.02	0.02	0.01	0.16	0.04	0.04	0.03
Zn	< LOD	410	100	100	52.1	413	200	90	0.03

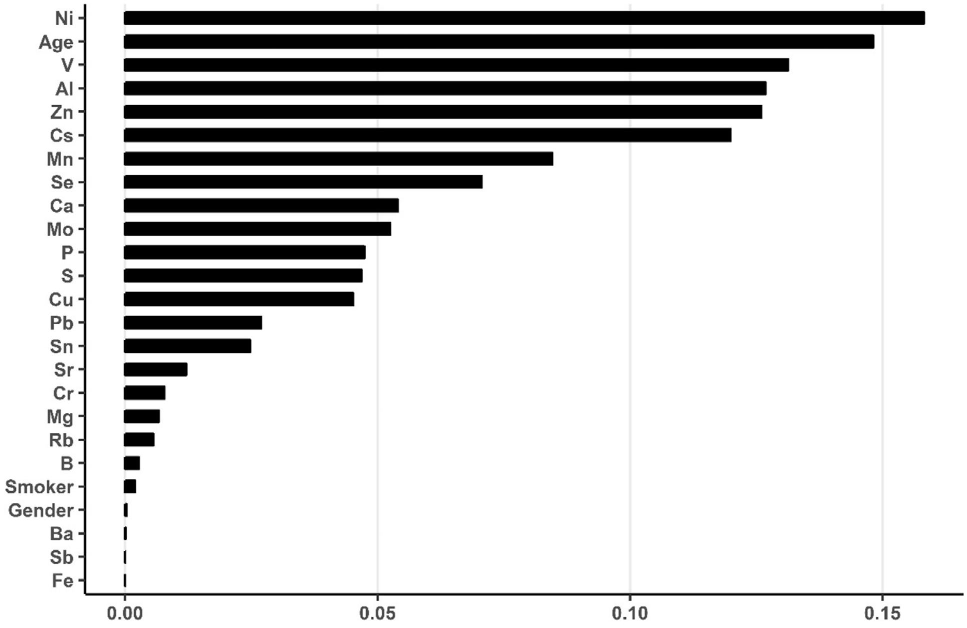

Fig. 1. *F*-scores for each feature including 22 elements, age, gender and smoking history.

from considering more elements than what the Student's *t*-test suggested as significant. Although a separation was present for seven disease observations and one control observation, the remaining thirty-two observations were not distinguishable. They clustered in one large group about the origin in Fig. 4. Therefore, a more flexible approach was needed beyond simple univariate and multivariate analyses to identify subtle patterns separating observations from diabetes and control groups.

Fig. 5 shows a feature importance plot generated from random forest feature importance determination. The features were ordered according to *gini* importance and scaled against the highest-ranking feature, *i.e.* age (Goldstein et al., 2011). Consistent with univariate analyses and PCA, age, Zn, Cs, Al and Ni were important for separating the observations according to response. Contrasting *t*-test, *F*-score and *chi*-square, random forest listed vanadium as the thirteenth most important feature. There is a sharp cutoff after age, and another cutoff after Cs in Fig. 5. This suggests a large portion of the observations were split into disease and control considering the features age, Cs and Zn alone. The remaining elements aided in identifying the patterns, further separating disease from control.

3.3. Machine learning training performance

Fig. 6 lists the resampling performance for all individual models and stacked ensembles according to AUC. Random forest before and after feature selection, and k-nearest neighbors with the full set of features (*i.e.* including all 22 elements plus age, gender and smoking history) performed the best with an average AUC of 0.73

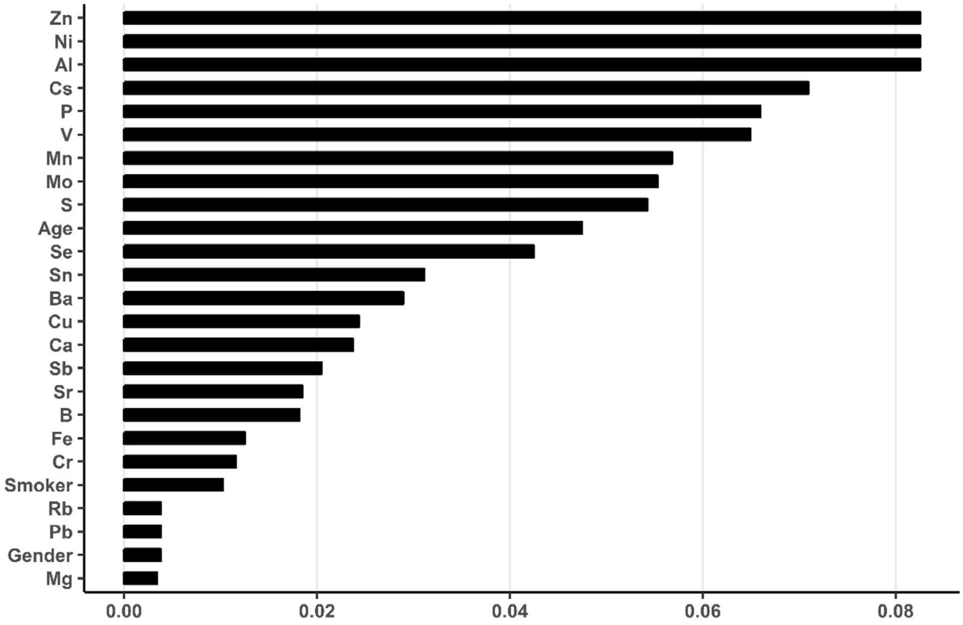

Fig. 2. Cramer's *V* coefficients determined from χ^2 statistics considering 22 elements, age, gender and smoking history.

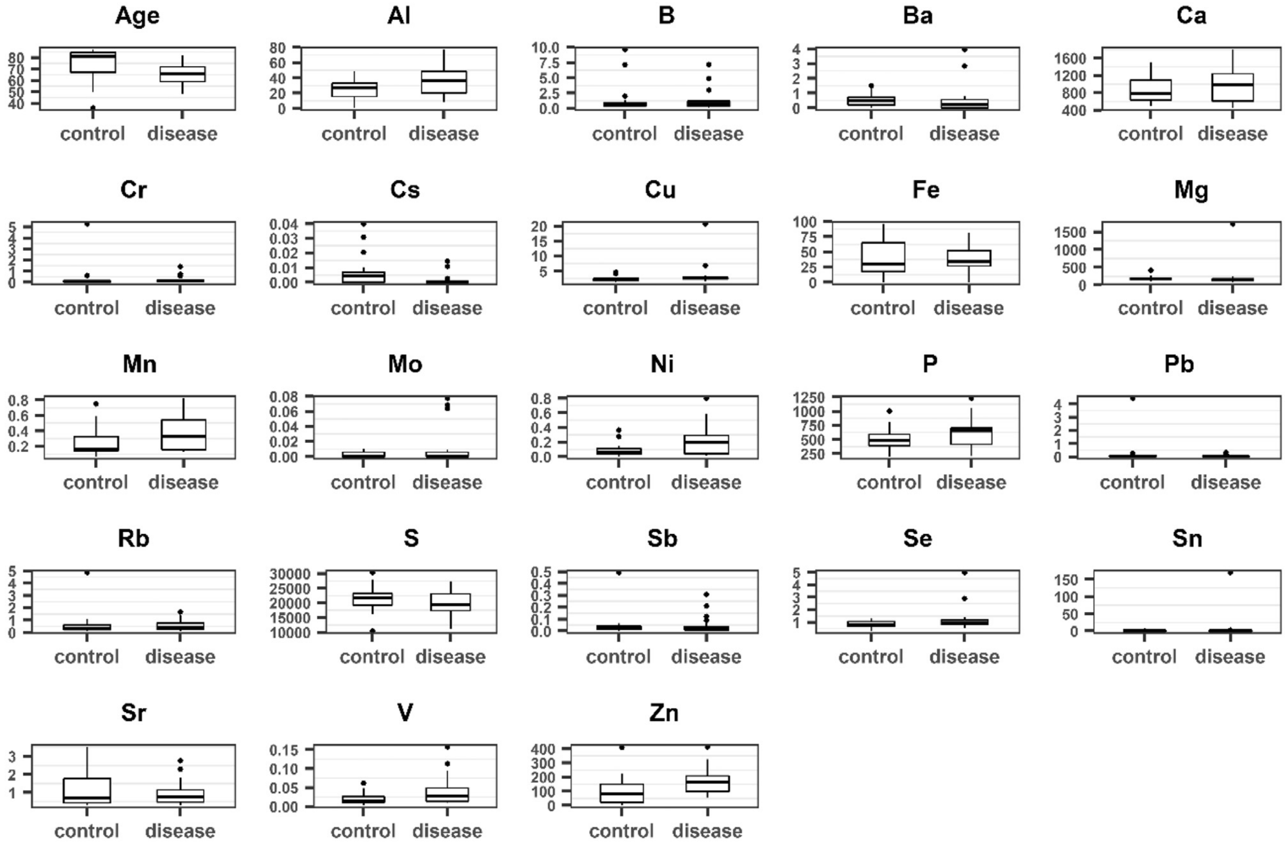

Fig. 3. Boxplots showing differences in element concentration (µg g^{-1}) and age for diabetic patients (disease) and healthy volunteers (control).

for the out-of-fold predictions. Support vector machine with the full set of features was the only other model to have an average AUC above 0.70. The non-linear individual models outperformed the linear individual models. The best stacked ensembles consisted of a support vector machine or a k-nearest neighbor as the model outside of the first layer. This suggests the performance of individual models are not enhanced through stacked ensembles since individual models within the ensemble outperformed the ensemble itself.

For individual models, feature selection contributed to no significant enhancement in model performance, except for the linear discriminant analysis and logistic regression models. Considering the stacked ensembles, the same trend was observed. The linear discriminant analysis and logistic regression stacked ensem-

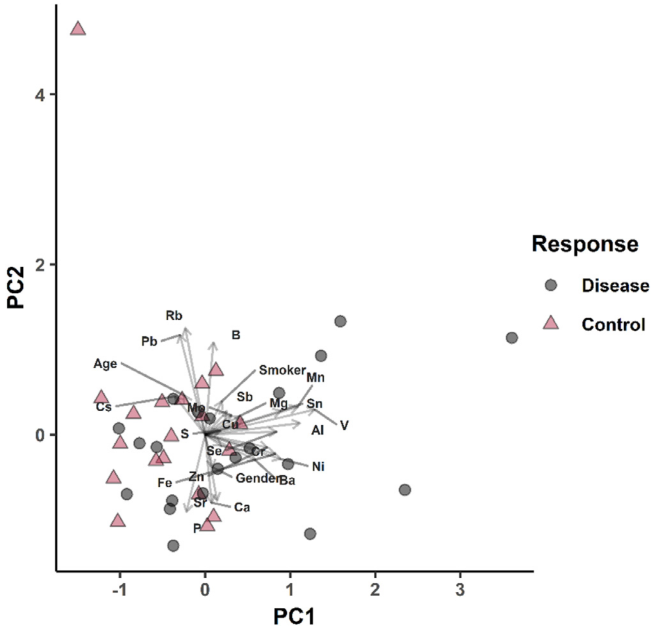

Fig. 4. Biplot of the scores and loadings from the first two principal components. In this case, “Control” and “DM” represents the healthy volunteers and type-2 diabetes groups, respectively.

bles were the only models to perform significantly better with a reduced set of features. The small difference observed for random forest before and after feature selection was expected considering the model is designed to be robust to noise (Breiman, 2001).

Fig. 7 lists the performance of ensembles from averaging individual model predictions. Ensembles of five individual models performed the best, with AUCs of 0.77. The five-membered ensembles consisted of random forest, k-nearest neighbor, support vector machine, naïve Bayes, and logistic regression. The differences between the ensembles were whether the models were trained on the full set of features or a reduced set after feature selection. These ensembles outperformed the individual models within the ensemble. This suggests the performance of individual models could be enhanced by simply averaging predictions across different individual models.

3.4. Machine learning performance based on testing data

Fig. 8 compares the AUC values on testing data for each individual model, all of the stacked ensembles, and the best average ensembles from training. Random forest after feature selection performed the best with an AUC of 0.90. Random forest performed almost equally well with either the full set of features or a reduced set of features. Only a 0.05 increase in AUC was observed with the latter approach. Three of the top five models from Fig. 8 consisted of non-linear models, including random forest and k-nearest neighbors. The other two were a linear discriminant analysis model after feature selection, and a penalized logistic regression model with the full set of features. The five-membered average ensembles performed poorly on the testing data suggesting these ensembles were overfit after training. The best-performing stacked ensembles were a naïve Bayes ensemble with the full set of model predictions, and a support vector machine ensemble with a reduced set of model predictions. Both of these performed much better on the testing data than the training data. The differences in performance between the testing and training data suggests the models are unstable and not able to generalize well across subsets of data. The best average ensembles on the testing data consisted of groups of two, eleven, twelve, and thirteen individual models with AUC values of 0.75. The two-membered ensemble consisting of averaging the predictions of the individual random forest and k-nearest neighbor models trained with the full set of features had an AUC of 0.74 on the training data. The consistency across training and testing implies the ensemble was well tuned across different subsets of data. However, it did not outperform both of the models within the ensemble. Random forest as an individual model outperformed the ensemble.

Table 4 lists additional performance metrics for the top five models from Fig. 8. The highest accuracy across the models was 0.78 for random forest with the full set of features, and random forest and linear discriminant analysis after feature selection.

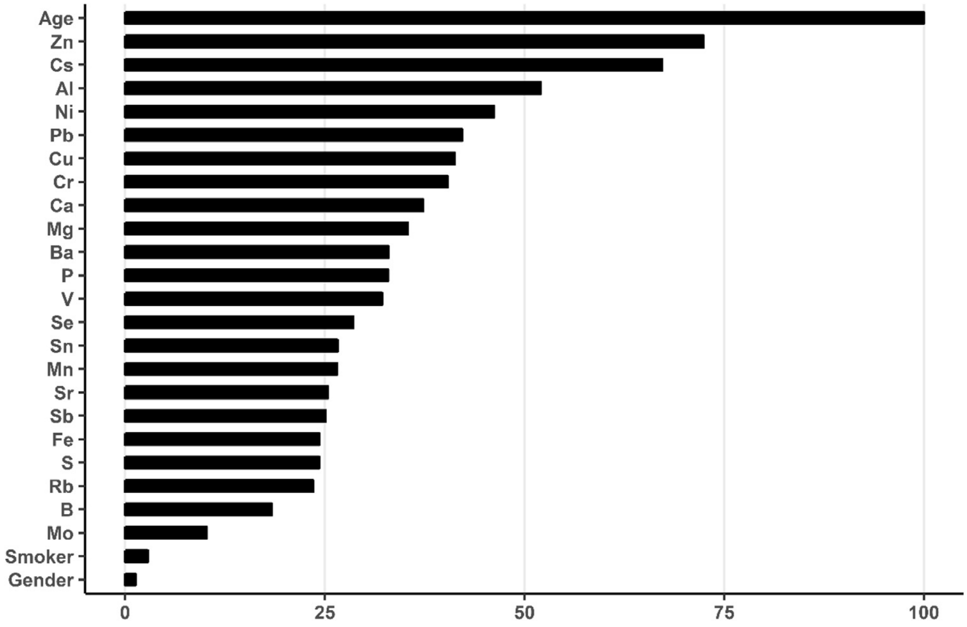

Fig. 5. Scaled feature importance after employing random forest feature importance determination. Values were normalized against age, since it was the highest-ranking feature according to *gini* feature importance.

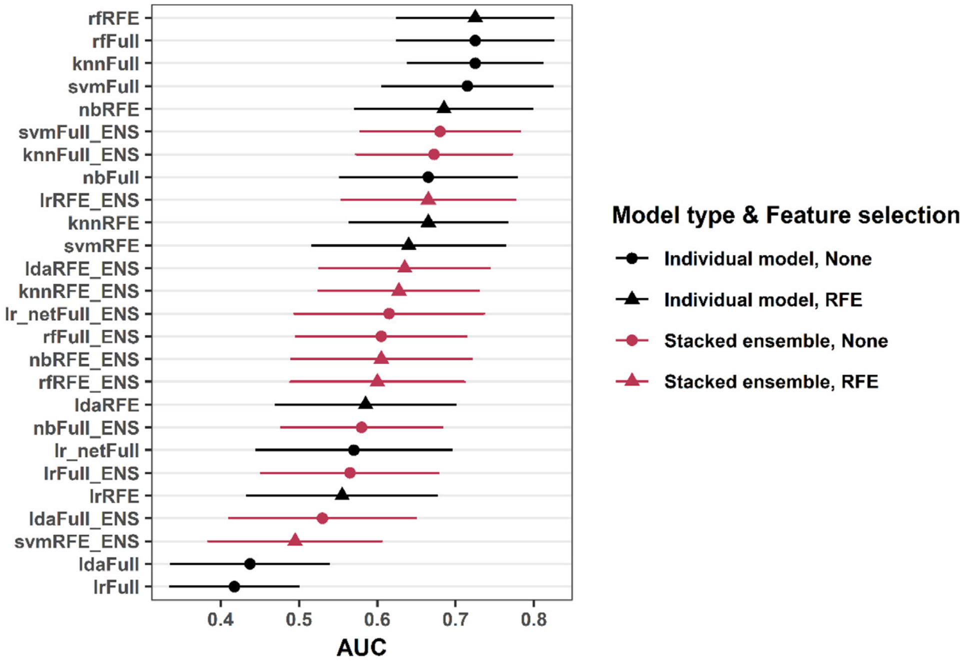

Fig. 6. Resampling performance from the training data (n = 31) after 10-fold cross validation repeated five times. Average AUC values and the corresponding confidence levels at a 95% confidence level are shown.

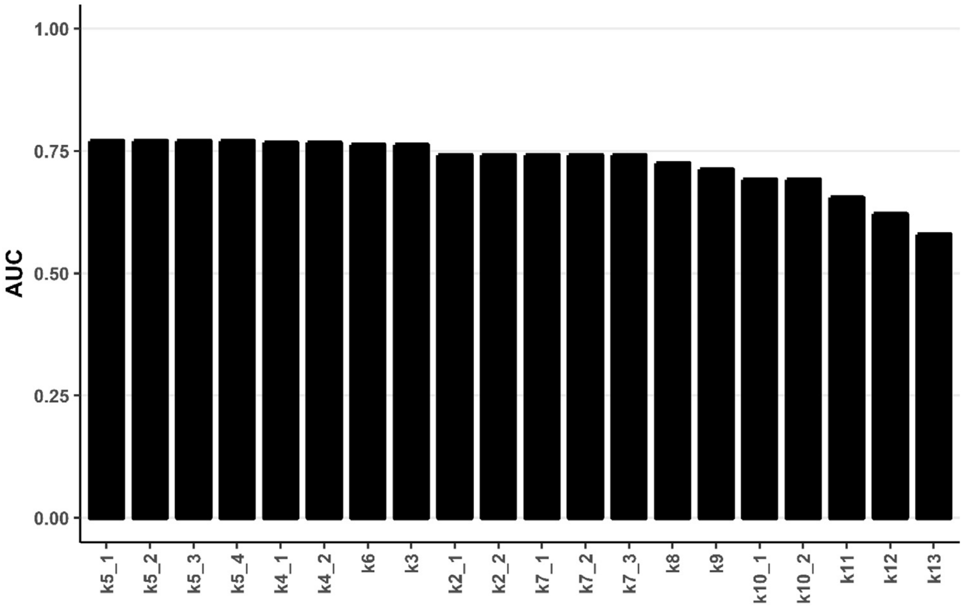

Fig. 7. AUC values after the resampling process for the best ensembles. The average of unique combinations and groups of individual model predictions were used here.

Table 4
Performance metrics on testing data (n = 9) for the best statistical models.

Model	AUC	Accuracy	Sensitivity	Specificity
rfRFE	0.90	0.78	1.00	0.50
rfFull	0.85	0.78	1.00	0.50
knnRFE	0.85	0.56	0.40	0.75
ldaRFE	0.85	0.78	0.80	0.75
lr_netFull	0.80	0.56	0.60	0.50

The most accurate models were more sensitive than specific. Both random forest models predicted two of the four control test observations correctly, and all five of the disease test observations correctly. The linear discriminant analysis model predicted three of the four control test observations correctly, and four of the five disease test observations correctly. Although the linear discriminant analysis model performed similar to both of the random forest models, it performed poorly during the training process suggesting the model does not generalize well across subsets of the data. Therefore, considering performances on both training and testing data, a random forest model, trained on the full set of features except for B and Mo, is the best model at this stage of the study.

4. Conclusions

Described for the first time is the use of elemental analysis of diabetic toenails and machine learning techniques for the robust

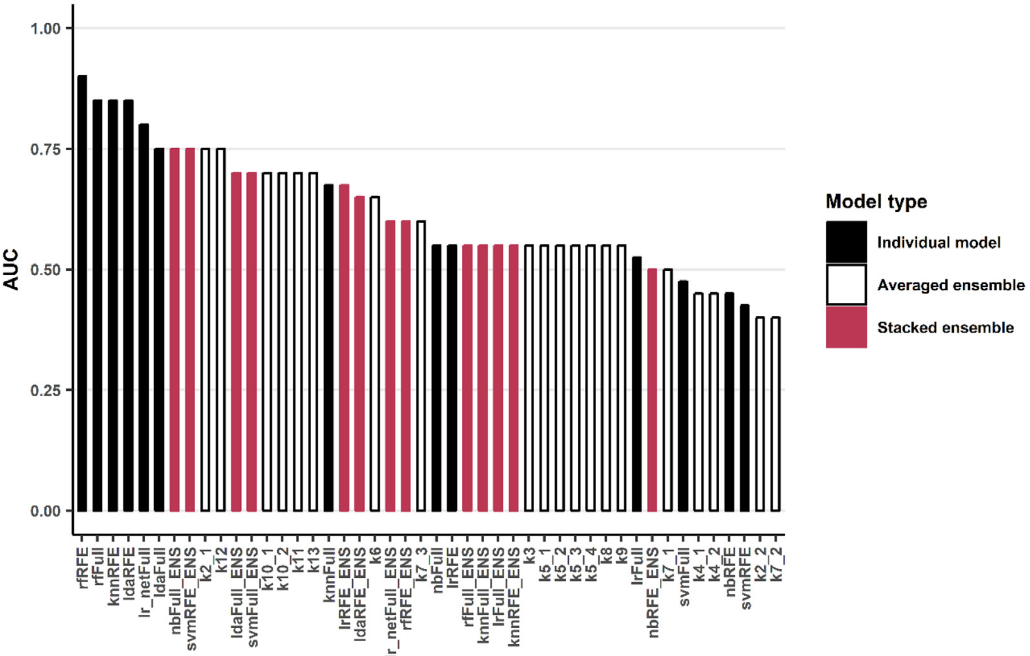

Fig. 8. AUC values calculated for each model after predictions on nine external test observations.

classification of type-2 diabetes. Different from other biological tests, trace element contamination of toenails during sampling and shipping is less likely. Toenails are chemically stable and relatively inert, require limited sample preparation and may be mailed out for analysis by patients with limited mobility. In addition, trace elements are more stable than the organic molecule analytes used in standard diabetes tests, which also facilitates sample shipping. This method may be used as a non-invasive diagnostic tool, and with proper sample collection and shipping, a health clinic equipped with common instrumentation, software, and trained algorithms similar to those described in the present study may be able to serve a large number of type-2 diabetes patients from across the world.

In this preliminary study, we have observed that age, Al, Cs, Ni, V and Zn were significantly ($p < 0.05$) different between the control and disease groups. Average concentrations of Al, Ni, V and Zn were higher in diabetic toenails. These results suggest elements in toenails undergo similar biological processes to those in urine of diabetic patients. Although outside the scope of this proof-of-concept study, additional research is required to understand such processes, which may eventually contribute to better understand the disease and to develop prevention and treatment strategies.

As a diagnostic method, seven different machine learning algorithms were studied, with elemental concentrations in toenails, age, gender and smoking history used as features. Model predictions were enhanced through feature selection and two different strategies of ensembling. Forty-six different machine learning models were developed to compare predictions across resampled training data and testing data. A random forest model, trained on the full set of features excluding B and Mo, had an average training AUC of 0.73, and predicted seven out of nine external test observations correctly, with an AUC of 0.90.

The results at this stage of the research prove the concept of combining elemental analysis of toenails and machine learning techniques for non-invasively diagnosing type-2 diabetes. It is important to highlight here that although relatively diverse, the participants in this study are all from a specific region, and future work should involve a larger sample set which is more representative of a wider population. Given more data, the significance of the statistical analysis may be enhanced, and more robust machine learning models may be trained. Additionally, with more data, outlier tests may be performed to further enhance the training of these models. Once the models have been trained on a large and diverse set of samples, and the methodology has matured beyond the proof of concept stage, easy to operate GUIs and apps may be developed for ease of use within a health clinic. Notwithstanding these limitations, the agreement of results from multiple statistical tests (*i.e.* univariate, multivariate and machine learning tests) suggests the elements and individual parameters found as significant in this report may hold true for a larger set of samples.

Finally, it is also important to note that the method described here is not as specific as the A1C test, as it can identify the disease, but presents no rank associated with its severity. A future work including more data and samples representing a wider section of the population will allow models to be trained beyond simple binary classifications. Machine learning regression models may then be developed to serve as predictors of severity of disease.

Credit author statement

Jake A. Carter: Software, Validation, Formal Analysis, Investigation, Writing – Original Draft, Visualization. **Christina S. Long:** Validation, Formal Analysis, Investigation, Resources, Writing – Review & Editing. **Beth P. Smith:** Methodology, Writing – Review & Editing. **Thomas L. Smith:** Methodology, Writing – Review & Editing. **George L. Donati:** Conceptualization, Methodology, Resources, Data Curation, Writing – Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition.

Declaration of interest

The authors have no competing interests to declare.

Acknowledgements

This work was supported by the National Science Foundation's Major Research Instrumentation Program (NSF MRI, grant CHE-1531698). We also thank the Department of Chemistry and the

Graduate School of Arts and Sciences at Wake Forest University for their support.

References

Ahmed, S. S., & Santosh, W. (2010). Metallomic profiling and linkage map analysis of early Parkinson's disease: a new insight to aluminum marker for the possible diagnosis. *PLoS ONE, 5*(6), e11252 https://doi:10.1371/journal.pone.0011252.

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences, 99*(10), 6562–6566.

American Diabetes Association. (2012). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care, 35*, S64–S71.

Badran, M., Morsy, R., Soliman, H., & Elnimr, T. (2016). Assessment of trace elements levels in patients with Type 2 diabetes using multivariate statistical analysis. *Journal of Trace Elements in Medicine and Biology, 33*, 114–119.

Balcaen, L., Bolea-Fernandez, E., Resano, M., & Vanhaecke, F. (2015). Inductively coupled plasma – Tandem mass spectrometry (ICP-MS/MS): A powerful and universal tool for the interference-free determination of (ultra)trace elements – A tutorial review. *Anal Chim Acta, 894*, 7–19.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior, in: IEE Conference on Computer Vision and Pattern Recognition, CVPR.

Batista, B. L., da Silva, L. R. S., Rocha, B. A., Rodrigues, J. L., Berretta-Silva, A. A., Bonates, T. O., et al. (2012). Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. *Food Research International, 49*(1), 209–215.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59.

Boucher, T. F., Ozanne, M. V., Carmosino, M. L., Dyar, M. D., Mahadevan, S., Breves, E. A., et al. (2015). A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy, 107*, 1–10.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Canizo, B. V., Escudero, L. B., Pérez, M. B., Pellerano, R. G., & Wuilloud, R. G. (2018). Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools. *Food Chem, 242*, 272–278.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble Selection from Libraries of Models. In *Proceedings of ICML-04, 21st International Conference on Machine learning* (pp. 18–25). ACM. pgs.

Castaldi, P. J., Dahabreh, I. J., & Ioannidis, J. P. A. (2011). An empirical assessment of validation practices for molecular classifiers. *Briefings in Bioinformatics, 12*(3), 189–202.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.

Chen, H., Tan, C., Lin, Z., & Wu, T. (2014). The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Computers in Biology and Medicine, 50*, 70–75.

Chen, Y., & Lin, C. (2006). Combining SVMs with various feature selection strategies. In *Feature Extraction* (pp. 315–324). New York: Springer-Verlag.

Chi, L., & Mu, Y. (2017). Deep steering: learning end-to-end driving model from spatial and temporal visual cues. In: ArXiv preprint arXiv:1708.03798, 1–12.

Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., Liu, J., Misko, T. P., et al. (2011). Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer's Disease Diagnosis and Prognosis. *PLoS ONE, 6*(4), e18850 https://doi:10.1371/journal.pone.0018850.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

Fawcett, R. S., Linford, S., & Stulberg, D. L. (2004). Nail abnormalities: Clues to systemic disease. *American Family Physician, 69*(6), 1417–1424.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fernández, S. D., Sugishama, N., Encinar, J. R., & Sanz-Medel, A. (2012). Triple Quad ICPMS (ICPQQQ) as a New Tool for Absolute Quantitative Proteomics and Phosphoproteomics. *Analytical Chemistry, 84*(14), 5851–5857.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res, 15*(1), 3133–3181.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*, Springer series in statistics, 1, New York.

Goldstein, B. A., Polley, E. C., & Briggs, F. B. S. (2011). Random forests for genetic association studies. *Stat. Appl. Genet. Mol. Biol., 10*(1) Article 32, 1–36.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Hozumi, I., Hasegawa, T., Honda, A., Ozawa, K., Hayashi, Y., Hashimoto, K., et al. (2011). Patterns of levels of biological metals in CSF differ among neurodegenerative diseases. *Journal of the Neurological Sciences, 303*(1–2), 95–99.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*: 103. New York: Springer.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on esearch and development in information retrieval* (pp. 154–161). Salvador, Brazil: SIGIR '05.

Kazi, T. G., Afridi, H. I., Kazi, N., Jamali, M. K., Arain, M. B., Jalbani, N., et al. (2008). Copper, Chromium, Manganese, Iron, Nickel, and Zinc Levels in Biological Samples of Diabetes Mellitus Patients. *Biol Trace Elem Res, 122*(1), 1–18.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Lowery, K. L., McSweeney, T., Adhikari, S. P., Lachgar, A., & Donati, G. L. (2016). Signal correction using molecular species to improve biodiesel analysis by microwave-induced plasma optical emission spectrometry. *Microchemical Journal, 129*, 58–62.

Maione, C., de Paula, E. S., Gallimberti, M., Batista, B. L., Campiglia, A. D., Jr, F. B., et al. (2016). Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. *Expert Systems with Applications, 49*, 60–73.

Makaram, P., Owens, D., & Aceros, J. (2014). Trends in Nanomaterial-Based Non-Invasive Diabetes Sensing Technologies. *Diagnostics, 4*(2), 27–46.

Malik, S., Khadgawat, R., Anand, S., & Gupta, S. (2016). Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. *Springer-Plus, 5*(1), 701 https://doi.org/10.1186/s40064-016-2339-6.

Mehra, R., & Juneja, M. (2005). Fingernails as biological indices of metal exposure. *Journal of Biosciences, 30*(2), 253–257.

Meyer, J. A., & Spence, D. M. (2009). A perspective on the role of metals in diabetes: Past findings and possible future directions. *Metallomics, 1*(1), 32–41.

Montaser, A. (1998). *Inductively coupled plasma mass spectrometry* (1st ed.). New York: Wiley.

NCD Risk Factor Collaboration. (2016). Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4· 4 million participants. *The Lancet, 387*(10027), 1513–1530.

Neiva, A. M., Chagas Jacinto, M. A., Mello de Alencar, M., Esteves, S. N., & Pereira–Filho, E. R. (2016). Proposition of classification models for the direct evaluation of the quality of cattle and sheep leathers using laser-induced breakdown spectroscopy (LIBS) analysis. *RSC Advances, 6*(106), 104827–104838.

IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught, A. D. and Wilkinson, A. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook.

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., et al. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice, 128*, 40–50.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing URL https://www.R-project.org/.

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine, 11*(1), 41–53.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.

Seuring, T., Archangelidi, O., & Suhrcke, M. (2015). The economic costs of Type 2 Diabetes: A global systematic review. *PharmacoEconomics, 33*(8), 811–831.

Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics, 2*(3-suppl), S75–S83.

Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 142 https://doi:10.1186/1471-2164-7-142.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(1), 91 https://doi:10.1186/1471-2105-7-91.

Warren, C., Denley, K., & Atchley, E. (2014). *Beginning Statistics* (2nd ed.). Charleston: Hawkes Learning Systems.

Witkowska Nery, E., Kundys, M., Jeleń, P. S., & Jönsson-Niedziółka, M. (2016). Electrochemical glucose sensing: is there still room for improvement. *Analytical Chemistry, 88*(23), 11271–11282.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52.

World Health Organization. (2016). Global report on diabetes Geneva, Switzerland.

Yaemsiri, S., Hou, N., Slining, M., & He, K. (2010). Growth rate of human fingernails and toenails in healthy American young adults. *Journal of the European Academy of Dermatology and Venereology, 24*(4), 420–423.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (pp. 412–420).

Zhang, P., Zhang, X., Brown, J., Vistisen, D., Sicree, R., Shaw, J., et al. (2010). Global healthcare expenditure on diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice, 87*(3), 293–301.