Metabolite Screening for Heart Disease Using Support Vector Machine-Based AI

Edward L. Boone[1], Ryad A. Ghanam[2], Faten S. Alamri[3], Elizabeth B. Amona[4]
Department of Statistical Sciences and Operations Research,
Virginia Commonwealth University, Richmond, Virginia, 23284, USA[1,4]
Department Liberal Arts and Sciences, Virginia Commonwealth University, School of the Arts in Qatar, Doha, Qatar[2]
Department of Mathematical Sciences-College of Science, Princess Nourah bint Abdulrahman University,
P.O. Box 84428, Riyadh 11671, Saudi Arabia[3]

***Abstract*—Algorithms for feature selection are growing in interest among researchers aiming to connect specific features in a dataset with specific classifications. Recent developments in machine learning, particularly Support Vector Machine-based artificial intelligence algorithms have demonstrated excellent classification performance in highly nonlinear data. However, identifying which features contribute most to classification remains challenging, especially when datasets include hundreds of variables. Initially, features must be screened to narrow down the set for deeper analysis. Metabolomics datasets are one such case, where many features must be examined to determine those associated with heart disease diagnosis. This work applies a Genetic Algorithm, incorporating a penalized likelihood approach with Support Vector Machines for mutation, to stochastically search the feature space. A large-scale simulation study demonstrates that the proposed method achieves a high true feature identification rate while maintaining a reasonable false identification rate. The method is then applied to a Qatar BioBank dataset focused on heart disease, reducing the number of candidate metabolites from 232 to 37.**

***Keywords*—Machine learning; genetic algorithm; support vector machines; classification; heart disease; metabolites**

I. Introduction

Metabolites are the intermediate or end products of metabolism — the process in which the body converts energy in food into energy available for running cellular processes [1]. Metabolism converts food into building blocks for proteins, lipids, nucleic acids, and carbohydrates, and also eliminates metabolic waste. The study of metabolites, or metabolomics [1], has garnered interest in the biomedical community in recent years due to its potential to improve early disease detection and intervention through the analysis of relevant biomarkers. Certain metabolites have been found to be correlated with the prediction or progression of diseases such as Alzheimer's [2], ovarian cancer [3], colon cancer [4], and breast cancer [5], and have shown potential for improving treatment efficacy in patients with rheumatoid arthritis [6], [7]. The study of metabolomics has been growing steadily due to its promise in advancing early disease diagnosis and accelerating the historically decades-long path to innovation in clinical interventions.

Using features of a dataset to determine correct classification is a well-studied area of statistics. A wide variety of techniques have been developed, including LDA (Breiman et al. [8]), Classification Trees (Mondon, Camille [9]), Random Forests (Ho [10]), Support Vector Machines (SVM) (Cortes [11]), as well as Artificial Neural Networks (ANN) (Sarker [12]). Applications of ANN to genomics research are discussed in Lopez [13], and a good overview is provided by Zou [14].

In addition to classification, feature selection is very important. Determining which features from a set of features contribute to classification is more difficult. The literature includes a few works in this area using SVM, but most are limited to smaller numbers of features. Heinemann [15] uses SVM with a reverse elimination scheme to identify features associated with classification. Li et al. [16] use a genetic algorithm with an SVM to search through a small set of genetic features for classification, although it is not likelihood-based and is applied to microarray data. Tapak et al. [17] use a two-line genetic algorithm with Support Vector Machines to analyze gene expression data for psoriasis classification.

Recent advances in biomedical feature selection have shown the potential of combining Support Vector Machines (SVM) with Genetic Algorithms (GA) to improve disease classification in high-dimensional datasets. These hybrid GA-SVM approaches have been successfully applied to problems such as cancer detection, gene expression analysis, and metabolite classification, often outperforming standard filter or wrapper methods when properly tuned [18], [19], [20], [21]. Recent studies have also shown that incorporating hybrid selection strategies can yield more compact and accurate feature sets [22], [23], motivating further innovation in the design of AI-driven feature search algorithms for biomedical data.

Currently, there are no likelihood-based methods that allow for posterior inclusion probabilities of the metabolites to be calculated. Posterior inclusion probabilities allow researchers to understand the relative importance of each metabolite. A novel artificial intelligence methodology is introduced to identify features that contribute to correct classification. The method uses Support Vector Machines combined with a Genetic Algorithm to perform a stochastic search through the feature space using a penalized likelihood-based approach. This results in inclusion probabilities for each feature. Furthermore, the Genetic Algorithm uses multiple parallel genetic lines that mutate at each time step, with crossover between lines at different intervals. Basic properties of the algorithm are studied through a simulation study with 72 different experimental conditions, each replicated 100 times.

This paper begins with an overview of the Qatar BioBank

data in Section II, which serves as the motivating example. Section III introduces the proposed algorithm, followed by Section IV, where we run simulation studies under varying conditions to assess the method's ability to identify metabolites linked to heart disease. In Section V, we compare our approach against existing methods. Section VI presents results from applying the method to the Qatar BioBank data, highlighting metabolites associated with heart disease. We conclude in Section VII with a discussion of the method and directions for future work.

II. Qatar BioBank Data

The Qatar Biobank was created in collaboration with Hamad Medical Corporation and Qatar's Ministry of Public Health in an effort to collect and consolidate Qatar health data, which includes metabolite samples. The Qatar Biobank makes the data accessible to allow scientists to use information about the region's healthcare landscape to guide medical innovation tailored to local needs, and to enable local healthcare workers to make informed decisions about patient care. The Biobank currently has 38,213 total participants, 30,570 of whom are Qatari citizens. To qualify for participation in the Qatar Biobank, individuals must either be Qataris or long-term Qatar residents who have lived in Qatar for at least 15 years. All participants must also be at least 18 years of age [24]. Metabolite samples collected by the Biobank are obtained from buffy coat, DNA, erythrocyte, PaxGene, plasma, RNA, saliva, saliva and RNA, serum, urine, and viable cell samples [24]. Samples are stored in a -80°C automated biostore or cryopreservation laboratory and are processed through a Waters ACQUITY ultra-performance liquid chromatography (UPLC) unit, a Thermo Scientific Q-Exactive mass spectrometer, heated electrospray ionization (HESI-II), and an Orbitrap mass analyzer [25].

The initial dataset includes 1,046 participants with 1,159 measured metabolites, all of whom are Qatari citizens. As many of the participants had multiple disease states, the data was reduced to 626 participants who had exactly one of the following disease states: Angina; Control; Heart Attack; Stroke; or High Cholesterol. Table I shows the distribution of these disease states across the sample, as well as the number of males and females in the sample.

TABLE I. Summaries of Disease State and Gender for the Qatar BioBank Data

Variable	Group	Count
Disease	Angina	9
	Control	291
	Heart Attack	8
	High Cholesterol	316
	Stroke	2
Gender	Female	303
	Male	323

Table II provides summary statistics for clinical and demographic variables such as Age, Systolic BP, Diastolic BP, HDL Cholesterol, LDL Cholesterol, HBA1C (%), Hemoglobin (g/dL), Creatinine (umol/L), Urea (mmol/L), Thyroid Stimulating Hormone (mIU/L), Red Blood Cell ($\times 10^6$/uL), White Blood Cell ($\times 10^3$/uL), and Pulse Wave Velocity for the Qatar BioBank data. The number of samples is $n_r = 626$ for all variables.

TABLE II. Summary Statistics for Clinical and Demographic Variables in the Qatar BioBank Dataset ($n_r = 626$) for all Variables

Variable	Mean	Median	StDev	Q_1	Q_3	Min	Max
Age	35.9	34.0	11.16	27.0	44.0	18.0	79.0
Systolic BP	110.4	110.0	11.22	102.0	118.2	85.0	139.0
Diastolic BP	71.6	71.0	8.07	66.0	99.0	43.0	89.0
HDL Cholesterol	1.36	1.31	0.36	1.09	1.56	0.41	2.87
LDL Cholesterol	3.32	3.40	0.99	2.55	4.00	1.00	8.84
HBA 1C%	5.37	5.40	0.40	5.10	3.60	4.20	6.40
Hemoglobin g/dl	13.66	13.70	1.73	12.50	15.00	6.80	19.70
Creatinine umol/L	67.98	67.00	15.36	56.00	79.00	34.00	140.00
Urea mmol/L	4.42	4.4	1.27	3.60	5.10	1.20	11.00
Thyroid Stimulating Hormone mIU/L	1.74	1.44	1.13	1.03	2.10	0.010	9.61
Red Blood Cell $x10^6$/ul	4.89	4.90	0.56	4.50	5.20	3.30	7.20
White Blood Cell $x10^3$/ul	6.61	6.40	1.74	5.50	7.50	2.80	14.60
PulseWave Velocity	10.68	10.30	2.55	9.10	11.60	0.00	25.80

III. Method

To identify which metabolites contribute to the correct classification of heart disease, computational techniques such as Support Vector Machines and Genetic Algorithms are used in combination. This section provides details of a multi-line Genetic Algorithm with crossover, implemented using Support Vector Machines as the classifier.

A. Support Vector Machines

Support Vector Machine (SVM) is a type of supervised learning algorithm used in machine learning to solve classification and regression problems. The main objective of the SVM algorithm is to find the best possible line, or decision boundary, that separates the data points of different classes. This boundary is called a hyperplane when working in higher-dimensional feature spaces. The main idea behind SVM is to maximize the margin, which is the distance between the hyperplane and the closest data points from each category that needs to be classified. This makes the data easier to classify.

Support Vector Machine (SVM) has different types, such as linear SVM, which separates the data with a straight line, and nonlinear SVM, which is used when the data cannot be separated by a line and is more complex. SVM transforms the input data into a higher-dimensional feature space. This transformation makes it easier to separate and classify the data. To achieve this, SVM uses a kernel function that enables it to implicitly calculate the dot product between the transformed feature vectors and avoid unnecessary computations. Examples of kernels include linear kernels, polynomial kernels, and radial basis functions (RBF). These kernels are very important and help capture complex relationships and patterns in the data.

In terms of a mathematical formulation, SVM separates the data by a hyperplane given by

$$< W, \Phi(x) > + b = 0, \tag{1}$$

corresponding to the decision function $f(x)$ given by:

$$f(x) = sign(< W, \Phi(x) > +b), \tag{2}$$

where W is the weight vector, Φ is an implicit mapping of the input data into a higher dimensional feature space defined by a kernel function subject to the decision function and b is the bias term, which is an offset parameter that allows the hyperplane to be shifted away from the origin.

Vapnik [26], [27] showed that the optimal, in terms of classification performance, is the hyperplane with maximal margin of separation between the classes. This can be done by constructing and solving a constrained quadratic optimization problem whose solution W has an expansion $W = \Sigma_i \alpha_i \Phi(x_i)$ in terms of the a subset of training patterns that lie on the margin, where $\alpha_i \geq 0$ is the Lagrange multiplier associated with the training samples x_i. The training patterns are called support vectors and carry all the relevant information about the classification problem. For more details, we refer the reader to [28], [29], [30], [31], [32], [33].

Moreover, the SVM can generate class probabilities as an output. In this case we use a sigmoid function given by

$$P(y = 1 \mid f) = \frac{1}{1 + e^{Af+B}}, \tag{3}$$

which is fitted to the decision values f of the binary SVM classifiers, A and B are parameters to be estimated by minimizing the negative log-likelihood function and f is the decision function. We can extend the class probabilities to the multi-class case where all binary classifiers class probabilities output can be combined as one problem [34], [35], [36], [37], [38], [39], [40]. In our case, we set up the SVM and consider the following optimization problem.

$$\begin{aligned} \text{minimize} \quad & t(\{\mathbf{w}_n\}, \zeta) = \frac{1}{2} \sum_{n=1}^{k} \|\mathbf{w}_n\|^2 + \frac{C}{m} \sum_{i=1}^{m} \zeta_i \\ \text{subject to} \quad & \langle \Phi(x_i), \mathbf{w}_{y_i} \rangle - \langle \Phi(x_i), \mathbf{w}_n \rangle \geq b_i^n - \zeta_i, \quad i = 1, \ldots, m \\ \text{where} \quad & b_i^n = 1 - \delta_{y_i,n} \end{aligned} \tag{4}$$

where $\mathbf{w_n}$ is the weight vector associated with class n,ζ is a slack variable introduced to allow some misclassifications, and ζ_i is a slack variable associated with each training sample x_i, and b_i^n is associated with margin between the decision hyperplane and the correct class n for the training example x_i. The decision function is given by:

$$\text{argmax}_{n=1..k} < \Phi(x_i), \mathbf{w_n} > \tag{5}$$

For the search algorithm the Bayesian Information Criterion (BIC) will be used which is given by (Schwarz [41]):

$$BIC(\mathbf{D}|m_i) = -2ln(L(\mathbf{D}|m_i)) - pln(n) \tag{6}$$

where $L(D|m_i)$ is the likelihood of the data $\mathbf{D}$ using model m_i, defined by the features in the model and p is the number of features in the model. This is a penalized likelihood approach that will favor a SVM utilizing a smaller number of features.

B. Genetic Algorithm

For the genetic algorithm, a population of five genetic lines is utilized. To initialize these lines, each metabolite is selected to be part of the model with probability 1/2. Once initialized, at each step, each model m_i is mutated by randomly selecting one metabolite and creating a candidate model m_i, where if the metabolite is already in the model, it is removed; if it is absent from the model, it is added. The model is then tested to determine if it increases the likelihood. If it does increase the BIC, then the metabolite will stay in the model; otherwise, it will stay with probability $\xi = \frac{BIC(\mathbf{D}|\mathbf{m_i})}{BIC(\mathbf{D}|\mathbf{m_i})}$. If not, model m_i will be retained at this step. This allows the algorithm to move between regions of high probability.

The crossover step for models m_i with m_j is done by randomly selecting metabolites with probability 1/2, and the chosen metabolite states (included or excluded) are copied from the other model. For example, if one of the randomly selected metabolites is in m_i and not in m_j, then this metabolite will be removed from m_i and added to m_j. Similarly, if the metabolite is in neither m_i nor m_j, then it will not appear in either m_i or m_j. Likewise, if the metabolite is in both m_i and m_j, then it will remain in both. Note that at each crossover step, there is no randomization step to determine if the crossover was advantageous to the BIC.

The following crossover schedule is used for the five models: every 11 steps, models m_1 with m_2, as well as models m_3 with m_4, undergo crossover steps. This crossover process is done at every 23rd step for m_1 with m_3 and m_2 with m_4. At every 37th step, m_1 with m_4 and m_2 with m_3 are crossed over. At every 47th step, the crossover process is done with m_1 and m_5, immediately followed by m_2 and m_5. This allows m_5 to have many more mutation steps between crossover, while both m_1 and m_2 impact m_5 at the same time. The multiple models with crossover at this schedule allow each model to have several mutation steps before a crossover step.

Fig. 1 shows the crossover schedule in visual format. Initialize only starts at the beginning, and then the schedule is repeated at each interval labeled by Steps. The diagram shows that crossovers at every 44, 46, and 47 are close together, meaning there is a large amount of mixing between all models.

This process is run for $n_{step} = 1{,}000$ steps. At each step, the metabolites included in or excluded from the model are recorded for each model. At the end of the 1,000 steps, the number of times each metabolite appeared in each model is determined, and the total number of times the metabolite appeared in any model is calculated. This total is then used to compute the Inclusion Probability by dividing it by 5,000, the total number of opportunities the metabolite had to appear in a model. Numerous test runs were performed, and $n_{step} = 1{,}000$ appeared sufficient, as by the final step, almost all models had converged to the same solution, meaning they included the same metabolites. If a metabolite has an inclusion probability greater than 0.5, it is deemed important for classification.

The method was coded in R (v4.3.0) [42], using support vector machines from the `e1071` (v1.7.13) package [43], [44]. The algorithm takes approximately 8.3 to 80.4 minutes to complete 1,000 steps on an Apple M2 Pro processor with 32GB of RAM, depending on the sample size n_r, with larger samples taking progressively longer.

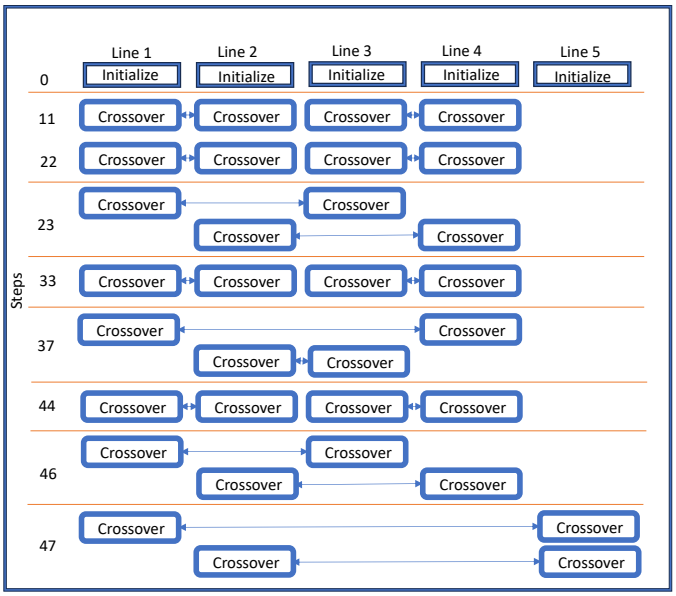

Fig. 1. Genetic algorithm crossover schedule for the first 50 steps. Crossover intervals are preserved across all iterations. Note that mutations occur at every step.

IV. Simulation Study

To study the performance of the algorithm, a simulation study is conducted by varying the sample size n_r, the number of significant metabolites n_s, the number of candidate metabolites n_c, the magnitude of effect (effect size), and the overall variance of the relationship, σ^2. To generate each simulated metabolite dataset, an $n_r \times n_c$ matrix $\mathbf{X}$ was constructed, where each element $\mathbf{X}ij$ follows a standard normal distribution: $\mathbf{X}ij \sim N(0,1)$. The following two effect profiles are used to assign the influence of each metabolite on classification:

$$\beta_1 = \begin{cases} M(1-1/(n_s+1))w & w < n_s, \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$\beta_2 = \begin{cases} M(1/n_s+1))w, & w < n_s \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $w = (1, 2, .., n_c)$. Here β_1 induces a decreasing effect across the first n_s metabolites while maintaining a positive effect. Similarly, β_2 induces an increasing effect across the first n_s metabolites. This is converted into a two linear relationships defined by:

$$\begin{aligned} z_1 &= \mathbf{X}\beta_1 \\ z_2 &= \mathbf{X}\beta_2 \end{aligned} \quad (9)$$

with two additional relationships with system noise added:

$$\begin{aligned} z_1^* &= \mathbf{X}\beta_1 + \epsilon_1 \\ z_2^* &= \mathbf{X}\beta_2 + \epsilon_2 \end{aligned} \quad (10)$$

with $\epsilon_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\epsilon_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with $Cov(\epsilon_1, \epsilon_2) = 0$. Here σ^2 represents the inherent noise in the system designed to confuse the classifier. To obtain the true simulated classifications the following rule is used:

$$y_{true} = \begin{cases} A & z_1 < -1, z_2 < -1 \\ S & z_1 > 1, z_2 > 1 \\ R & z_1 > 1, z_2 < -1 \\ T & z_1 < -1, z_2 > 1 \\ C & \text{otherwise} \end{cases} \quad (11)$$

Fig. 2 shows an example dataset for this simulation. Here, the number of observations is $n_r = 1{,}000$, the number of candidate metabolites is $n_c = 200$, the number of significant metabolites is $n_s = 5$, the magnitude of the effect is $M = 5$, and the system variance is $\sigma^2 = 0.5$, using equation (11) as the true classification boundaries. Notice that this includes a large number of controls, class A, and class S, with smaller numbers of R and T. This inequality in distribution is desired for this simulation, as it reflects the inequality typically found in metabolomic data. Furthermore, the simulation allows mixing of A, R, S, and T with the control C. However, there is rare mixing across the disease states A, R, S, and T. This is a property of the metabolomic data under consideration, as the participants were chosen based on exhibiting only one of the disease states or being a control. Hence, the data does not include individuals with multiple disease states. Furthermore, it is assumed that a larger increase among a number of metabolites corresponds to an additive effect toward classification. Note that the goal of the study is not to achieve correct disease classification but to identify which metabolites contribute to the correct classification.

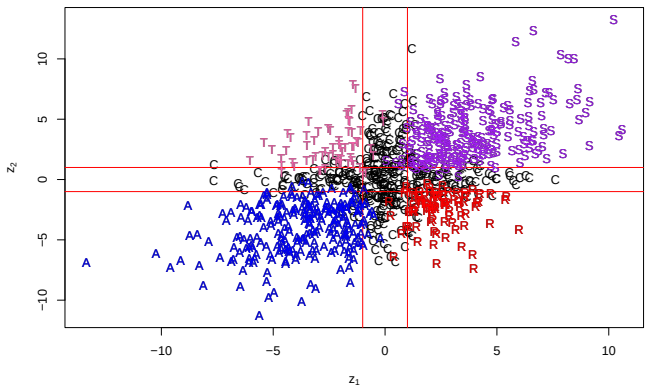

Fig. 2. Example simulated dataset $n_r = 1,000$, $n_c = 200$, $n_s = 10$, Magnitude $= 2$, and $\sigma^2 = 0.5$. The red lines demark the true classification boundaries set by equation (11).

A simulation study is performed to consider sample sizes of $n_r = 300,\ 500$, and 1,000, number of significant metabolites $n_s = 5$ and 10, number of candidate metabolites $n_c = 100$ and 200, magnitudes $= 1, 2$, and 4, and variance $\sigma^2 = 0.1$ and 0.5. This gives 7,200 simulated datasets and analyses, which were conducted on an AMD Ryzen Threadripper PRO 3975WX 3.5GHz 32-Core sWRX8 processor with 64GB RAM. For each of the combinations, 100 datasets were simulated, the proposed algorithm was run, and the Correct Identification Rate (CIR)

was calculated by counting the number of metabolites correctly identified in the model by the algorithm, divided by the number of true metabolites n_s in the model. The resulting correct metabolite identification rates were then averaged over the 100 datasets. The False Identification Rate (FIR) is also calculated by adding the number of incorrect metabolites and dividing by the total number of metabolites selected by the algorithm, then averaging across all 100 datasets.

Table III shows the results for the simulation study. Results are reported as CIR (FIR) in the table. Notice for $n_s = 5$ and $n_c = 100$, the correct identification rate was near 1.000 for all magnitudes and variances, with the exception of $M = 1$, $\sigma^2 = 0.5$, and $n_r = 300$. Even in this exception, a 0.998 correct identification rate was achieved. When $n_s = 5$ and $n_c = 200$, the correct identification rate goes down to 0.918 for $n_r = 300$, $M = 1$, and $\sigma^2 = 0.5$, but is near 1.000 for $n_r = 500$ and 1,000. It should be noted that in this case, there is a small sample size $n_r = 300$ with a relatively large number of candidates $n_c = 200$, small effect size $M = 4$, and large variance $\sigma^2 = 0.5$. One would expect, in this situation, the method to have a more difficult time correctly identifying all the metabolites. When $n_s = 10$ and $n_r = 10$, the method seems to struggle to correctly identify all the contributing metabolites, with a rate as low as 0.826. While this is not terrible, it shows the sensitivity of the algorithm to cases where the pattern is not as clearly defined. Whereas when $n_r = 500$, the lowest correct identification rate is 0.935, and when $n_r = 1{,}000$, the lowest correct identification rate is 0.993, neither of which would be considered bad. This corresponds to the general idea that the larger the sample size, the more information is available for detecting patterns.

Table III also shows the FIR for each experimental combination. Note that FIR values are in parentheses. Notice that there is a general increase in the false positives as the number of candidate metabolites is increased. For example, when $n_s = 5$, $n_c = 100$, $M = 1$, and $\sigma^2 = 0.1$ for $n_r = 300$, the false positive rate is 0.054 compared to the same setting with $n_r = 200$, where the false positive rate is 0.177. This is persistent across all simulation settings. This should be expected, as the space that needs to be searched through is much larger and, hence, there are more opportunities for a Type I error to occur. Also notice that as the variance σ^2 increases, the false positive rate also increases. For example, when $n_s = 10$, $n_c = 200$, $M = 1$, and $\sigma^2 = 0.1$, the false positive rate is 0.261, and when $\sigma^2 = 0.5$, the false positive rate is 0.268. This is also to be expected, as in general, in any classification algorithm, the larger the variance (noise), the more difficult it becomes for the algorithm to detect the correct predictors. It also appears that as the magnitude increases, the false positive rate also increases. However, the difference between the false identification rates due to magnitude seems to be less than 0.05, whereas the number of candidates n_c tends to produce a difference of 0.2.

As a screening algorithm, this method seems to be a good choice, as the correct identification rates are high and the number of false identifications is reasonable. The goal of any screening method is to ensure that the correct metabolites are identified. A simulation study should be conducted to determine the correct number of steps for the Genetic Algorithm, the threshold cutoff, the effect size that can be detected, etc.

TABLE III. PROPORTIONS OF CORRECTLY AND FALSELY IDENTIFIED METABOLITES (CIR/FIR) ACROSS SIMULATION SETTINGS. PROPORTION OF CORRECTLY IDENTIFIED METABOLITES ARE BASED ON 100 SIMULATED DATASETS

				n_r		
n_s	n_c	Magnitude	σ^2	300	500	1,000
5	100	1	0.1	1.000 (0.054)	1.000 (0.066)	1.000 (0.090)
		1	0.5	0.998 (0.120)	1.000 (0.139)	1.000 (0.185)
		2	0.1	1.000 (0.086)	1.000 (0.103)	1.000 (0.139)
		2	0.5	1.000 (0.119)	1.000 (0.145)	1.000 (0.203)
		4	0.1	1.000 (0.137)	1.000 (0.158)	1.000 (0.223)
		4	0.5	1.000 (0.149)	1.000 (0.164)	1.000 (0.239)
	200	1	0.1	0.950 (0.177)	0.986 (0.189)	1.000 (0.181)
		1	0.5	0.918 (0.247)	0.980 (0.250)	1.000 (0.203)
		2	0.1	0.996 (0.214)	1.000 (0.184)	1.000 (0.175)
		2	0.5	0.988 (0.237)	1.000 (0.197)	1.000 (0.202)
		4	0.1	0.996 (0.264)	1.000 (0.241)	1.000 (0.233)
		4	0.5	0.988 (0.276)	1.000 (0.249)	1.000 (0.239)
10	100	1	0.1	0.958 (0.106)	0.998 (0.084)	1.000 (0.099)
		1	0.5	0.951 (0.141)	0.997 (0.131)	1.000 (0.178)
		2	0.1	0.989 (0.135)	1.000 (0.154)	1.000 (0.195)
		2	0.5	0.979 (0.151)	1.000 (0.158)	1.000 (0.213)
		4	0.1	0.984 (0.169)	0.999 (0.177)	1.000 (0.234)
		4	0.5	0.975 (0.181)	0.999 (0.177)	1.000 (0.232)
	200	1	0.1	0.837 (0.261)	0.945 (0.199)	0.997 (0.178)
		1	0.5	0.826 (0.268)	0.938 (0.232)	0.993 (0.210)
		2	0.1	0.913 (0.274)	0.982 (0.238)	0.995 (0.247)
		2	0.5	0.907 (0.287)	0.983 (0.259)	1.000 (0.249)
		4	0.1	0.892 (0.309)	0.973 (0.293)	0.998 (0.276)
		4	0.5	0.876 (0.304)	0.971 (0.292)	0.998 (0.285)

However, this is beyond the scope of this study, as it is exploratory and would detract from the presentation of the method presented. More on extended simulation studies is given in the Discussion.

V. COMPARISON WITH OTHER METHODS

To determine the performance of our Support Vector Machine Artificial Intelligence (SVMAI) approach compared with other modern techniques designed to answer this type of question, such as Random Forests (RF) and Partial Least Squares - Discriminant Analysis (PLSDA). Both of these approaches are designed to perform both feature selection and classification simultaneously, which is the goal of the SVMAI approach.

The RF technique can be traced back to the basic classification trees of Breiman [8]. The big issue with classification trees is that they are sensitive to the initial feature selected to cut along. Hence the need for a forest of classification trees, which use randomization to determine which feature to cut along at each step. This is a very popular approach for feature selection as well as prediction. It has been used widely in metabolomics for feature selection; see [45] for use in urinary metabolomics, [46], who used the technique on Omega Fatty Acid Pathways, and [47], who provides a very good overview of the technique applied to lipid metabolites. For this work,

the randomForest version 4.7-1.1 package [48] in R 4.4.1 [42] is used.

PLSDA is an extension of standard Linear Discriminant Analysis, which seeks to first find the Partial Least Squares lines to fit each group, then determine the best discriminant line for classification. A review and comparison of RF, PLSDA, and standard SVM algorithms in metabolomics is provided by [49]. For more details on the PLSDA algorithm, see [50], [51], and a great simplistic discussion is given by [52].

A small simulation study was designed with sample sizes of $n = 300, 400$, and 500, with the number of true significant metabolites $n_t = 5$ and 10, and 100 metabolites to search through. The magnitude of the effect is 2.0, and the magnitude standard deviation is 0.5. The simulated data uses the same protocol as the simulation study in Section IV. To evaluate the performance, the %Correct, %Incorrect, and Overall Classification Accuracy are considered. Here, %Correct is the proportion of truly significant metabolites detected, and %Incorrect is the proportion of truly insignificant metabolites detected. Hence, a high %Correct indicates that the approach has the ability to find the true features. A high %Incorrect corresponds to a considerable number of insignificant metabolites being selected.

Table IV shows the results of this simulation study. Notice that SVMAI has extremely high correct detection rates, with the smallest at 0.997. In addition, the Classification Accuracy of the SVMAI technique is consistently 1.0, meaning 100% correct at classifying the heart health status. The %Incorrect for the SVMAI approach tends to be lower than the RF approach but much higher than the PLSDA technique. Of all the techniques considered here, the RF approach performs the worst, with low %Correct rates, high %Incorrect rates, and low Classification Accuracy. From this study, one can see that SVMAI is very good at determining the correct metabolites with high classification accuracy. However, it does select incorrect metabolites at a higher rate than PLSDA, but generally lower than RF.

TABLE IV. RESULTS OF SIMULATION STUDIES TO COMPARE ACROSS SVMAI, RF AND PLSDA TECHNIQUES FOR SCREENING. SAMPLE SIZES $n_s = 300, 400, 500$, NUMBER OF TRUE SIGNIFICANT METABOLITES $n_t = 5, 10$ WITH % CORRECT METABOLITES DETECTED, % INCORRECT METABOLITES DETECTED AND % CLASSIFICATION ACCURACY. BASED ON 100 SIMULATED DATASETS

		% Correct			% Incorrect			Classification Accuracy		
n_s	n_t	SVMAI	RF	PLSDA	SVMAI	RF	PLSDA	SVMAI	RF	PLSDA
300	5	1	0.8	0.798	0.123	0.553	0.23	1	0.539	0.851
	10	0.988	0.857	0.716	0.158	0.175	0.19	1	0.456	0.853
400	5	1	0.8	0.8	0.128	0.454	0.05	1	0.545	0.803
	10	0.997	0.887	0.761	0.143	0.156	0.05	1	0.470	0.806
500	5	1	0.8	0.8	0.141	0.398	0.06	1	0.552	0.778
	10	0.999	0.897	0.806	0.157	0.117	0.03	1	0.482	0.777

VI. QATAR BIOBANK RESULTS

For the Qatar BioBank data, there were $n_c = 223$ fully observed metabolites across $n_r = 626$ participants. The algorithm was performed on the dataset using 1,000 steps in the genetic algorithm. The computation time was approximately 12 minutes on an Apple M2 Pro processor with 32GB of RAM. Table V shows the metabolites with Inclusion Probabilities greater than 0.9 for the Qatar BioBank data concerning Heart Disease. Out of the 223 candidate metabolites, 37 had an inclusion probability above 0.9, and 48 had an inclusion probability above 0.5 (not shown). Notice that many of the metabolites have an inclusion probability above 0.999, indicating they were present in almost all of the models across the five genetic algorithm lines. Of particular interest is the fact that there are eight metabolites with the "X" prefix, which indicates an unknown chemical identity [53]. Also of note is that circulating BCAAs (leucine/isoleucine, valine, glutamate/glutamine, proline, and methionine) have been associated with predicting risk of coronary artery disease [54], [55], indicating the validity of our method in capturing previously known metabolites associated with heart disease.

TABLE V. METABOLITES AND THEIR INCLUSION PROBABILITIES FOR THE QATAR BIOBANK HEART DISEASE DATASET. ALL METABOLITES WITH AN INCLUSION PROBABILITY GREATER THAN 0.9 ARE PRESENTED IN DESCENDING ORDER OF INCLUSION PROBABILITY

Metabolite	Inclusion Probability
X - 23636	1.000
N1-methyladenosine	0.999
5alpha-pregnan-3beta,20alpha-diol disulfate	0.999
2,3-dihydroxy-5-methylthio-4-pentenoate (DMTPA)*	0.999
X - 11372	0.999
carnitine	0.999
2-hydroxyoctanoate	0.999
6-bromotryptophan	0.999
alanine	0.999
p-cresol sulfate	0.999
phosphoethanolamine	0.999
X - 14056	0.999
ornithine	0.999
deoxycarnitine	0.999
arginine	0.999
metabolonic lactone sulfate	0.999
X - 11880	0.999
X - 23639	0.999
1-arachidonoyl-GPI (20:4)*	0.999
choline	0.999
X - 21258	0.999
proline	0.999
dimethylarginine (SDMA + ADMA)	0.999
alpha-hydroxyisocaproate	0.999
indoleacetate	0.999
X - 24425	0.999
1-oleoyl-GPE (18:1)	0.999
N-delta-acetylornithine	0.999
5alpha-androstan-3beta,17beta-diol disulfate	0.999
taurocholenate sulfate*	0.999
pregnenediol sulfate (C21H34O5S)*	0.999
2-hydroxypalmitate	0.999
gamma-glutamylcitrulline*	0.999
X - 21364	0.999
hydroxy-CMPF*	0.998
creatinine	0.995
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	0.931

VII. Conclusion

This work shows the ability to use a genetic algorithm approach with multiple lines and frequent crossbreeding in conjunction with support vector machines using a likelihood approach to create an AI method to screen metabolites for heart disease. The method proposed performs well under simple simulation studies across a large number of scenarios. The method allows for high correct identification rates of metabolites (above 85%). The false positive rate is moderately high, ranging from 5% to 30%. Hence, the algorithm would be good for screening purposes to locate candidate metabolites for further study. The results of the algorithm, when applied to the Qatar BioBank data, allow for 223 metabolites to be screened down to 37, all of which have an inclusion probability above 0.9. As the algorithm has a reasonably high false discovery rate, one should be careful to make declarative statements that all of the metabolites identified are truly contributing to the disease state. Instead, one should state that there is preliminary evidence that these metabolites may be related to these disease states. Furthermore, since the data is dominated by Control and High Cholesterol participants, the results most likely have selected metabolites that are biased toward these two states. However, future studies can verify or refute these claims. Knowing which metabolites contribute to heart disease will allow for assays to be developed that can indicate which people are likely to have heart disease, which can lead to preventative care to reduce the risk of heart disease.

Additional work could be done to study the performance of the algorithm on more complex datasets. The work here only considered a small number of disease states to mimic the Qatar BioBank data; however, a higher number of disease states and multiple disease states should be studied. Also, the interaction of metabolites was not considered here. Work by Boone et al. [56] and Lee and Boone [57] shows how to explore a restricted space when interactions are present for linear regression models. A similar approach could be developed here.

In the simulation study presented, the Genetic Algorithm only took 1,000 steps due to computational times. Another simulation study should be conducted to determine if the number of steps can reduce the false positive rates. This study should attempt to determine a rule that one could apply to determine the adequate number of steps to achieve an acceptable false positive rate. In this study, one should consider the ratio of candidate metabolites to number of participants, $\frac{n_c}{n_r}$, to look for a general pattern. Also, the threshold to be deemed important could be adjusted from 0.5 to a higher value to become more stringent and potentially reduce the number of false positives. Furthermore, the SVMAI method is superior to PLSDA and RF for Classification Accuracy and %Correct metabolites selected and should be considered when conducting metabolomic studies where classification is of interest.

Another item that could be considered is how to address the issue of missing data. Many of the metabolites have missing values for various participants, which in this study have been omitted. Only metabolites observed in all participants were used here. As the computational complexity here is high, several approaches could be used to *impute* the missing values, such as mean imputation, the EM algorithm, and possibly multiple imputation.

Acknowledgment

Ryad Ghanam and Edward Boone would like to thank Qatar Foundation and Virginia Commonwealth University in Qatar for their support through the Mathematical Data Science Lab. Faten S. Alamri would like to thank Princess Nourah bint Abdulrahman University for their support through grant number PNURSP2024R346. The authors would to acknowledge the data received from Qatar Biobank.

Funding

This research was funded by Virginia Commonwealth University in Qatar through the Mathematical Data Science Lab HQ1260 and Princess Nourah bint Abdulrahman University through Project number PNURSP2024R346

Authors' Contribution

Conceptualization, E.L., R.A. and F.S.; methodology, E.L.; software, E.L. and R.A.; validation, E.L., R.A., F.S., and E.B; data curation, R.A.; writing—original draft preparation, E.L, R.A., F.S. and E.B. All authors have read and agreed to the published version of the manuscript.

References

[1] J. R. Idle and F. J. Gonzalez, "Metabolomics," *Cell metabolism*, vol. 6, no. 5, pp. 348–351, 2007.

[2] A. Yilmaz, T. Geddes, B. Han, R. O. Bahado-Singh, G. D. Wilson, K. Imam, M. Maddens, and S. F. Graham, "Diagnostic biomarkers of alzheimer's disease as identified in saliva using 1h nmr-based metabolomics," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 355–359, 2017.

[3] H. Shang, J. Zheng, and J. Tong, "Integrated analysis of transcriptomic and metabolomic data demonstrates the significant role of pyruvate carboxylase in the progression of ovarian cancer," *Aging (Albany NY)*, vol. 12, no. 21, p. 21874, 2020.

[4] D. Luo, Z. Shan, Q. Liu, S. Cai, Q. Li, and X. Li, "A novel seventeen-gene metabolic signature for predicting prognosis in colon cancer," *BioMed Research International*, vol. 2020, no. 1, p. 4845360, 2020.

[5] M. Huang, H.-Y. Li, H.-W. Liao, C.-H. Lin, C.-Y. Wang, W.-H. Kuo, and C.-H. Kuo, "Using post-column infused internal standard assisted quantitative metabolomics for establishing prediction models for breast cancer detection," *Rapid Communications in Mass Spectrometry*, vol. 34, p. e8581, 2020.

[6] I. Dudka, A. Chachaj, A. Sebastian, W. Tański, H. Stenlund, G. Gröbner, and A. Szuba, "Metabolomic profiling reveals plasma glyca and glycb as a potential biomarkers for treatment efficiency in rheumatoid arthritis," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 197, p. 113971, 2021.

[7] S. Takahashi, J. Saegusa, A. Onishi, and A. Morinobu, "Biomarkers identified by serum metabolomic analysis to predict biologic treatment response in rheumatoid arthritis patients," *Rheumatology*, vol. 58, no. 12, pp. 2153–2161, 2019.

[8] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[9] C. Mondon, "Classification and regression trees," 1984.

[10] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[12] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN computer science*, vol. 2, no. 6, pp. 1–20, 2021.

[13] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of artificial neural networks and deep learning," in *Multivariate statistical machine learning methods for genomic prediction*. Springer, 2022, pp. 379–425.

[14] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[15] J. Heinemann, A. Mazurie, M. Tokmina-Lukaszewska, G. J. Beilman, and B. Bothner, "Application of support vector machines to metabolomics experiments with limited replicates," *Metabolomics*, vol. 10, pp. 1121–1128, 2014.

[16] L. Li, W. Jiang, X. Li, K. L. Moser, Z. Guo, L. Du, Q. Wang, E. J. Topol, Q. Wang, and S. Rao, "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, 2005.

[17] L. Tapak, S. Afshar, M. Afrasiabi, M. K. Ghasemi, and P. Alirezaei, "Application of genetic algorithm-based support vector machine in identification of gene expression signatures for psoriasis classification: a hybrid model," *BioMed Research International*, vol. 2021, no. 1, p. 5520710, 2021.

[18] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, "Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information," in *2016 International Conference on Interactive Technologies and Games (ITAG)*. IEEE, 2016, pp. 70–76.

[19] Y. Zheng, X. Yang, M. Siddique, and G. Beddoe, "Simultaneous feature selection and classification based on genetic algorithms: an application to colonic polyp detection," in *Medical Imaging 2008: Computer-Aided Diagnosis*, vol. 6915. SPIE, 2008, pp. 123–131.

[20] N. Y. Moteghaed, K. Maghooli, and M. Garshasbi, "Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine," *Journal of medical signals and sensors*, vol. 8, no. 1, p. 1, 2018.

[21] L. Ali, I. Wajahat, N. Amiri Golilarz, F. Keshtkar, and S. A. C. Bukhari, "Lda–ga–svm: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Computing and Applications*, vol. 33, pp. 2783–2792, 2021.

[22] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8144–8150, 2011.

[23] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, p. 562, 2023.

[24] H. Al Kuwari, A. Al Thani, A. Al Marri, A. Al Kaabi, H. Abderrahim, N. Afifi, F. Qafoud, Q. Chan, I. Tzoulaki, P. Downey *et al.*, "The qatar biobank: background and methods," *BMC public health*, vol. 15, pp. 1–9, 2015.

[25] K. Suhre, N. Stephan, S. Zaghlool, C. R. Triggle, R. J. Robinson, A. M. Evans, and A. Halama, "Matching drug metabolites from non-targeted metabolomics to self-reported medication in the qatar biobank study," *Metabolites*, vol. 12, no. 3, p. 249, 2022.

[26] V. Vapnik and V. Vapnik, "Statistical learning theory wiley," *New York*, vol. 1, no. 624, p. 2, 1998.

[27] H. Cevikalp, "Best fitting hyperplanes for classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1076–1088, 2016.

[28] S. Abe, *Support vector machines for pattern classification*. Springer, 2005, vol. 2.

[29] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox. perception systèmes et information, insa de rouen, rouen, france (2005)."

[30] M. Awad and L. Khan, "Support vector machines," in *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2008, pp. 1138–1146.

[31] B. Caputo, K. Sim, F. Furesjo, and A. Smola, "Appearance-based object recognition using svms: which kernel should i use?" in *Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler*, vol. 2002, 2002.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[33] A. Gammerman, N. Bozanic, B. Schölkopf, V. Vovk, V. Vapnik, L. Bottou, A. Smola, C. Watkins, Y. LeCun, C. Saunders *et al.*, "Royal holloway support vector machines," *URL http://svm. dcs. rhbnc. ac. uk/dist/index. shtml*, 2001.

[34] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.

[35] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE transactions on neural networks*, vol. 13, no. 2, pp. 464–471, 2002.

[36] C.-J. Lin and J. J. Moré, "Newton's method for large bound-constrained optimization problems," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1100–1127, 1999.

[37] C.-J. Lin, R. C. Weng *et al.*, "Simple probabilistic predictions for support vector regression," *National Taiwan University, Taipei*, 2004.

[38] A. C. Lorena, A. C. De Carvalho, and J. M. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artificial Intelligence Review*, vol. 30, pp. 19–37, 2008.

[39] Y. Shiraishi and K. Fukumizu, "Statistical approaches to combining binary classifiers for multi-class classification," *Neurocomputing*, vol. 74, no. 5, pp. 680–688, 2011.

[40] A. Rocha and S. K. Goldenstein, "Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 289–302, 2013.

[41] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[42] R. C. Team, "Ra language and environment for statistical computing, r foundation for statistical," *Computing*, 2020.

[43] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin, "Misc functions of the department of statistics," *Probability Theory Group (Formerly: E1071), TU Wien*, 2015.

[44] A. Karatzoglou and I. Feinerer, "Kernel-based machine learning for fast text mining in r," *Computational Statistics & Data Analysis*, vol. 54, no. 2, pp. 290–297, 2010.

[45] N. Chen, H.-B. Wang, B.-Q. Wu, J.-H. Jiang, J.-T. Yang, L.-J. Tang, H.-Q. He, and D.-D. Linghu, "Using random forest to detect multiple inherited metabolic diseases simultaneously based on gc-ms urinary metabolomics," *Talanta*, vol. 235, p. 122720, 2021.

[46] V. H. Oza, J. K. Aicher, and L. K. Reed, "Random forest analysis of untargeted metabolomics data suggests increased use of omega fatty acid oxidation pathway in drosophila melanogaster larvae fed a medium chain fatty acid rich high-fat diet," *Metabolites*, vol. 9, no. 1, p. 5, 2018.

[47] A. Acharjee, J. Larkman, Y. Xu, V. R. Cardoso, and G. V. Gkoutos, "A random forest based biomarker discovery and power analysis framework for diagnostics research," *BMC medical genomics*, vol. 13, pp. 1–14, 2020.

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[49] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre, "A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding," *Analytica chimica acta*, vol. 879, pp. 10–23, 2015.

[50] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 1, pp. 3–25, 2010.

[51] K.-A. Lê Cao, S. Boitard, and P. Besse, "Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems," *BMC bioinformatics*, vol. 12, pp. 1–17, 2011.

[52] D. Ruiz-Perez, H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan, "So you think you can pls-da?" *BMC bioinformatics*, vol. 21, pp. 1–10, 2020.

[53] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski *et al.*, "Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information," 2012.

[54] S. H. Shah, J. R. Bain, M. J. Muehlbauer, R. D. Stevens, D. R. Crosslin, C. Haynes, J. Dungan, L. K. Newby, E. R. Hauser, G. S. Ginsburg *et al.*, "Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events," *Circulation: Cardiovascular Genetics*, vol. 3, no. 2, pp. 207–214, 2010.

[55] R. Yang, S. Wang, L. Sun, J. Liu, H. Li, X. Sui, M. Wang, H. Xiu, S. Wang, Q. He *et al.*, "Association of branched-chain amino acids with coronary artery disease: A matched-pair case–control study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 25, no. 10, pp. 937–942, 2015.

[56] E. L. Boone, S. J. Simmons, K. Ye, and A. E. Stapleton, "Analyzing quantitative trait loci for the arabidopsis thaliana using markov chain monte carlo model composition with restricted and unrestricted model spaces," *Statistical methodology*, vol. 3, no. 1, pp. 69–78, 2006.

[57] A. H. Lee and E. L. Boone, "A frequentist assessment of bayesian inclusion probabilities for screening predictors," *Journal of Statistical Computation and Simulation*, vol. 81, no. 9, pp. 1111–1119, 2011.