VNU Journal of Science: Comp. Science & Com. Eng., Vol. 40, No. 2 (2024) 62–80

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

An Integrated Method for Disease Risk Prediction Using Next-Generation Sequencing Data

Ta Van Nhan[1]*, Dang Trung Du[1], Nguyen Thi Hong Minh[1]

[1] *VNU University of Science, 334 Nguyen Trai, Hanoi, Vietnam*

Received 03 May 2024;
Revised 27 June 2024; Accepted 29 August 2024

Abstract: Disease screening has seen increased adoption owing to heightened health awareness among individuals. Traditionally, wet labs have served as the conventional approach for testing; however, recent strides in bioinformatics have facilitated genetic testing and disease risk detection through computational analysis of data. This study presents a preprocessing methodology tailored for next-generation sequencing (NGS) data, integrating advanced computational tools. Leveraging the inherent advantages of NGS technology, this methodology ensures the acquisition of high-quality data essential for model training. Consequently, machine learning algorithms and neural networks are deployed to accurately predict disease risk and identify significant genetic variants. The performance of the proposed methods is higher than that of previous research. Through rigorous analysis, we have identified a subset of the most significant 8 SNPs linked to obesity and 61 SNPs associated with type 2 diabetes, with 50 SNPs previously reported in studies. These findings contribute to an understanding of the genetic factors underlying these complex diseases.

Keywords: Disease Screening, Genetic Testing, Bioinformatics Analysis, Next-Generation Sequencing, Machine Learning, Feature Selection, Obesity, Type 2 Diabetes.

1. Introduction

Screening tests are commonly used in medicine to evaluate the likelihood of a particular disease within a defined population, and as health awareness grows, disease screening is becoming increasingly popular [1, 2]. While wet lab testing has been a longstanding practice globally, the diagnostic paradigm is experiencing a notable shift. The rapid evolution of sequencing technologies and computational methodologies propels this transformation. Consequently, the diagnostic process has transitioned towards dry laboratories, where bioinformatics analyses precede disease risk stratification and identification.

*Corresponding author.
E-mail address: tavannhan@gmail.com

https://doi.org/10.25073/2588-1086/vnucsce.2687

Some approaches exist for identifying disease risk, ranging from gene panels to polygenic risk scores (PRS), and employing machine learning techniques. One can readily provide predictions based on gene panels utilizing next-generation sequencing (NGS) for single-gene Mendelian diseases [3, 4]. An example of such a disease is breast cancer with the presence of BRCA1/BRCA2 mutant genes [5]. However, the accuracy of these predictions tends to be low when dealing with polygenic diseases, which involve multiple genes. Conditions like schizophrenia, autism, obesity, and diabetes fall into this category, involving numerous variants. To address this, scientists are developing a polygenic risk score method (PRS) that utilizes genome-wide association study (GWAS) data to stratify polygenic risks [6, 7]. While GWAS technology utilizes SNP arrays to obtain a large set of biallelic variants, it still falls short compared to the number of variants generated by NGS technology. Moving on to NGS data, which is generated from sequencing platforms such as Illumina, Ion Torrent, PacBio, and Nanopore [8]. To identify variants, the raw data from these sequencers must undergo a specific workflow, such as the Genome Analysis Toolkit (GATK) [9], DeepVariant [10, 11], or DRAGEN [12]. Within this workflow, it is possible to annotate the variants to gain insights into related gene functions. This information may be used to predict disease risk and provide recommendations in areas such as pharmacogenomics and nutrigenomics. However, simply stopping at annotation makes it challenging to comprehend complex disease risks. Therefore, we introduce an efficient workflow for complex human disease prediction. Throughout the data preprocessing and model training stages, two distinct challenges arise that require resolution. Firstly, the data may contain missing genotypes following the determination of genotypes for each variant, attributable to errors in the sequencing process or variant calling procedures. Consequently, preprocessing of this data is necessary to address and impute missing genotypes. Here, we integrate the GATK-based variant calling procedure with BEAGLE's data imputation procedure [13] for this purpose. Secondly, with the advent of high-throughput sequencing technologies, the availability of genomic data has increased exponentially, leading to datasets with thousands to millions of genetic features [14]. The issue at hand revolves around the selection of informative genetic features. Feature selection not only improves the accuracy and interpretability of predictive models [15] but also helps uncover genetic markers and pathways underlying complex traits. Henceforth, we employed feature selection techniques to reduce the dimensionality of genomic datasets and identify the subset of genetic variants linked to traits or diseases of interest. Empirically, we have validated that feature selection mitigates computational complexity and augments the performance of prediction models utilizing NGS data.

This paper presents a comprehensive workflow designed to integrate the GATK-based variant calling procedure, BEAGLE's data imputation protocol, and machine learning methods for predicting human disease risk. The proposed workflow is structured into distinct phases to ensure robustness and efficacy in processing raw data, conducting genotype data analysis, and constructing prediction models. In the prediction stage, we have employed established and novel feature selection methodologies to identify significant SNPs associated with the respective diseases. To validate the conjecture that model performance is enhanced when trained on datasets post-feature selection, we applied feature selection techniques to two datasets: obesity and type 2 diabetes. As a result, a model constructed using obesity data exhibits performance that surpasses that of previous research conducted on the same

dataset [16]. Moreover, this process identifies a subset of the most significant 8 SNPs associated with obesity and 61 SNPs associated with type 2 diabetes. These findings deepen our understanding of genetic drivers behind complex diseases and could enable precise interventions and personalized therapies.

In the following sections of the paper, we will present the materials and methods in section II. Implementation with two datasets is given in section III. Finally, we conclude the article in section IV.

2. Material and methods

2.1. Workflow description

We have proposed a comprehensive workflow integrating the GATK-based variant calling procedure [17], BEAGLE's data imputation procedure [18], and machine learning methods for human disease risk prediction. In the raw data processing phase, raw nucleotide sequences (in FASTQ format) with a quality score below 99% are trimmed. Subsequently, these sequences are mapped to a reference genome, generating BAM/SAM files. The bases in these files undergo recalibration to ensure accuracy in preparation for subsequent analysis. The Haplotype caller calls variants from all BAM/SAM files in the variant calling stage, producing genotype variant call format (GVCF) files [19]. Our workflow further consolidates all the GVCF files into a cohesive cohort dataset, represented as a VCF file.

The processing of target genotype data involves key steps to ensure data integrity and compatibility with downstream analyses[1]. Firstly, variant alleles are aligned with the reference human genome using alignment algorithms ensuring alignment consistency. Concurrently, filtering is applied to retain only bi-allelic sites. Secondly, procedures are implemented to identify and remove duplicate individuals within the target genotype data and between the target and reference panels. Additionally, duplicate variant detection addresses potential duplications arising from dataset-specific anomalies or errors post-alignment. Thirdly, a comprehensive comparison of allele frequencies (AF) is conducted between the target genotype data and the reference panel. Variants exhibiting significant discrepancies in allele frequencies are identified, with those exceeding 10 percentage points in AF difference or a log_2 fold change greater than 5 or less than -5 earmarked for exclusion. Fourthly, variants demonstrating highly discordant allele frequencies or those absent from the imputation reference panel are excluded, as they may introduce bias or inaccuracies, thereby ensuring data integrity and reliability. Lastly, genotype imputation for each chromosome was performed individually using BEAGLE 5.2.

Following the target data processing stage, we obtain the genotype-phenotype data, which is then utilized to construct a prediction model. Genotypes are encoded using the values 0, 1, and 2, representing the total differences between alleles one and two compared to the reference allele. Additionally, we incorporate the covariate of sex, assigning a value of 1 for males and 2 for females. The model's label is determined by the phenotype, which is encoded as 1 or 2, denoting whether an individual does not carry or carries a specific disease.

The final stage of the process revolves around predicting human diseases. Covariates, such as gender, are integrated into the target data and labeled based on metadata. The dataset is then partitioned into training and testing subsets. Subsequently, a range of feature selection techniques, encompassing Lasso, Recursive Feature Elimination (RFE), and hybrid methodologies, are used to discern significant features [20]. Following this, the model is trained with the selected features. It is important to emphasize that the current workflow is optimized

[1]https://www.protocols.io/run/genotype-imputation-workflow-v3-0-xbgfijw?step=1

for bi-allelic data and predominantly concentrates on SNPs as the primary features (see Figure 1).

2.2. Feature selection

Significant SNP selection enhances the accuracy and interpretability of predictive models and facilitates the identification of genetic markers and pathways associated with complex traits. These SNPs shall be chosen utilizing feature selection methodologies. Filter methods assess the relevance of features independently of the chosen learning algorithm. They typically rely on statistical measures or heuristics to rank features based on their correlation with the target variable [21]. Examples include the Pearson Correlation Coefficient, Mutual Information, Chi-Square Test, and ANOVA. Besides, wrapper methods select features based on their performance with a specific learning algorithm. They employ a search strategy, such as forward selection or backward elimination, to evaluate subsets of features by training and testing a model iteratively [22]. Common examples of wrapper methods include Recursive Feature Elimination (RFE), Forward Selection, and Backward Elimination (see Algorithm 1). Additionally, embedded methods integrate feature selection into the model training process. These methods optimize feature selection and model parameters jointly. Techniques such as Lasso, Ridge, and Dropout Regularization fall under this category [23]. Moreover, hybrid methods integrate different techniques to achieve better performance than individual methods [24] (see Figure 2).

In this study, we employ several methods categorized as follows. Penalized Logistic Regression belongs to embedded methods, while Decision Tree and Support Vector Machine Recursive Feature Elimination (RFE) are methods used by the wrapper group. Additionally, a Simple, Fast, and Efficient Feature Selection Algorithm (SFE) for high-dimensional data falls within hybrid methods.

2.2.1. Logistic Regression (LR)

LR incorporates two types of regularization. The first is L2 regularization, also known as "Ridge," which reduces the coefficients' magnitude. The second is L1 regularization, or "LASSO," which forces specific coefficients to zero. These regularization techniques are helpful for variable selection during the learning process. When dealing with a large number of SNPs compared to the number of samples, the combination of L1 and L2 regularization, referred to as "Elastic-Net," is particularly effective [25].

More specifically, the problem entails estimating the coefficients β_0 and β in order to minimize the loss function:

$$L(\lambda, \alpha) = -\sum_{i=1}^{n}(y_i log(z_i) + (1 - y_i) log(1 - z_i)) + \lambda((1 - \alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1) \quad (1)$$

where $z_i = 1/(1 + exp(-(\beta_0 + x_i\beta)))$, x represents genotype and covariates, n is the sample size, y is the disease status, λ and α are the two regularization hyperparameters.

In the case of PLR, features can be eliminated during the training process, but the retained number of SNPs cannot be directly controlled. Instead, PLR can automatically select the appropriate number of SNPs based on hyperparameters, specifically the balance parameter α between L1 and L2 regularization.

2.2.2. Decision Tree-Recursive Feature Elimination (DT-RFE)

In Decision Trees, information entropy is a crucial criterion for feature selection. Given a training dataset with samples requiring classification, the Decision Tree algorithm calculates the information entropy and progressively splits the dataset [26]. This process eventually separates each sample type

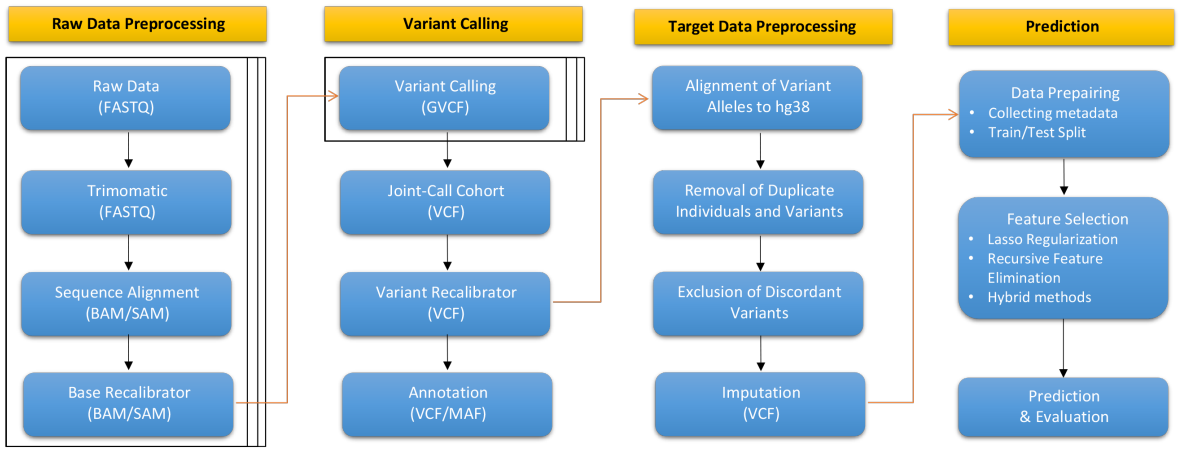

Figure 1. **Integration of GATK-based Variant Calling, BEAGLE's Data Imputation, and Machine Learning for Human Disease Risk Prediction**. A comprehensive workflow including raw data processing, variant calling, genotype data processing, and disease prediction stages. Data integrity is ensured through alignment algorithms, bi-allelic site filtering, duplicate removal, and allele frequency comparison. Disease prediction involves feature selection and training focusing on bi-allelic data and SNPs.

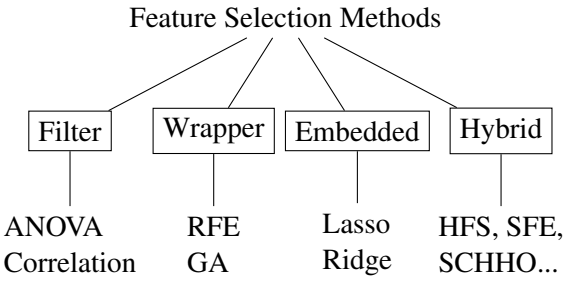

Figure 2. **Some categories of feature selection methods.** These comprise four groups: filter methods, wrapper methods, embedded methods, and hybrid methods. Below, various techniques belonging to each category are listed.

individually. Entropy quantifies the uncertainty of a random variable. Assume X is a random variable with a finite set of values, with a probability distribution represented as:

$$P(X = x_i) = p_i \tag{2}$$

Here, each value x_i corresponds to its respective probability p_i. The entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{3}$$

When there is a greater difference among the probabilities p_i, the entropy $H(X)$ is higher.

Assume the joint probability distribution of the random variables (X, Y) is:

$$P(X = x_i, Y = y_j) = p_{ij} \tag{4}$$

The conditional entropy $H(Y|X)$, representing the uncertainty of Y given X, is calculated as:

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i) \tag{5}$$

The information gain $G(D|A)$ of feature A with respect to the dataset D is defined as:

$$G(D|A) = H(D) - H(D|A) \tag{6}$$

Information gain measures the reduction in uncertainty about Y after learning feature X. However, information gain can bias partitioning

toward features with more values. To address this, the information gain ratio is used:

$$G_R = \frac{G(D|A)}{H(D)} \tag{7}$$

Consider a Decision Tree T with $|T|$ leaf nodes, where each leaf node t contains N_t samples, with N_{tk} samples belonging to class k. The entropy at node t is H_t, and α (≥ 0) is an optional parameter related to the penalty term. Thus, the loss function for T is:

$$L_\alpha(T) = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha|T| \tag{8}$$

where $H_t(T)$ is:

$$H_t(T) = -\sum_k \frac{N_{tk}}{N_t} \log\left(\frac{N_{tk}}{N_t}\right) \tag{9}$$

The first term of the loss function in (8) is:

$$C(T) = \sum_{t=1}^{|T|} N_t H_t(T) = -\sum_{t=1}^{|T|} \sum_{k=1}^{K} N_{tk} \log \frac{N_{tk}}{N_t} \tag{10}$$

Therefore, the loss function simplifies to:

$$C_\alpha(T) = C(T) + \alpha|T| \tag{11}$$

where $C(T)$ represents the model's prediction error, and $|T|$ reflects model complexity, and can be seen as a penalty term. The parameter α balances model complexity and prediction error.

As depicted in Algorithm 1, the Recursive Feature Elimination (RFE) method uses Decision Trees for training in multiple iterations. Through the weight coefficients obtained during training, better features are retained for subsequent training rounds. The RFE technique, which incorporates feature weights into prediction models, systematically reduces the size of the feature set under evaluation to select the most relevant features. The prediction models are initially trained using the original features, assigning a weight to each feature. Subsequently, the feature set is streamlined by removing the features with the smallest absolute weight values. This recursive process continues until the desired number of remaining features is reached.

2.2.3. Support Vector Machine-Recursive Feature Elimination (SVM-RFE)

Considering a training set of n points, denoted as $\{x_i, y_i\}$ $i = 1, \ldots, n$. Here, $y_i \in \{-1, 1\}$ represents the class label of the point x_i. The representation of the hyperplane is as follows:

$$x_i w + b = 0 \tag{12}$$

In the given equation, the weight vector is denoted as w, and the constant b represents the bias or displacement of the hyperplane.

The optimization process for improving the discriminatory function of the hyperplane can be formulated as a quadratic programming problem:

$$Minimize \;\; L_p = \frac{1}{2}\|w\|^2 \tag{13}$$

$$Subject\ to \;\; y_i(x_i w + b) - 1 \geq 0, \quad \forall i = 1, ..., N \tag{14}$$

We can transform it into a Lagrangian problem. Consequently, we can reframe the problem as follows:

$$Minimize\ L_p(\alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^{N} \alpha_i y_i (x_i w + b) + \sum_{i=1}^{N} \alpha_i \tag{15}$$

where $\alpha = [\alpha_1, ..., \alpha_N]^T$ and $\alpha_i \geq 0, \forall i = 1, ..., N$.

We can interpret this as a convex quadratic programming problem (15) with the corresponding dual formulation:

$$Maximize\ L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{16}$$

$$Subject\ to \;\; \sum_{i=1}^{N} \alpha_i y_i = 0, \forall i = 1, ..., N \tag{17}$$

$$\alpha_i \geq 0, \forall i = 1, ..., N \tag{18}$$

By employing the Karush-Kuhn-Tucker (KKT) conditions for optimality, we get:

$$w = \sum_{j=1}^{N} \alpha_j y_j x_j \tag{19}$$

SVM-RFE calculates ranking weights for all features and arranges them based on weight vectors [27]. It involves an iterative process where features are sequentially eliminated in a backward (see Algorithm 1).

2.2.4. Neural Network-Recursive Feature Elimination

Integrated Gradients is a method that provides a principled approach to quantify feature importance by attributing contributions to the input features based on their gradients concerning the model's output [28].

For a given function $F : \mathbb{R}^n \rightarrow [0, 1]$, with $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, the attribution of the prediction input x with respect to the baseline input x' is represented by a vector $A_F(x, x') = (a_1, ..., a_n) \in \mathbb{R}^n$, where a_i denotes the contribution or importance of x_i to the prediction $F(x)$. The computation of a_i can be achieved using the integrated gradients (ID) method through path integration. ID involves aggregating the gradients along a straight line connecting the baseline input x' to the input x.

Consider $\gamma = (\gamma_1, ..., \gamma_n) : [0, 1] \rightarrow \mathbb{R}^n$ as a smooth function that represents a path in $\mathbb{R}^n$ from the baseline x' to the input x. For instance, one has $\gamma(0) = x'$ and $\gamma(1) = x$. This path functions γ is obtained by integrating the gradients along the path $\gamma(\alpha)$, where $\alpha \in [0, 1]$. The path-integrated gradients along the i^{th} dimension for the input x are defined as follows:

$$ID_i^{\gamma}(x) = \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha \tag{20}$$

where $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F along i^{th} dimension at x. Early neural network interpretability used gradients to assign feature importance scores, refer to Algorithm 1 for the NN RFE using these scores.

Algorithm 1 Algorithm Generic RFE

Input: Training sample set X_0.
Output: Feature sort set R.

Initialisation : The original feature set S and feature ordering set $R = \{\}$.

while (S <> {}) **do**
  if using Decision Tree **then**
    Train the Decision Tree classifier and obtain the feature selection results using F-test (ANOVA) for individual variables;
    Calculate ranking criterion score;
    Identify the feature with the lowest ranking score;
  else if using SVM **then**
    Obtain the new training sample matrix based on the features: $X = X_0(1 : |S|)$;
    Train the SVM classifier;
    Calculate the weight: $w = \Sigma_k \alpha_k y_k x_k$;
    Calculate the sorting criteria: $c_i = (w_i)^2$;
    Find the feature with the minimum weight: $f = argmin(c)$.
  else if using Neural Network **then**
    Train the neural network model on the given dataset;
    Calculate Integrated Gradients for feature importance scores;
    Rank the features based on their importance scores;
    Select the top-k features or set a threshold on the importance scores.
  end if
  Update the sorted feature set: $R = \{R, S(f)\}$;
  Remove other features from S: $S = \{S/S(f)\}$.
end while
return Feature sort set R.

2.2.5. Simple, Fast, and Efficient (SFE)

SFE proposed a binary optimization problem where each feature can exist in one of two states: selected or non-selected. In particular, the algorithm uses a search agent M that shows each feature's state. Specifically, $m_j = 0$ indicates that the jth feature is not selected, while $m_j = 1$ indicates that it is selected. The algorithm employs two primary operators: the selection and non-selection operators. The selection operator transitions a feature's state from a non-selected state to a selected state, while the non-selection operator shifts a feature from a selected state to a non-selected state.

During the search process of SFE, both selection and non-selection operators are applied to the search agent M iteratively to improve the solution quality. The determination of whether to use the selection operator depends on the characteristics of the problem's search space and the current position of M within it. Typically, the non-selection operator is initially utilized on M during the feature selection process to conduct a global search, aiming to identify and convert irrelevant features into the non-selected mode.

The non-selection operator's effectiveness is controlled by the non-selection operator rate UR, which dictates the number of features subjected to its operation. From this rate, one can calculate UN, which represents the number of pseudo-random numbers generated for the operation of the non-selection operator using Equation (21), where *nvar* represents the dimensionality of the search space or the number of features in the dataset.

$$UN = \lceil UR \times nvar \rceil \tag{21}$$

In short, the non-selection operator is initially employed for a global search at the onset of the search process, placing the algorithm in the exploration phase. In subsequent search steps, the algorithm executes the exploitation phase in conjunction with the exploration phase [29]. To balance between the exploration and exploitation phases, a linear decrease in the value of UR is implemented. Equation (22) delineates the calculation of UR in the SFE algorithm, where UR_{max} and UR_{min} represent the initial and final values of UR, respectively.

$$UR = (UR_{max} - UR_{min}) \times \left(\frac{Max\_FEs - FEs}{FEs} \right) + UR_{min} \tag{22}$$

where Max_FEs represents the maximum number of function evaluations, and FEs denotes the current number of function evaluations performed by the SFE algorithm. The evaluation takes place after the training phase of machine learning models. In the initial approach, SFE was only applied to KNN. This study incorporates decision tree and SVM models with SFE.

Following a global search utilizing the non-selection operator, the selection operator is invoked if all features are transformed into the non-selected state. This operator enables the algorithm to re-select pertinent features whose state has changed. The process terminates after a pre-defined number of iterations, where each iteration involves a predetermined maximum number of function evaluations, referred to as Max_FEs.

2.2.6. Performance metrics

We use Sensitivity, Specificity, Accuracy, Matthews correlations coefficient (MCC), and The Area Under the Curve (AUC) for evaluating the model performance. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the outcomes used to define these metrics.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

MCC is a measure to assess the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. MCC values range from -1 to +1, where +1 indicates a perfect prediction, 0 is no better than a random prediction, and -1 indicates total disagreement between prediction and observation.

$$\text{MCC} = (TP \times TN - FP \times FN)/ \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

The Area Under the Curve (AUC) is a metric used to evaluate the performance of a binary classification model. Specifically, the AUC refers to the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) for various threshold settings of the classifier [30].

The confidence interval (CI) for a proportion can be computed based on the Wilson score interval [31, 32]. Given r is a number of successes, n is the total number of trials, α is a confidence level (for example 0.95 for 95% confidence), and z is Z-value corresponding to the desired confidence level (for example $z = 1.96$ for 95% confidence). The Wilson score interval for a proportion is defined as:

$$z = \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Here, Φ^{-1} denotes the inverse of the Cumulative Distribution Function (CDF) (or the quantile function) of the standard normal distribution.

$$A = 2r + z^2$$

$$B = z\sqrt{z^2 + 4r\left(1 - \frac{r}{n}\right)}$$

$$C = 2(n + z^2)$$

The lower and upper bounds of the confidence interval are then given by:

$$\text{CI}_{\text{lower}} = \frac{A - B}{C}$$

$$\text{CI}_{\text{upper}} = \frac{A + B}{C}$$

Here, r corresponds to the numerator and n corresponds to the denominator in the formulas identifying Sensitivity, Specificity, Accuracy, and MCC.

3. Implemetation

We perform experiments on two datasets: one on obesity and the other on type 2 diabetes.

3.1. Data Description

3.1.1. Obesity

The dataset to be used in this study comprises 139 individuals, as employed in the research conducted by H. Y. Wang et al. [16]. Among them, 75 individuals had obesity, defined as having a BMI (Body Mass Index) equal to or exceeding 27 kg/m², while 64 individuals were nonobese, with a BMI below 24 kg/m². It is important to note that all participants had no history of metabolic or endocrine disorders, were not undergoing steroid or surgical treatment for obesity, and no pregnant individuals were included.

The genomic DNA was isolated from whole blood samples using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) per the manufacturer's guidelines. The resulting raw data from Ion Torrent sequencing consists of single-end reads in FASTQ format.

3.1.2. Diabetes type 2

A population-specific genome for the indigenous Arab population of Qatar (QTRG) was obtained from the sequencing of 1,161 Qataris. Human subjects were recruited from Hamad Medical Corporation (HMC) and HMC Primary Health Care Centers in Doha, Qatar, following written informed consent and under protocols approved by the Institutional Review Boards of Hamad Medical Corporation and Weill Cornell Medical College in Qatar. 1,376 subjects underwent genome ($n = 108$) or exome ($n = 1,268$) sequencing, with 31 individuals sequenced by both methods for validation. Sequencing was performed using Illumina paired-end sequencing technology, with exome sequencing involving target enrichment utilizing the Agilent SureSelect Human All Exon 38Mb (Exome38Mb) and Agilent SureSelect Human All Exon 51Mb (Exome51Mb) platforms. Genotypes were determined using the GATK Best Practices workflow [33].

In this dataset, we curated 780 whole-exome sequencing samples, comprising 270 control samples and 510 samples diagnosed with type 2 diabetes. The complete original sequences of these 780 samples were aligned to the GRCh37 reference genome to detect variants. Therefore, we utilized variant information in the Variant Call Format (VCF) for subsequent analyses.

3.2. Data Preprocessing

The obesity data undergoes the complete processing workflow 1, whereas the type 2 diabetes data is subjected only to the final two stages: target data preprocessing and prediction. The reason is that the original type 2 diabetes data were genotyped using GATK, following the same procedure as our first two steps. Although exact similarity cannot be guaranteed, but we still reused the genotype calling results from the previous study, which were also carefully processed.

3.2.1. The raw data preprocessing

The raw obesity data preprocessing involves several steps. First, the raw data undergoes base trimming using Trimomatic version 0.36 to remove sequences with a sequencing quality below 99%. Subsequently, the data is aligned to the GRCh38 genome using the BWA-MEM version 0.7.17. A comprehensive base quality score recalibration (BQSR) is performed using extensive databases following alignment.

3.2.2. Variant calling

Obesity variants are then called from the BQSR data using the Haplotype Caller of GATK version 4.2.0.0. The individual variant results are combined to create a cohort variant file. This file is further subjected to variant quality score recalibration using the VQSR tool, and variants that fail to meet the quality criteria are removed. Specifically, only SNPs are retained for downstream analysis. The resulting dataset consists of 605 SNPs.

3.2.3. The target data preprocessing

For obesity disease, the target data post-variant calling (in VCF format) consists of 625 SNPs and 139 samples, comprising 88 males and 51 females. It was observed that 291 SNPs were absent in the reference panel, while 197 SNPs exhibited discordant allele frequencies (AF) compared to the reference panel. Following filtering criteria, 137 SNPs were retained (see Figure 3 a)), despite having a missing genotype rate of 27%. Finally, the imputation process is executed to furnish the full genotype dataset for model training.

For type 2 diabetes, the initial data in VCF format has undergone variant calling. As these data files do not conform to the GVCF format, they have been merged utilizing GATK 3.8. Subsequently, the data were aligned to the GRCh37 reference genome, necessitating a liftover to the GRCh38 genome build using

GATK 4.3. After merging, the dataset comprises 564,052 SNPs, of which 291,347 SNPs do not align with the reference genome, and 261,197 SNPs exhibit discordant allele frequencies compared to the variants in the panel genome. Only 11,508 SNPs are retained for subsequent analysis (see Figure 3 b). The dataset, which exhibited a missing genotype rate of 12%, underwent an imputation process to acquire a complete genotype dataset.

3.3. Prediction

For the prediction stage, we performed experiments on a high-performance computer with 72 CPU cores and 128 GB of memory. The preprocessed data were divided into 80% for training per 20% for testing. For all methods, the selected feature sets were evaluated using 5-fold cross-validation. Models were trained in turn on each 80% of the training data, and performance measurements were obtained by evaluating the remaining 20% of the training data. The final evaluation results were derived by averaging the Area Under the Curve (AUC), helping the models avoid high bias. Additionally, we employed grid search to determine the best hyperparameters. For imbalanced data such as that seen in type 2 diabetes, employing a 5-fold cross-validation technique and using sensitivity, Matthews correlations coefficient (MCC), and AUC as evaluation metrics are suitable measures. In the case of Neural Networks, the activation function for the nodes in the hidden layers was sigmoid, while the output layer utilized the softmax function. The prediction results for each case or control were determined by comparing the probabilities generated by the softmax function. The data was trained with 200 epochs. For the SFE method, we set $UR_{max} = 0.3$, $UR_{min} = 0.001$, and $Max\_FEs = 777$ for all datasets. Regarding *nvar* (the number of features), $nvar = 137$ for obesity and $nvar = 11,508$ for type 2 diabetes.

The feature selection models are compared

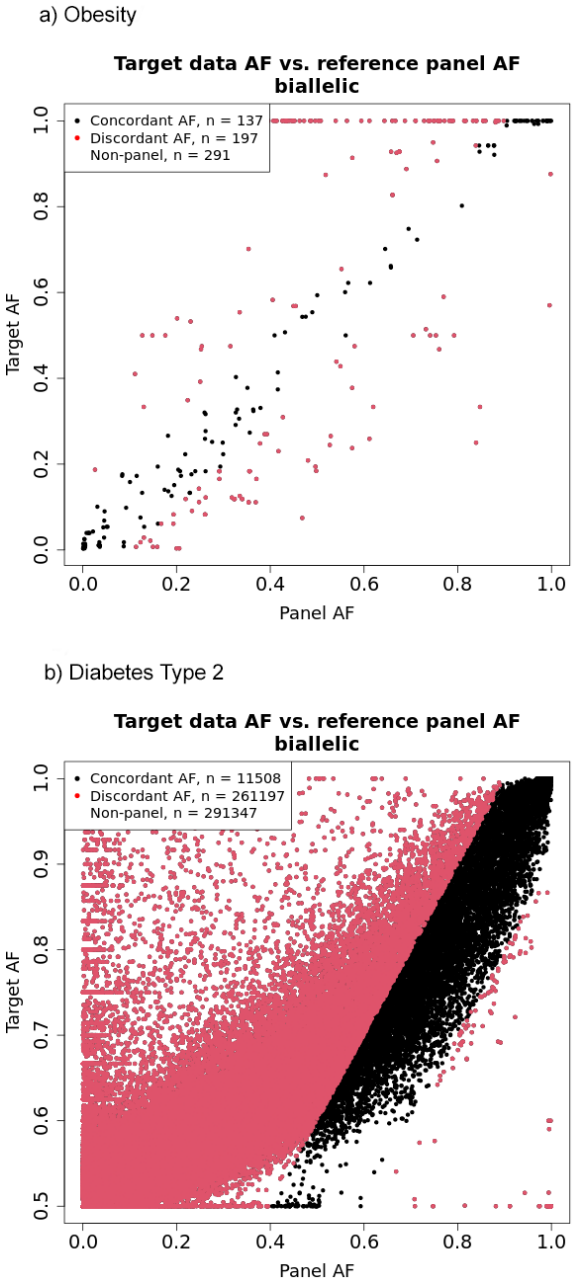

Figure 3. **The preprocessing steps for both obesity disease and type 2 diabetes datasets.** The dataset underwent filtration to remove variants not present in the gene panel and those with allele frequencies discordant with the reference allele frequencies. a) For obesity disease, 137 SNPs were retained b) For type 2 diabetes, 11,508 SNPs were retained for subsequent analysis.

by assessing their performance on the test set. Subsequently, the models are retrained using three datasets: the original data, the data post-dimensionality reduction using common methods (specifically, principal component analysis), and the dataset incorporating selected features. This iterative process aims to reassess the model performance with datasets containing selected features, thereby reinforcing the hypothesis that models trained on this dataset attain optimal performance.

3.4. Results

3.4.1. Obesity

In our obesity disease analysis workflow, the predictive models exhibit competitive performance. Specifically, NN RFE demonstrates the highest sensitivity of 100% (95% CI: 0.57-1.0), surpassing the sensitivity values reported by H. Y. Wang et al. for SVM (80%) and KNN (76%). Moreover, our analysis across various models and feature selection methods reveals consistently high specificity values, with the majority achieving specificity above 70%. In contrast, H. Y. Wang et al.'s findings indicate lower specificity values for SVM (63%) and KNN (50%), implying reduced discriminative power. Additionally, our examination highlights competitive accuracy values, with KNN SFE achieving the highest accuracy of 86% (95% CI: 0.69-0.94) compared to the lower accuracy values reported for SVM (71%) and KNN (63%). The MCC values for DT methods (RFE and SFE) 60% (0.53-0.67) and 66% (0.59-0.72) respectively, are significantly higher than the MCC of 0.16 (-0.06 to 0.23) reported by Wang et al. Similarly, the Support Vector Machine (SVM) method shows a notable improvement in MCC, with a value of 64% (0.57-0.71) using feature selection (SFE) compared to the MCC of 41% (0.22-0.50) from Wang et al. KNN SFE demonstrates a much higher MCC of 71% (0.64-0.77) compared to the 17% (0.10-0.25) reported by Wang et al. The complete set of results is presented in Table 1.

Besides, AUC quantifies the model's ability to differentiate between classes by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold values. In Figure 4, distinct AUC values are shown for the LR method, as well as for the RFE and SFE methods. Notably, the KNN SFE method emerges with the highest AUC of 0.84, indicating its good discriminatory ability compared to other techniques. Therefore, KNN SFE would be the preferred feature selection method for predicting obesity disease in this scenario.

The inquiry arises as to whether training on the complete dataset or the principal components of the original data would enhance prediction performance with feature selection data. To assess feature selection efficacy using genotype-phenotype data, we compare the AUC and execution time of methods across three datasets: the original dataset (All), the principal components (PCA) derived from the original data, and the dataset comprising only selected features (FS) based on KNN SFE (8 SNPs). Here, we perform PCA with the number of principal components equal to the number of selected features from KNN SFE. The outcomes reveal uniformity across all model training methods; models achieve the highest AUC with the dataset containing the significant 8 SNPs, notably reaching 0.84 with KNN. Furthermore, employing KNN results in the shortest model training time with the feature-selected data (see Figure 5). Consequently, the model's performance is optimal when utilizing data comprising selected features.

Finally, we provided information about 8 significant SNPs obtained from KNN SFE (see Table 2). We can see the gene located near the SNP or the gene affected by the SNP in the GENE column. Besides, most SNPs are annotated as intergenic variants (located

Table 1. Comparing Feature Selection Methods for Obesity Disease.

Model	**Sensitivity** (95% CI)	**Specificity** (95% CI)	**Accuracy** (95% CI)	**MCC** (95% CI)	**AUC**
LR	0.46 (0.23-0.71)	0.60 (0.36-0.80)	0.54 (0.36-0.70)	0.06 (0.04-0.11)	0.53
DT RFE	0.69 (0.44-0.86)	**0.92** (0.65-0.99)	0.79 (0.60-0.90)	0.60 (0.53-0.67)	0.80
SVM RFE	0.62 (0.39-0.82)	0.83 (0.55-0.95)	0.71 (0.53-0.85)	0.46 (0.39-0.53)	0.73
NN RFE	**1.00** (0.57-1.00)	0.70 (0.49-0.84)	0.75 (0.57-0.87)	0.54 (0.46-0.62)	0.71
KNN SFE	0.90 (0.60-0.98)	0.83 (0.61-0.94)	**0.86** (0.69-0.94)	**0.71** (0.64-0.77)	**0.84**
DT SFE	0.73 (0.48-0.89)	**0.92** (0.67-0.99)	0.82 (0.64-0.92)	0.66 (0.59-0.72)	0.83
SVM SFE	0.89 (0.57-0.98)	0.79 (0.57-0.91)	0.82 (0.64-0.92)	0.64 (0.57-0.71)	0.80

The table compares various feature selection methods for predicting obesity disease. Each method's performance metrics, including sensitivity, specificity, accuracy, Matthews correlations coefficient (MCC), and area under the receiver operating characteristic curve (AUC), are reported along with their 95% confidence intervals (CI). The highest performance was observed with NN RFE, which achieved a sensitivity of 100% (0.57-1.0); DT RFE and SFE showed a specificity of 92%; and KNN SFE, which demonstrated an accuracy of 86% (0.69-0.94) and an MCC of 71% (0.64-0.77). KNN SFE also attained the highest AUC of 0.84, indicating strong discriminative power.

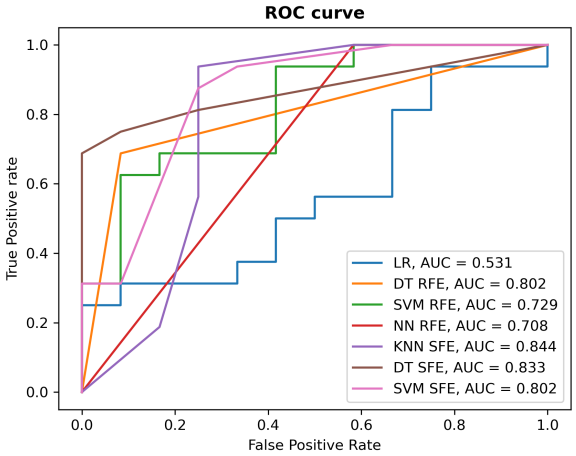

Figure 4. **Comparing AUC for feature selection methods in predicting obesity disease.** KNN SFE method achieved the highest AUC of 0.84. This indicates its best discriminatory ability compared to other methods, making it the preferred choice for feature selection of the obesity disease.

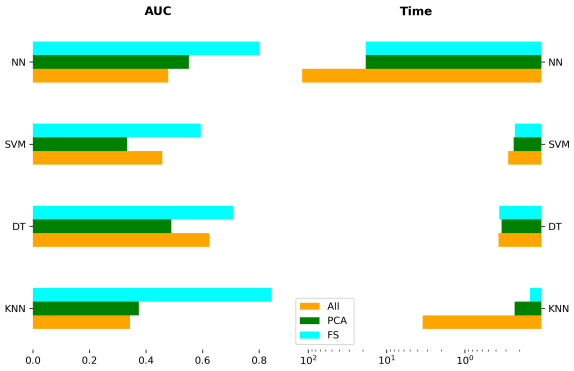

Figure 5. **Comparing feature selection methods' impact on prediction performance for obesity disease**. Three datasets—original (All), principal components (PCA), and selected features (FS) from KNN SFE (8 SNPs)—are evaluated for AUC and execution time across NN, SVM, DT, and KNN methods. Notably, models perform best with the dataset containing the top 8 SNPs, achieving an AUC of 0.84 and the shortest training time with KNN.

in regions between genes) or intronic variants (within introns of genes). The SNPs are associated with various obesity-related traits, including body fat distribution, body mass index (BMI), and hyperlipidemia. Key genes include **NISCH**, **FAM13A**, **MTCH2**, **KCNA5-LINC02443**, **PRKCH**, **RBFOX1**, and **FTO**, which are involved in pathways affecting obesity traits. The findings provide insights into the genetic factors influencing obesity and potential targets for precision medicine.

3.4.2. Type 2 diabetes

After the first two steps of the workflow, the obesity and type 2 diabetes datasets share the same format and characteristics. The features are all single nucleotide polymorphisms, the samples are all human, and the values are all genotypes. Moreover, the output labels remain as case and control. Additionally, based on the findings obtained concerning obesity disease, three feature selection methods were employed, namely KNN, DT, and SVM SFE, to identify significant SNPs because these algorithms perform better than the other group of methods. Therefore, we selected the three best methods applied to obesity to apply to type 2 diabetes.

Each method underwent seven iterations of algorithm execution, with Max_FEs = 777 set for model evaluation in each iteration. The results indicate that the AUC value increases proportionally with the number of function evaluations. Specially, the acceleration of KNN SFE surpasses that of other methods notably from iteration 200 onwards (see Figure 6).

The table 3 compares three feature selection methods for predicting Type 2 Diabetes. The KNN SFE method has an MCC of 58% (0.57-0.60), indicating a good balance between true and false predictions. This method shows the highest sensitivity (74%) and specificity (85%), leading to the highest accuracy of 82%. In contrast, the Decision Tree (DT) method with SFE has a lower MCC of 29% (0.28-0.30), suggesting

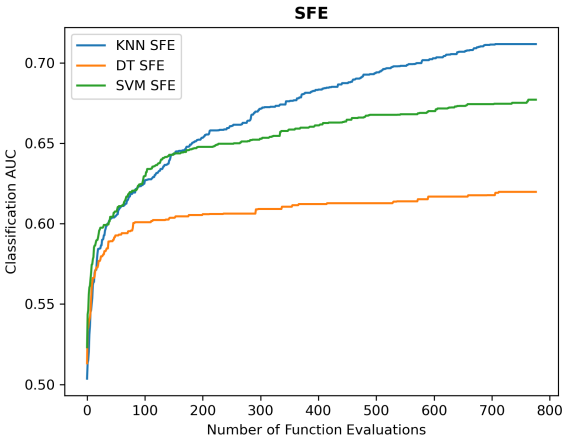

Figure 6. **Comparative Analysis of Feature Selection Methods for Identifying Significant SNPs in Type 2 Diabetes Prediction.** KNN SFE, DT SFE, and SVM SFE were employed. Each method underwent seven iterations with Max_FEs = 777. The results demonstrate that the AUC value increases with the number of function evaluations, with KNN SFE exhibiting notably faster acceleration from iteration 200 onwards.

a less balanced performance than KNN SFE. DT SFE has sensitivity (52%) and specificity (77%) are also lower, resulting in an accuracy of 69%. SVM SFE demonstrates a moderate MCC of 52% (0.51-0.54). This model has sensitivity (67%) and specificity (85%) well, achieving an accuracy of 79%. The AUC values summarize the overall discriminatory power of each method, with KNN SFE having the highest AUC (0.79), followed by SVM SFE (0.76) and DT SFE (0.65). Therefore, we choose the features based on the KNN SFE model, resulting in the selection of 61 significant SNPs. The list of SNPs is provided in Supplementary 1. To identify the total number of SNPs related to type 2 diabetes from previous research, we used the Type 2 Diabetes Knowledge Portal, a genetic resource dedicated to type 2 diabetes and related traits [41], to query information about these SNPs. As a result, we found that 50 out of the 61 SNPs were reported in previous research. The table 4

Table 2. Annotation of 8 significant single nucleotide polymorphisms for obesity disease.

SNP	CHROM	POS	REF	ALT	GENE	EFFECT	OBESITY TRAIT	REFERENCE
rs6445486	Chr 3	52472475	A	G	NISCH	Intron	Fat Distribution	C. Sun, 2021 [34]
rs34749134	Chr 4	88958337	C	T	FAM13A	Intron	Fat Distribution	M. Fathzade, 2020 [35]
rs3817334	Chr 11	47629441	C	T	MTCH2	Intron	BMI	J. A. Fischer, 2023 [36]
rs657538	Chr 12	5135527	T	C	KCNA5-LINC02443	Intergenic	Hyperlipidemia	W. Bi et al., 2019 [37]
rs79090609	Chr 14	61441381	G	A	PRKCH	Intron	BMI	Q.-Y. Song, 2017 [38]
rs1957894	Chr 14	61441393	T	G	PRKCH	Intron	BMI	Q.-Y. Song, 2017 [38]
rs147340331	Chr 16	6114440	T	C	RBFOX1	Intron	BMI	K. Y. He, 2017 [39]
rs1421085	Chr 16	53767042	T	C	FTO	Intron	BMI	K. A. Fawcett, 2010 [40]

Each row represents a SNP and includes the following attributes: ID: The unique identifier for the SNP; CHROM: The chromosome where the SNP is located; POS: The position of the SNP on the chromosome; REF: The reference allele for the SNP; ALT: The alternate allele for the SNP; AF: The allele frequency of the SNP; GENE: The gene associated with the SNP; EFFECT: The type of genomic region or effect of the SNP; OBESITY TRAIT: a genetically influenced predisposition towards obesity.

in Appendix A provides annotations for 50 SNPs associated with type 2 diabetes (T2D), detailing their chromosome locations, positions, genes, and related traits. SNPs span various chromosomes, with notable genes including **ADGRL4**, **TTN**, **NGEF**, **TECRL**, **ZFYVE16**, **OR12D2**, **CPNE5**, **TMEM184A**, **ATP6V1H**, **LAMC3**, **PRLHR**, **MYBPC3**, **ALDH3B2**, **GYS2**, and **OSM**. Associated traits cover general T2D susceptibility and specific complications such as cardiovascular diseases (e.g., coronary artery disease, peripheral vascular disease), renal conditions (e.g., end-stage renal disease, chronic kidney disease), metabolic conditions (e.g., triglyceride levels, NAFLD), youth-onset and obesity-related T2D, and other diabetes-related traits.

4. Conclusion

Our study presents a systematic workflow for human disease risk prediction, based on next-generation sequencing technologies, computational methodologies, and machine learning techniques. By integrating the Genome Analysis Toolkit (GATK), BEAGLE's data imputation procedure, and feature selection methodologies, we have developed a workflow to identify genetic variants associated with complex diseases such as obesity and T2D. The workflow encompasses several stages, including raw data preprocessing, variant calling, target data preprocessing, and prediction. This comprehensive approach not only enhances the accuracy of disease risk prediction models but also improves their interpretability.

Furthermore, our study highlights the significance of feature selection in improving the accuracy and interpretability of predictive models. By employing a variety of feature selection methodologies, we have identified subsets of genetic variants associated with obesity and type 2 diabetes. For obesity, we not only compared models but also considered data encoding. Wang et al. assigned "1" (the SNP is used) and "0" (the SNP is not used) to all attributes, whereas we encoded using the values 0, 1, and 2, representing the total differences between alleles one and two compared to the reference allele. Our encoding method generated post-processing data that resulted in better model performance. Notably, the KNN SFE method emerged as particularly effective, demonstrating high sensitivity, specificity, accuracy, and AUC for both diseases.

Table 3. Comparing Feature Selection Methods for Type 2 Diabetes.

Model	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	MCC (95% CI)	AUC
KNN SFE	**0.74** (0.60-0.85)	**0.85** (0.77-0.91)	**0.82** (0.75-0.87)	**0.58** (0.57-0.60)	**0.79**
DT SFE	0.52 (0.39-0.65)	0.77 (0.68-0.84)	0.69 (0.61-0.75)	0.29 (0.28-0.30)	0.65
SVM SFE	0.67 (0.54-0.78)	**0.85** (0.76-0.90)	0.79 (0.72-0.85)	0.52 (0.51-0.54)	0.76

The table compares various feature selection methods for predicting type 2 diabetes. Each method's performance metrics, including sensitivity, specificity, accuracy, Matthews correlations coefficient (MCC), and area under the receiver operating characteristic curve (AUC), are reported along with their 95% confidence intervals (CI).

Our study provides insights into the genetic landscape of type 2 diabetes and obesity by identifying significant SNPs. We identified 61 SNPs associated with type 2 diabetes, with 50 of these SNPs previously reported in studies. Additionally, we identified 8 SNPs significantly linked to obesity. Several SNPs are associated with obesity and T2D, illustrating their interconnected nature. These genetic overlaps highlight the role of obesity as a significant risk factor for T2D, underscoring the need for integrated prevention and treatment strategies addressing both conditions.

Our findings contribute to understanding how genetic variants influence disease risk in obesity and type 2 diabetes. This knowledge has the potential to lead to improved treatments tailored to an individual's genetic profile. The data regarding obesity and the source code for this study are available at https://github.com/nhanta/HDRP.

Acknowledgement

We would like to thank the REMOSAT laboratory at the University of Science and Technology of Hanoi (USTH) for providing the computing resources utilized in this study.

References

[1] L. D. Maxim, R. Niebo, M. J. Utell, Screening Tests: A Review with Examples, Vol. 26, No. 13, pp. 811–828. doi:10.3109/08958378.2014.955932.

[2] M. Speechley, A. Kunnilathu, E. Aluckal, M. S. Balakrishna, B. Mathew, E. K. George, Screening in Public Health and Clinical Care: Similarities and Differences in Definitions, Types, and Aims – A Systematic Review, Vol. 11, No. 3, pp. LE01–LE04. doi:10.7860/JCDR/2017/24811.9419.

[3] Y. Xue, A. Ankala, W. R. Wilcox, M. R. Hegde, Solving the Molecular Diagnostic Testing Conundrum for Mendelian Disorders in the Era of Next-Generation Sequencing: Single-Gene, Gene Panel, or Exome/Genome Sequencing, Vol. 17, No. 6, pp. 444–451, publisher: Nature Publishing Group. doi:10.1038/gim.2014.122.

[4] Z. Liu, L. Zhu, R. Roberts, W. Tong, Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We?, Vol. 35, No. 11, pp. 852–867, publisher: Elsevier. doi:10.1016/j.tig.2019.08.006.

[5] S. C. Rubin, M. A. Blackwood, C. Bandera, K. Behbakht, I. Benjamin, T. R. Rebbeck, J. Boyd, BRCA1, BRCA2, and Hereditary Nonpolyposis Colorectal Cancer Gene Mutations in an Unselected Ovarian Cancer Population: Relationship to Family History and Implications for Genetic Testing, Vol. 178, No. 4, pp. 670–677. doi:10.1016/S0002-9378(98)70476-4.

[6] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, E. M. Derks, A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis, Vol. 27, No. 2, pp. e1608. doi:10.1002/mpr.1608.

[7] S. W. Choi, T. S.-H. Mak, P. F. O'Reilly, Tutorial: A Guide to Performing Polygenic Risk Score Analyses, Vol. 15, No. 9, pp. 2759–2772, number: 9 Publisher: Nature Publishing Group. doi:10.1038/s41596-020-0353-1.

[8] B. E. Slatko, A. F. Gardner, F. M. Ausubel, Overview of Next-Generation Sequencing Technologies, Vol. 122, No. 1, pp. e59. doi:10.1002/cpmb.59.

[9] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data, Vol. 20, No. 9, pp. 1297–1303. doi:10.1101/gr.107524.110.

[10] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks, Vol. 36, No. 10, pp. 983–987, publisher: Nature Publishing Group. doi:10.1038/nbt.4235.

[11] T. Yun, H. Li, P.-C. Chang, M. F. Lin, A. Carroll, C. Y. McLean, Accurate, Scalable Cohort Variant Calls Using DeepVariant and GLnexus, Vol. 36, No. 24, pp. 5582–5589. doi:10.1093/bioinformatics/btaa1081.

[12] A. Goyal, H. J. Kwon, K. Lee, R. Garg, S. Y. Yun, Y. H. Kim, S. Lee, M. S. Lee, Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGENTM Bio-IT Processor for Precision Medicine, Vol. 7, No. 1, pp. 9–19, number: 1 Publisher: Scientific Research Publishing. doi:10.4236/ojgen.2017.71002.

[13] B. L. Browning, Y. Zhou, S. R. Browning, A One-Penny Imputed Genome from Next-Generation Reference Panels, Vol. 103, No. 3, pp. 338–348. doi:10.1016/j.ajhg.2018.07.015.

[14] E. R. Mardis, The Impact of Next-Generation Sequencing Technology on Genetics, Vol. 24, No. 3, pp. 133–141. doi:10.1016/j.tig.2007.12.007.

[15] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Vol. 3, No. 7, pp. 1157–1182, place: US Publisher: MIT Press. doi:10.1162/153244303322753616.

[16] H.-Y. Wang, S.-C. Chang, W.-Y. Lin, C.-H. Chen, S.-H. Chiang, K.-Y. Huang, B.-Y. Chu, J.-J. Lu, T.-Y. Lee, Machine Learning-Based Method for Obesity Risk Evaluation Using Single-Nucleotide Polymorphisms Derived from Next-Generation Sequencing, Vol. 25, No. 12, pp. 1347–1360. doi:10.1089/cmb.2018.0002.

[17] K. R. Franke, E. L. Crowgey, Accelerating Next Generation Sequencing Data Analysis: An Evaluation of Optimized Best Practices for Genome Analysis Toolkit Algorithms, Vol. 18, No. 1, pp. e10. doi:10.5808/GI.2020.18.1.e10.

[18] B. L. Browning, X. Tian, Y. Zhou, S. R. Browning, Fast Two-Stage Phasing of Large-Scale Sequence Data, Vol. 108, No. 10, pp. 1880–1890. doi:10.1016/j.ajhg.2021.08.005.

[19] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The Variant Call Format and VCFtools, Vol. 27, No. 15, pp. 2156–2158. doi:10.1093/bioinformatics/btr330.

[20] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, J. M. O'Sullivan, A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction, Vol. 2, , publisher: Frontiers.

[21] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang, Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data, Vol. 143, pp. 106839. doi:10.1016/j.csda.2019.106839.

[22] J. Maldonado, M. C. Riff, B. Neveu, A Review of Recent Approaches on Wrapper Feature Selection for Intrusion Detection, Vol. 198, pp. 116822. doi:10.1016/j.eswa.2022.116822.

[23] H. Hamla, K. Ghanem, Comparative Study of Embedded Feature Selection Methods on Microarray Data, in: I. Maglogiannis, J. Macintyre, L. Iliadis (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, pp. 69–77. doi:10.1007/978-3-030-79150-6_6.

[24] T. Dokeroglu, A. Deniz, H. E. Kiziloz, A Comprehensive Survey on Recent Metaheuristics for Feature Selection, Vol. 494, pp. 269–296. doi:10.1016/j.neucom.2022.04.083.

[25] F. Privé, H. Aschard, M. G. B. Blum, Efficient Implementation of Penalized Regression for Genetic Risk Prediction, Vol. 212, No. 1, pp. 65–74. doi:10.1534/genetics.119.302019.

[26] W. Lian, G. Nie, B. Jia, D. Shi, Q. Fan, Y. Liang, An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning, Vol. 2020, pp. e2835023, publisher: Hindawi. doi:10.1155/2020/2835023.

[27] M.-L. Huang, Y.-H. Hung, W. M. Lee, R. K. Li, B.-R. Jiang, SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier, Vol. 2014, pp. 795624. doi:10.1155/2014/795624.

[28] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, pp. 3319–3328.

[29] B. Ahadzadeh, M. Abdar, F. Safara, A. Khosravi, M. B. Menhaj, P. N. Suganthan, SFE: A Simple, Fast, and Efficient Feature Selection Algorithm for High-Dimensional Data, Vol. 27, No. 6, pp. 1896–1911, conference Name: IEEE Transactions on Evolutionary Computation. doi:10.1109/TEVC.2023.3238420.

[30] W. Zhu, N. Zeng, N. Wang, Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations.

[31] E. B. Wilson, Probable Inference, the Law of Succession, and Statistical InferencePublisher: Taylor

& Francis Group.

[32] D. Altman, D. Machin, T. Bryant, M. Gardner, Statistics with Confidence: Confidence Intervals and Statistical Guidelines, John Wiley & Sons, google-Books-ID: HmnIBAAAQBAJ.

[33] K. A. Fakhro, M. R. Staudt, M. D. Ramstetter, A. Robay, J. A. Malek, R. Badii, A. A.-N. Al-Marri, C. A. Khalil, A. Al-Shakaki, O. Chidiac, D. Stadler, M. Zirie, A. Jayyousi, J. Salit, J. G. Mezey, R. G. Crystal, J. L. Rodriguez-Flores, The Qatar Genome: A Population-Specific Tool for Precision Medicine in the Middle East, Vol. 3, No. 1, pp. 1–7, publisher: Nature Publishing Group. doi:10.1038/hgv.2016.16.

[34] C. Sun, P. Kovacs, E. Guiu-Jurado, Genetics of Body Fat Distribution: Comparative Analyses in Populations with European, Asian and African Ancestries, Vol. 12, No. 6, pp. 841, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/genes12060841.

[35] M. Fathzadeh, J. Li, A. Rao, N. Cook, I. Chennamsetty, M. Seldin, X. Zhou, P. Sangwung, M. J. Gloudemans, M. Keller, A. Attie, J. Yang, M. Wabitsch, I. Carcamo-Orive, Y. Tada, A. J. Lusis, M. K. Shin, C. M. Molony, T. McLaughlin, G. Reaven, S. B. Montgomery, D. Reilly, T. Quertermous, E. Ingelsson, J. W. Knowles, FAM13A Affects Body Fat Distribution and Adipocyte Function, Vol. 11, No. 1, pp. 1465, publisher: Nature Publishing Group.

[36] J. A. Fischer, T. O. Monroe, L. L. Pesce, K. T. Sawicki, M. Quattrocelli, R. Bauer, S. D. Kearns, M. J. Wolf, M. J. Puckelwartz, E. M. McNally, Opposing Effects of Genetic Variation in MTCH2 for Obesity Versus Heart Failure, Vol. 32, No. 1, pp. 15–29. doi:10.1093/hmg/ddac176.

[37] W. Bi, Z. Zhao, R. Dey, L. G. Fritsche, B. Mukherjee, S. Lee, A Fast and Accurate Method for Genome-wide Scale Phenome-wide G × E Analysis and Its Application to UK Biobank, Vol. 105, No. 6, pp. 1182–1192. doi:10.1016/j.ajhg.2019.10.008.

[38] Q.-Y. Song, J.-Y. Song, Y. Wang, S. Wang, Y.-D. Yang, X.-R. Meng, J. Ma, H.-J. Wang, Y. Wang, Association Study of Three Gene Polymorphisms Recently Identified by a Genome-Wide Association Study with Obesity-Related Phenotypes in Chinese Children, Vol. 10, No. 3, pp. 179–190. doi:10.1159/000471487.

[39] K. Y. He, H. Wang, B. E. Cade, P. Nandakumar, A. Giri, E. B. Ware, J. Haessler, J. Liang, J. A. Smith, N. Franceschini, T. H. Le, C. Kooperberg, T. L. Edwards, S. L. R. Kardia, X. Lin, A. Chakravarti, S. Redline, X. Zhu, Rare Variants in FOX-1 Homolog A (RBFOX1) Are Associated With Lower Blood Pressure, Vol. 13, No. 3, pp. e1006678. doi:10.1371/journal.pgen.1006678.

[40] K. A. Fawcett, I. Barroso, The Genetics of Obesity: FTO Leads the Way, Vol. 26, No. 6, pp. 266–274.

[41] M. C. Costanzo, M. v. Grotthuss, J. Massung, D. Jang, L. Caulkins, R. Koesterer, C. Gilbert, R. P. Welch, P. Kudtarkar, Q. Hoang, A. P. Boughton, P. Singh, Y. Sun, M. Duby, A. Moriondo, T. Nguyen, P. Smadbeck, B. R. Alexander, M. Brandes, M. Carmichael, P. Dornbos, T. Green, K. C. Huellas-Bruskiewicz, Y. Ji, A. Kluge, A. C. McMahon, J. M. Mercader, O. Ruebenacker, S. Sengupta, D. Spalding, D. Taliun, G. Abecasis, B. Akolkar, B. R. Alexander, N. D. Allred, D. Altshuler, J. E. Below, R. Bergman, J. W. J. Beulens, J. Blangero, M. Boehnke, K. Bokvist, E. Bottinger, A. P. Boughton, D. Bowden, M. J. Brosnan, C. Brown, K. Bruskiewicz, N. P. Burtt, M. Carmichael, L. Caulkins, I. Cebola, J. Chambers, Y.-D. I. Chen, A. Cherkas, A. Y. Chu, C. Clark, M. Claussnitzer, M. C. Costanzo, N. J. Cox, M. d. Hoed, D. Dong, M. Duby, R. Duggirala, J. Dupuis, P. J. M. Elders, J. M. Engreitz, E. Fauman, J. Ferrer, J. Flannick, P. Flicek, M. Flickinger, J. C. Florez, C. S. Fox, T. M. Frayling, K. A. Frazer, K. J. Gaulton, C. Gilbert, A. L. Gloyn, T. Green, C. L. Hanis, R. Hanson, A. T. Hattersley, Q. Hoang, H. K. Im, S. Iqbal, S. B. R. Jacobs, D.-K. Jang, T. Jordan, T. Kamphaus, F. Karpe, T. M. Keane, S. K. Kim, A. Kluge, R. Koesterer, P. Kudtarkar, K. Lage, L. A. Lange, M. Lazar, D. Lehman, C.-T. Liu, R. J. F. Loos, R. C.-w. Ma, P. MacDonald, J. Massung, M. T. Maurano, M. I. McCarthy, G. McVean, J. B. Meigs, J. M. Mercader, M. R. Miller, B. Mitchell, K. L. Mohlke, S. Morabito, C. Morgan, S. Mullican, S. Narendra, M. C. Y. Ng, L. Nguyen, C. N. A. Palmer, S. C. J. Parker, A. Parrado, A. Parsa, A. C. Pawlyk, E. R. Pearson, A. Plump, M. Province, T. Quertermous, S. Redline, D. F. Reilly, B. Ren, S. S. Rich, J. B. Richards, J. I. Rotter, O. Ruebenacker, H. Ruetten, R. M. Salem, M. Sander, M. Sanders, D. Sanghera, L. J. Scott, S. Sengupta, D. Siedzik, X. Sim, P. Singh, R. Sladek, K. Small, P. Smith, P. Stein, D. Spalding, H. M. Stringham, Y. Sun, K. Susztak, L. M. art, D. Taliun, K. Taylor, M. K. Thomas, J. A. Todd, M. S. Udler, B. Voight, M. v. Grotthuss, A. Wan, R. P. Welch, D. Wholley, K. Yuksel, N. A. Zaghloul, P. Smith, M. K. Thomas, B. Akolkar, M. J. Brosnan, A. Cherkas, A. Y. Chu, E. B. Fauman, C. S. Fox, T. N. Kamphaus, M. R. Miller, L. Nguyen, A. Parsa, D. F. Reilly, H. Ruetten, D. Wholley, N. A. Zaghloul, G. R. Abecasis, D. Altshuler, T. M. Keane, M. I. McCarthy, K. J. Gaulton, J. C. Florez, M. Boehnke, N. P. Burtt, J. Flannick, The Type 2 Diabetes Knowledge Portal: An Open Access Genetic Resource Dedicated to Type 2 Diabetes and Related Traits, Vol. 35, No. 4, pp. 695–710.e6, publisher: Elsevier. doi:10.1016/j.cmet.2023.03.001.

Appendix A Variant Annotation for T2D

Table 4. Annotation of 50 single nucleotide polymorphisms related to type 2 diabetes in previous researches.

SNP	CHROM	POS	GENE	TYPE 2 DIABETES TRAIT
rs1061728	Chr1	78889911	ADGRL4	Type 2 diabetes
rs4512713	Chr1	204247340	PLEKHA6	Claudication in type 2 diabetes
rs2251987	Chr2	178701419	TTN	Type 2 diabetes
rs895432	Chr2	232892987	NGEF	Type 2 diabetes
rs13018934	Chr2	233799983	MROH2A	End-stage renal disease vs. no ESRD in type 2 diabetes (T2D)
rs1483711	Chr4	64280237	TECRL	Coronary artery disease in type 2 diabetes (CAD in T2D)
rs2544600	Chr5	80437260	ZFYVE16	Triglyceride levels in individuals without type 2 diabetes
rs3128853	Chr6	29397010	OR12D2	Youth-onset type 2 diabetes (T2D)
rs763046	Chr6	36745486	CPNE5	Chronic kidney disease (CKD) in type 2 diabetes (T2D)
rs3814481	Chr7	1548755	TMEM184A	Triglyceride levels in individuals with type 2 diabetes
rs1584614	Chr7	32490324	LSM5	NAFLD in type 2 diabetes (T2D)
rs2392572	Chr7	38429095	AMPH	Chronic kidney disease in type 2 diabetes
rs6463449	Chr7	47813900	PKD1L1, C7orf69	Type 2 diabetes (T2D)
rs11784716	Chr8	12015214	DEFB134-DEFB130A	Type 2 diabetes (T2D) adjusted BMI
rs6468093	Chr8	30069784	SARAF	Microalbuminuria in type 2 diabetes (T2D)
rs6991513	Chr8	53832978	ATP6V1H	eGFRcreat (serum creatinine) in type 2 diabetes
rs10780871	Chr9	70168692	MAMDC2,MAMDC2-AS1	Cardiovascular disease in type 2 diabetes (T2D)
rs2045732	Chr9	97432124	TDRD7	Chronic kidney disease in type 2 diabetes
rs1949755	Chr9	104504493	OR13F1	Type 2 diabetes (T2D)
rs3739512	Chr9	131009433	LAMC3	Type 2 diabetes (T2D)
rs3818581	Chr9	131590011	RAPGEF1	Coronary artery disease in type 2 diabetes (CAD in T2D)
rs2797491	Chr10	5746645	TASOR2	Peripheral vascular disease in type 2 diabetes
rs1613448	Chr10	118594398	PRLHR	Type 2 diabetes (T2D)
rs892336	Chr11	5581449	OR52B6	Macroalbuminuria in type 2 diabetes (T2D)
rs2956109	Chr11	34916718	PDHX,APIP	Type 2 diabetes (T2D)
rs10838693	Chr11	47329002	MYBPC3,MADD	Triglyceride levels in individuals with type 2 diabetes
rs4646823	Chr11	67666945	ALDH3B2	Cardiovascular disease in type 2 diabetes (T2D)
rs2306180	Chr12	21560468	GYS2	Mild obesity-related type 2 diabetes
rs7132431	Chr12	55320988	OR6C1	Peripheral vascular disease in type 2 diabetes
rs1284467	Chr12	57511904	MIR6758,DDIT3,MARS1	Chronic kidney disease (CKD) and diabetic nephropathy in T2D
rs6488867	Chr12	123309475	SBNO1	Peripheral vascular disease in type 2 diabetes
rs3812896	Chr13	41328803	NAA16	Type 2 diabetes (T2D)
rs1042631	Chr15	88859008	ACAN	End-stage renal disease vs. no ESRD in type 2 diabetes (T2D)
rs8064024	Chr16	4805278	GLYR1,ROGDI	Severe insulin-deficient type 2 diabetes
rs1376041	Chr16	57655971	ADGRG1	Type 2 diabetes (T2D) adjusted BMI
rs2236375	Chr17	2040594	DPH1,OVCA2	Stroke in type 2 diabetes
rs61075345	Chr17	18790964	TVP23B	Microalbuminuria in type 2 diabetes (T2D)
rs6587220	Chr17	19328765	EPN2	Youth-onset type 2 diabetes (T2D)
rs8068049	Chr17	31856838	COPRS,UTP6	Severe insulin-deficient type 2 diabetes
rs4602	Chr17	41528069	KRT19	Type 2 diabetes (with no history of pregnancy)
rs1156287	Chr17	54999438	STXBP4	Coronary heart disease/stroke/peripheral vascular disease in T2D
rs2382647	Chr17	74370556	GPR142	Microalbuminuria in type 2 diabetes (T2D)
rs12970083	Chr18	24130357	TTC39C	Microalbuminuria in type 2 diabetes (T2D)
rs8085482	Chr18	63659368	SERPINB3	Youth-onset type 2 diabetes (T2D)
rs803665	Chr20	11810360	LINC00687	Peripheral artery disease in type 2 diabetes (T2D)
rs2424217	Chr20	18490336	RBBP9	End-stage renal disease vs. no ESRD in type 2 diabetes (T2D)
rs910152	Chr20	62836546	TCFL5,COL9A3	Macroalbuminuria in type 2 diabetes (T2D)
rs4809287	Chr20	63310395	COL20A1	Youth-onset type 2 diabetes (T2D)
rs5760472	Chr22	24557603	GUCD1,SNRPD3	Microalbuminuria in type 2 diabetes (T2D)
rs1476576	Chr22	30264528	OSM	Type 2 diabetes (T2D) adjusted BMI

Each row represents a SNP and includes the following attributes: ID: The unique identifier for the SNP; CHROM: The chromosome where the SNP is located; POS: The position of the SNP on the chromosome; GENE: The gene associated with the SNP; EFFECT: The type of genomic region or effect of the SNP; TYPE 2 DIABETES TRAIT: a genetically influenced predisposition towards type 2 diabetes.