Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2025;17(1):

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2025-108361

Prediction of Polygenic Risk Score by Machine Learning and Deep Learning Methods in Genome-wide Association Studies

Genom-boyu İlişki Çalışmalarında Poligenik Risk Skorunun Makine Öğrenimi ve Derin Öğrenme Yöntemleri ile Tahmin Edilmesi

Ragıp Onur ÖZTORNACI[a], Erdal COŞGUN[b], Cemil ÇOLAK[c], Bahar TAŞDELEN[d]

[a]University of Bristol, Department of Population Health Sciences, The MRC Integrative Epidemiology Unit, Bristol, UK
[b]Private, USA
[c]İnönü University Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye
[d]Mersin University Faculty of Medicine, Department of Biostatistics, Mersin, Türkiye

This study was presented as an oral presentation at 23rd National and 6th International Biostatistics Congress, October, 26-29, 2022, Ankara, Türkiye.

ABSTRACT Objective: We aimed to investigate whether machine learning (ML) and deep learning (DL) methods, utilizing individual-level data from genome-wide association studies (GWAS), could serve as a viable alternative to traditional polygenic risk score (PRS) calculation methods, which rely on odds ratios as weights. PRS is widely used to estimate genetic susceptibility to diseases, but its accuracy and generalizability can be affected by variations in allele frequencies and sample sizes. Given the advancements in ML and DL techniques, we explored their potential for improving risk prediction. **Material and Methods:** We generated GWAS datasets using the PLINK program, simulating genetic data under various conditions by varying allele frequencies and sample sizes. This process was repeated 100 times to assess the robustness of the approaches. We applied 2 ML algorithms-Support Vector Machine and Random Forest alongside a DL approach. The predictive performance of these methods was compared to the traditional PRS calculation, which uses odds ratios as weights. **Results:** Our findings showed that ML and DL methods provided more consistent case-control separation than the classical approach. Additionally, they exhibited reduced bias and greater stability across different genetic conditions. **Conclusion:** ML and DL approaches present a promising alternative to odds ratio-based PRS calculations, offering enhanced reliability and consistency in genetic risk prediction.

Keywords: Genome-wide association studies; polygenic risk score; deep learning; machine learning; precision medicine

ÖZET Amaç: Bu çalışmada, genom-boyu ilişkilendirme çalışması [genome-wide association studies (GWAS)] verilerinden elde edilen bireysel düzey bilgileri kullanarak, poligenik risk skoru (PRS) hesaplamasında, olasılık oranlarını ağırlık olarak kullanan geleneksel yaklaşımlara alternatif olarak, makine öğrenimi [machine learning (ML)] ve derin öğrenme [deep learning (DL)] yöntemlerinin uygulanabilirliğini araştırmayı amaçladık. **Gereç ve Yöntemler:** PLINK programı kullanılarak farklı alel frekansları ve örneklem büyüklüklerinde 100 kez tekrarlanan GWAS veri setleri oluşturuldu. Ardından, bu veri setleri üzerinde 2 farklı ML algoritması (Destek Vektör Makinesi ve Rastgele orman) ile bir DL yaklaşımı uygulandı. Bu yöntemlerin performansı, olasılık oranlarını ağırlık olarak kullanan klasik PRS hesaplama yöntemiyle karşılaştırıldı. **Bulgular:** ML ve DL yaklaşımları, klasik yönteme kıyasla vaka-kontrol ayrımında daha tutarlı sonuçlar üretti. Ayrıca, farklı alel frekansları ve örneklem büyüklükleri altında daha az yanlılık ve daha yüksek kararlılık sergiledikleri gözlendi. **Sonuç:** ML ve DL tabanlı yöntemler, PRS hesaplamasında geleneksel olasılık oranına dayalı yaklaşımlara kıyasla daha güvenilir ve tutarlı bir risk tahmini sunarak alternatif bir yöntem olarak öne çıkmaktadır.

Anahtar kelimeler: Genom-boyu ilişki çalışmaları; poligenik risk skoru; derin öğrenme; makine öğrenimi; hassas tıp

Correspondence: Ragıp Onur ÖZTORNACI
University of Bristol, Department of Population Health Sciences, The MRC Integrative Epidemiology Unit, Bristol, UK
E-mail: onur.oztornaci@bristol.ac.uk

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 14 Jan 2025 ***Received in revised form:*** 25 Feb 2025 ***Accepted:*** 07 Mar 2025 ***Available online:*** 03 Apr 2025

Genome-wide association studies (GWAS) are methods used to identify and investigate genes and genomic regions potentially responsible for a specific disease. By analysing entire chromosome each containing hundreds of thousands of gene GWAS can assess both gene-gene and gene-environment interactions. Therefore, GWAS are not restricted to specifically selected genes; rather, they aim to identify differences between case and control groups by analysing large-scale genetic data. Changes in single nucleotides within the genome are known as single nucleotide polymorphisms (SNPs). SNPs help explain why some individuals are healthier than others, why the same disease progresses differently across individuals, and why some people respond positively to treatment while others do not. GWAS typically involve analysing SNP data. Today, numerous disease-associated SNPs have been identified, but further studies beyond GWAS are still needed. Approaches such as polygenic risk scores (PRS) may help clinicians by contributing to precision and personalized medicine before disease onset.[1,2] PRS is a measure that utilizes multiple SNPs simultaneously to calculate an individual's genetic disease risk. A low PRS indicates lower susceptibility to genetic disease, whereas a high PRS indicates a higher predisposition to disease. The objectives of this study are to identify the most accurate model for distinguishing case-control groups and predicting genetic risk, as well as to validate traditional PRS calculations using machine learning (ML) and deep learning (DL).

MATERIAL AND METHODS

In this study, for the purpose of finding out the best model of PRS, a raw GWAS data set (which is bad, bim, and fam files) that has been simulated from 1000 genome project real datasets, consisting of 251 cases and 232 controls, as well as 489805 SNPs that are associated with obesity, is used. To determine cases and controls, a body mass index above 30 was used as a criterion.[3] In our simulation scenarios, we created different odds ratios for each case and control group and different sample sizes. If compared with all SNPs, a number of the SNPs associated with disease have a 1% rate in all datasets. For the purposes of separating cases from controls, the SNPs associated with disease have been created by a less than 0.05 p value obtained from logistic regression analysis. This study was planned as a methodological study. This study was conducted in accordance with the principles of the Helsinki Declaration.

PRS

A PRS is a method used to estimate the genetic risk of a complex disease. By design, a PRS incorporates all variants across the genome, thereby potentially providing more information than any single mutation. Mathematically, it is calculated as the sum of weighted genotypes across relevant SNPs:

$$\text{PRS}= \sum_{i=1}^{n} w_i * X_i \ , \qquad (1)$$

where w_i is the weight assigned to the i-th SNP, and X_i indicates the genotype (e.g., the number of risk alleles) for the i-th SNP.[3] In this context, the odds ratio of each SNP typically serves as its weight.

ML METHODS

ML refers to a system of algorithms designed to identify patterns within a dataset and autonomously improve their own predictive performance. In other words, ML algorithms are self-improving: they enhance their prediction accuracy over time by analysing and learning from data.[4] ML is increasingly popular today because, unlike classical statistical methods, it does not rely on strict assumptions such as normality or specific sample sizes. Moreover, its capacity to handle numerous tuning parameters for solving non-linear problems is a key advantage.[5,6] Genetic epidemiological datasets often fall under the umbrella of "big data", so any methods used need to be able to handle large, complex datasets. Given the strong predictive power of ML approaches, they have become more commonly applied in big-data analyses.[7]

SUPPORT VECTOR MACHINES

Support vector machines (SVM) were first introduced by Vladimir Vapnik in 1992. They use Lagrange multipliers to solve classification problems, thereby identifying the decision boundaries that separate classes with minimal error. Although SVMs often have longer training times compared to some other algorithms, they are notably robust against overfitting.[8,9]

$$\min_{\omega,b,\xi} \frac{1}{2}\omega^T * \omega + C * \sum_{i=1}^{n} \xi_i \quad (2)$$

ω_m, b, ξ, V for the variables,

$$y_i * f_c(x_i) \geq 1 - \xi_i \quad (3)$$

$\forall_i \; and \; \xi_i \geq 0$ where, C is regulation parameter, ξ is slack variable V is tuning parameter and ω is weights. If Lagrange multipliers are used to find the smallest sums to be used to obtain the optimization solution in the equation,

$$\mathscr{L}_C = \frac{1}{2}\sum_{m=1}^{p}\|\omega_m\|^2 + C * \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i * (y_i * f_C(x_i) - 1 + \xi_i) - \sum_{i=1}^{n}\beta_i * \xi_i \quad (4)$$

$\mathscr{L}_C$ is the weight vector of the variable. ω_m, if the derivative is taken according to w;

$$\frac{d\mathscr{L}_C}{d\omega_m} \; ; \; \omega_m = \sum_{i=1}^{n}\alpha_i * y_i * \eta_m(x_i|V_i) * \emptyset_m(x_i) \quad (5)$$

(5) obtained. Similarly, if the derivative operation is continued,

$$\frac{d\mathscr{L}_C}{db} \; ; \; \sum_{i=1}^{n}\alpha_i * y_i = 0 \quad (6)$$

and

$$\frac{d\mathscr{L}_C}{d\xi_i}; \; C = \alpha_i + \beta_i \quad (7)$$

(7) equality is achieved. Dual formulation for $\mathscr{L}_C$ Lagrange multiplier,

$$\max \; \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i * \alpha_j * y_i * y_j * K_\eta(x_i, x_j) \quad (8)$$

(8) It is calculated by the equation, and

$$\sum_{i=1}^{n}\alpha_i * y_i = 0 \quad (9)$$

Therefore, the interval, $C \geq \alpha_i \geq 0$ exists and the kernel function is defined as follows.

$$K_\eta(x_i, x_j) = \sum_{i=1}^{n}\eta_m(x_i|V_i) * K_m(x_i, x_j) * \eta_m(x_j|V_i) \quad (10)$$

To determine the maximum distance between two classes for classification with the help of weight vectors, the equations, $wx_k + b = -1 \; ve \; wx_l + b = +1$ are used and these equations are subtracted from each other and maximized. Hence, these weight vectors are unbiased. Thus, the weights (represented by "w") will be unbiased when used for calculating the PRS.

α_i (Lagrange Multipliers) determine the influence of each training sample on the decision boundary. Only samples with nonzero α_i (support vectors) affect the final model. **β_i e**nsure that the slack variables **ξ_i** are non-negative. **C (Regularization Parameter)** balances the trade-off between maximizing the margin and minimizing classification errors. **ξ_i (Slack Variables)** allow flexibility by accommodating margin violations or misclassifications. **V (Tuning Parameters) a**ppear in the feature mapping functions and influence how

the data is transformed. They affect the behaviour of the kernel and the decision boundary. **b (Bias Term) s**hifts the decision boundary so that it is correctly positioned relative to the data points. **K_ψ(x_i, x_j) (Kernel Function)** allows the SVM to operate in a high-dimensional feature space efficiently, by implicitly computing inner products without explicitly transforming the data.

RANDOM FOREST

Random Forest (RF) is an algorithm composed of multiple decision trees. It can be applied to both categorical and continuous data in classification or regression tasks. RF is also useful for missing data imputation, feature selection, and determining each variable's contribution to the model-often referred to as variable importance.[10-13]

$$\sum \sum_{i=1}^{n} (f(G_i.T)|T|)(f(G_i,T)|T|) \tag{11}$$

Here (11), while T represents the training set, the function of the probability of being selected as belonging to the class G_i is$(f(G_i.T)|T|)$. Another strength of the random forest algorithm is that the variables contributing to the model can be calculated with the "variable importance" score. This score can be calculated in three different ways, with the out-of-bag sampling method, the bootstrap sampling method, and the calculation method based on the z-score.

$$VI^{(t)}(x_j) = \frac{\sum_{i\in\overline{\mathfrak{B}}(t)} I(y_i,=\widehat{y_i}^{(t)})}{\overline{\mathfrak{B}}(t)} - \frac{\sum_{i\in\overline{\mathfrak{B}}(t)} I(y_i,=\widehat{y_{i,\pi_j}}^{(t)})}{\overline{\mathfrak{B}}(t)} \tag{12}$$

Here (12), the expression, $\overline{\mathfrak{B}}^{(t)}$is the sample obtained by out-of-bag sampling method for a tree t, with t∈{1, 2, ..., n.tree}, $\widehat{y_i}^{(t)} = f^{(t)}(X_i)$. Here, i in the prediction class. Before the observation is included in the prediction class and $\widehat{y_{i,\pi_j}}^{(t)} = f^{(t)}(X_i,\pi_j)$. $\widehat{y_{i,\pi_j}}^{(t)} = f^{(t)}(X_i,\pi_j)$ i. your observation j. permutation up to observation and $X_i,\pi_j = (X_{i,1},X_{i,2},\ldots\ldots,X_{\pi_j(i),j},X_{\pi_j,j+1},..,X_{i,p})$. $VI^{(t)}$ Is the significance score of the variable. If $VI^{(t)}$ is equal to zero, then the variable $X_j$is not included in the t tree. If we continue with the Bootstrap method,

$$VI(x_j) = \frac{\sum_{t=1}^{n-number\ of\ tree} VI^{(t)}(x_j)}{n-number\ of\ tree} \tag{13}$$

Here (13), the score $VI^{(t)}(x_j)$is obtained from the number of independent n-trees. Finally, if it is desired to calculate with the z-score method,

$$\widetilde{VI}(x_j) = \frac{VI(x_j)}{\frac{\hat{\sigma}}{\sqrt{n-number\ of\ tree}}} \tag{14}$$

It is calculated by its formulation (14). Thus, the weights (represented by "i") will be unbiased when used for calculating the PRS.

DL METHODS

DL is a specialized subset of ML. While one key difference between DL and traditional ML lies in how data is processed particularly regarding feature selection and labelling DL does still require data preparation. However, DL is especially adept at automatically learning features from raw data, making it highly suitable for complex tasks.[14] Major breakthroughs in DL have been achieved in image and speech recognition, as well as in natural language processing and language translation. The success of DL can be attributed to its ability to learn hierarchical representations of data by progressively increasing the level of abstraction.[15,16] DL architectures typically involve artificial neural networks with multiple nonlinear layers, where each layer

processes the output of the preceding layer. Each layer comprises one or more artificial neurons, which receive inputs from the neurons in the previous layer. These neurons compute a weighted sum of their inputs and apply an activation function such as the sigmoid or rectified linear unit (ReLU) to produce an output. During training, the most suitable hierarchical representations are learned by optimizing the weight parameters in each layer. After the forward pass sequentially propagates signals through all layers, the loss function in the final output layer calculates the error between the predicted output and the true label. To minimize this error, a backward pass back-propagates the error signals, updating the weight parameters using optimization algorithms based on stochastic gradient descent.[17]

$$\Phi(X, W^1, \ldots\ldots W^K) = \Psi_K(\Psi_{K-1}(\Psi_2(X * W^1) * W^2) \ldots W^{K-1}) * W^K \quad (15)$$

Here (15), Φ is a matrix of size N×C, where C=d_K is the dimensionality of the network's output (equal to the number of classes in a classification task). The notation W=$\{W^K\}_{k=1}^K$, represents the set of weight matrices in the network, and X denotes the input data. The formulation used to optimize these weights during the training phase is given by:

$$\min_{W=\{W^K\}_{k=1}^K} \ell(Y, \Phi(X, W^1, \ldots\ldots W^K)) + \lambda\, \theta(W^1, \ldots\ldots W^K) \quad (16)$$

It is defined as (16). Here (16), the term $\ell(Y, \Phi)$ is defined as the loss function. Φ is a regulation term used to eliminate the overfitting problem and is calculated with $\Phi(W)=\sum_{k=1}^K ||W^k||_F^2$ and the loss function $\ell(Y, \Phi) = ||Y - \Phi||_F^2$ and $\min_w \ell(Y, \sum_i h_i(X) w_i) + \lambda\, ||w||$ (17)

Classification is made according to the result of equation (17) where $h_i(X)$ represents a possible latent unit activation and X is the training dataset and $\lambda > 0$ is the equilibrium parameter.

BIAS SOURCES IN GWAS ODDS RATIO ESTIMATION

GWAS face various biases in odds ratio (OR) estimation, which can result in distorted associations, false-positive findings, or overestimated effect sizes.

POPULATION STRATIFICATION

Population stratification occurs when genetic variations reflect ancestry differences rather than associations with a trait. For example, allele frequency differences between subpopulations (e.g., African versus European ancestry) can confound results, making it appear as if certain variants are associated with a disease, while they simply reflect population structure.[18]

WINNER'S CURSE

This bias refers to the inflation of effect sizes for genetic variants discovered in initial studies. Variants that reach statistical significance in a discovery sample often exhibit larger-than-true effect sizes due to sampling variation. Subsequent replication studies typically observe smaller effect size, reflecting the true underlying association.[19]

ASCERTAINMENT BIAS

Ascertainment bias arises when the selection of participants for GWAS is not representative of the general population. For example, cases might be recruited from hospitals while controls are sampled from the general population, introducing environmental or demographic differences that confound the genetic analysis.[20]

LINKAGE DISEQUILIBRIUM (LD)

LD refers to the non-random association of alleles at nearby loci. When a non-causal variant is in strong LD with a causal variant, it may appear to be associated with the disease. This can complicate efforts to pinpoint the true causal variant and interpret its functional impact.[21]

MULTIPLE TESTING BIAS

GWAS tests millions of SNPs simultaneously, increasing the probability of false-positive findings. For instance, even with stringent significance thresholds (e.g., $p<5\times10^{-8}$), spurious associations may arise due to the sheer number of comparisons. Proper statistical corrections, such as Bonferroni correction or false discovery rate control, are necessary to mitigate this issue.[22]

CONFOUNDING VARIABLES

Confounding occurs when external factors (e.g., environment, lifestyle) influence both the genetic variant and the phenotype of interest, creating spurious associations. For example, socio-economic factors might correlate with both disease risk and genetic variation, leading to biased OR estimates if not properly accounted for.[23]

CRYPTIC RELATEDNESS

Cryptic relatedness refers to hidden familial relationships among participants in a GWAS. Such relationships inflate test statistics by violating the assumption of independence between individuals, leading to an increased risk of false-positive associations. Proper quality control, such as identity-by-descent analysis, can help identify and exclude related individuals.[24]

ALTERNATIVE METHODS FOR POLYGENIC RISK SCORE PREDICTION

ML models, including RF, Support Vector Machines (SVM), and Deep Neural Networks, can mitigate biases in PRS formulation by accounting for population stratification through ancestry-adjusted features, addressing winner's curse with cross-validation and regularization, avoiding arbitrary significance thresholds by learning SNP contributions directly from data, incorporating non-linear relationships and epistatic interactions that traditional GWAS-based PRS methods overlook, handling missing data without imputation errors, and integrating environmental and genetic confounders into a unified predictive framework, all of which contribute to better-calibrated, generalizable, and less biased PRS for complex traits and diseases.

We propose an alternative approach for predicting PRS by identifying weight vectors using different ML methods. In the SVM and DL approaches, the weight matrices used to classify samples serve as weight vectors for computing PRS. For RF, we utilize the variable importance measurements as the weight vector. In all cases, each SNP is multiplied by the corresponding weight, yielding an individual risk score. The formulas are as follows:

$$PRS_{SVM} = \sum_{i=1}^{n} SVMw_i * SNP_i \quad (18)$$

$$PRS_{RF} = \sum_{i=1}^{n} RFi_i * SNP_i \quad (19)$$

$$PRS_{DL} = \sum_{i=1}^{n} DLw_i * SNP_i \quad (20)$$

Here, $SVMw_i$, RFi_i and DLw_irepresent the weights assigned to the i-th SNP by the respective methods, while SNP_i denotes the genotype for the i-th SNP.

SIMULATION STEPS

Because it is both challenging and costly to acquire GWAS data for specific diseases, we simulated data using the PLINK software. The resulting raw files (bed, bim, and fam) were then used to develop our new method and to compare the performance of different approaches. We selected case-control sample sizes of 500-500, 1,000-1,000, and 2,000-2,000 individuals. The number of SNPs studied included 5,000, 10,000, 50,000, and 100,000. The primary software and platforms used were PLINK, R, Python, Linux, and Microsoft Azure.[25-29]

TABLE 1: Tuning parameters

Model	Parameter	Value	Description
SVM (SVC)	C	1000	Regularization parameter
	Kernel	Linear	Kernel type for SVM
Randomforestclassifier	N_estimators	1000	Number of trees in the forest
	Max_depth	30	Maximum depth of the tree
	Min_samples_leaf	10	Minimum samples per leaf node
	Min_samples_split	10	Minimum samples to split a node
	Oob_score	TRUE	Use out-of-bag samples for evaluation
	Random_state	0	Random seed for reproducibility
Train_test_split	Test_size	0.2	Proportion of test data
Deep Learning	Layers	[n, 64, 1]	Layer sizes (number of neurons). N is the number of SNPs
	Activation	ReLU/sigmoid	Activation functions
	Optimizer	Adam	Optimization algorithm
	Learning rate	0.001	Learning rate
	Epochs	50	Number of training epochs
	Batch size	32	Mini-batch size
	Validation split	0.2	Proportion of validation data

SVM: Support Vector Machine; ReLU: Rectified linear unit; SNP: Single nucleotide polymorphism

RESULTS

In our simulation study, we compared PRS calculated by 4 methods: the classical (log-odds) approach, SVM, RF, and DL, across various sample sizes and SNP counts. A statistically significant difference was observed between case and control groups with PRSs derived from the classical method at all SNP counts and sample sizes ($p<0.001$). Although the classical method produced the lowest PRS values in both patient and control groups for all sample sizes, DL yielded the highest PRS values. Meanwhile, the RF method displayed a more homogeneous variation, with SVM showing a similar pattern. Overall, all newly developed methods (SVM, RF, and DL) appear suitable alternatives to the classical approach when the number of SNPs is small. For instance, with 10,000 SNPs in the 500-500 case-control group, classical, SVM, and RF methods detected a statistically significant difference in genetic risk scores ($p<0.001$), whereas DL did not ($p=0.361$). Similarly, at 50,000 SNPs, the classical, SVM, and RF methods again showed significant case-control differences ($p<0.001$), whereas DL did not ($p=0.642$). Finally, at 100,000 SNPs, the classical, SVM, and RF methods successfully distinguished case and control groups ($p<0.001$), while DL still did not ($p=0.803$). These findings suggest that SNP count is a crucial factor influencing the DL method's ability to separate cases from controls. When working with a small number of SNPs, DL performed best at distinguishing case and control groups; however, when the SNP count increases, RF, SVM, and the classical method tend to outperform DL. Notably, RF and SVM can replace the classical method for any SNP count, and they yield narrower confidence intervals while offering superior discrimination of case-control status as the number of SNPs rises (Table 2).

TABLE 2: Comparison of classical PRS and risk scores weighted according to other models in case and control groups (n=500-500)

Number of the SNPs	Method	500 Case-500 Control PRS Table		p value
		Control	Case	
		X±SD	X±SD	
		[Minimum:maximum]	[Minimum:maximum]	
5,000 SNP	PRS	0.081±0.269	0.212±0.267	<0.001
		[-0.645:0.916]	[-0.712:0.863]	
	SVM	0.089±0.129	0.172±0.077	<0.001
		[-0.178:0.428]	[-0.125:0.389]	
	RF	0.078±0.025	0.083±0.019	<0.001
		[0.033:0.128]	[0.045:0.121]	
	DL	0.489±0.318	0.505±0.296	<0.001
		[-0.098:1.247]	[-0.044:1.234]	
10,000 SNP	PRS	-0.513±0.451	-0.476±0.471	<0.001
		[-1.104:0.039]	[-1.119:0.079]	
	SVM	0.019±0.074	0.062±0.049	<0.001
		[-0.155:0.162]	[-0.075:0.156]	
	RF	0.039±0.012	0.042±0.009	<0.001
		[0.018:0.063]	[0.025:0.061]	
	DL	0.640±0.317	0.643±0.306	0.361
		[-0.016:1.441]	[0.012:1.443]	
50,000 SNP	PRS	-0.001±0.113	0.130±0.163	<0.001
		[-0.234:0.354]	[-0.241:0.587]	
	SVM	0.053±0.096	0.137±0.063	<0.001
		[-0.103:0.252]	[-0.102:0.241]	
	RF	0.077±0.024	0.084±0.017	<0.001
		[0.036:0.120]	[0.050:0.119]	
	DL	0.616±0.351	0.618±0.348	0.642
		[-0.005:1.468]	[0.004:1.468]	
100,000 SNP	PRS	0.286±0.358	0.452±0.266	<0.001
		[-0.246:0.963]	[0.045:0.986]	
	SVM	0.029±0.047	0.072±0.029	<0.001
		[-0.052:0.111]	[-0.028:0.111]	
	RF	0.038±0.013	0.042±0.008	<0.001
		[0.015:0.060]	[0.024:0.058]	
	DL	0.669±0.373	0.670±0.371	0.803
		[-0.004:1.339]	[0.002:1.351]	

SNP: Single nucleotide polymorphism; PRS: Polygenic Risk Score; SVM: Support Vector Machine; RF: Random Forest; DL: Deep Learning; SD: Standard deviation

When the sample size is 2,000, SVM is the most powerful model for distinguishing cases from controls. At 5,000 SNPs, all methods can effectively separate cases and controls, indicating they are all suitable for predicting PRS in this scenario. However, at 10,000 SNPs, the classical method, RF, and SVM continue to yield statistically significant differences ($p<0.001$), whereas DL does not ($p=0.467$). For 50,000 SNPs, the classical method, SVM, and RF again show significant separation between cases and controls ($p<0.001$), but DL does not ($p=0.652$). A similar pattern emerges at 100,000 SNPs: the classical method, SVM, and RF remain significant ($p<0.001$), while DL fails to distinguish cases from controls ($p=0.825$; Table 3).

TABLE 3: Comparison of classical PRS and risk scores weighted according to other models in case and control groups (n=1,000-1,000)

Number of the SNPs	Method	1,000 Case- 1,000 Control Polygenic Risk Score Table		p value
		Control	Case	
		X±SD	X±SD	
		[Minimum:maximum]	[Minimum:maximum]	
5,000 SNP	PRS	-0.003±0.112	0.066±0.147	<0.001
		[-0.400:0.463]	[-0.403:0.777]	
	SVM	0.195±0.151	0.282±0.071	<0.001
		[-0.139:0.540]	[-0.011:0.545]	
	RF	0.080±0.026	0.085±0.018	<0.001
		[0.027:0.136]	[0.044:0.128]	
	DL	0.673±0.320	0.689±0.295	<0.001
		[-0.043:1.432]	[0.022:1.457]	
10,000 SNP	PRS	-0.550±0.498	-0.516±0.519	<0.001
		[-1.130:0.087]	[-1.127:0.132]	
	SVM	0.073±0.095	0.119±0.062	<0.001
		[-0.168:0.261]	[-0.069:0.243]	
	RF	0.038±0.012	0.042±0.009	<0.001
		[0.015:0.069]	[0.025:0.062]	
	DL	0.902±0.430	0.905±0.419	0.467
		[0.003:1.640]	[0.061:1.657]	
50,000 SNP	PRS	-0.023±0.079	0.109±0.182	<0.001
		[-0.209:0.303]	[-0.164:0.815]	
	SVM	0.039±0.085	0.123±0.068	<0.001
		[-0.128:0.259]	[-0.076:0.245]	
	RF	0.077±0.024	0.084±0.017	<0.001
		[0.037:0.127]	[0.046:0.118]	
	DL	0.828±0.382	0.830±0.379	0.652
		[-0.002:1.551]	[0.008:1.544]	
100,000 SNP	PRS	0.094±0.189	0.234±0.172	<0.001
		[-0.174:0.545]	[-0.079:0.552]	
	SVM	0.021±0.045	0.063±0.037	<0.001
		[-0.041:0.116]	[-0.043:0.113]	
	RF	0.039±0.013	0.043±0.008	<0.001
		[0.013:0.065]	[0.028:0.060]	
	DL	0.892±0.461	0.893±0.459	0.825
		[0.000:1.727]	[0.005:1.732]	

SNP: Single nucleotide polymorphism; PRS: Polygenic risk score; SVM: Support vector machine; RF: Random forest; DL: Deep learning; SD: Standard deviation

For the largest sample size of 2,000, the DL method yielded the highest PRS values (for both cases and controls) at 5,000 SNPs, followed by SVM (range: [0.20,0.40]). The classical method produced the lowest PRS values. Although the classical method, SVM, and RF all showed significant differences between cases and controls ($p<0.001$), DL did not ($p=0.843$). At 50,000 SNPs, PRS values from the classical method were again lower for both case and control groups compared to the other methods. In contrast, RF and SVM produced similar results. While the classical method, SVM, and RF indicated a statistically significant difference between cases and controls ($p<0.001$), DL did not ($p=0.760$). This pattern persisted at 100,000 SNPs (Table 4).

TABLE 4: Comparison of classical PRS and risk scores weighted according to other models in case and control groups (n=2,000-2,000)

Number of the SNPs	Method	2,000 Case-2,000 Control Polygenic Risk Score Table		p value
		Control	Case	
		X±SD	X±SD	
		[Minimum:maximum]	[Minimum:maximum]	
5,000 SNP	PRS	-0.515±0.459	-0.485±0.472	<0.001
		[-1.135:0.029]	[-1.123:0.046]	
	SVM	0.254±0.248	0.354±0.182	<0.001
		[-0.384:0.757]	[-0.152:0.773]	
	RF	0.075±0.027	0.081±0.022	<0.001
		[0.018:0.132]	[0.030:0.129]	
	DL	1.214±0.526	1.224±0.501	0,028
		[-0.076:2.150]	[0.126:2.155]	
10,000 SNP	PRS	-0.534±0.477	-0.504±0.495	<0.001
		[-1.086:0.048]	[-1.084:0.084]	
	SVM	0.072±0.125	0.117±0.096	<0.001
		[-0.205:0.342]	[-0.210:0.343]	
	RF	0.039±0.013	0.042±0.009	<0.001
		[0.014:0.068]	[0.023:0.065]	
	DL	1.506±0.681	1.507±0.667	0.843
		[-0.072:2.448]	[0.030:2.435]	
50,000 SNP	PRS	-0.013±0.087	0.060±0.191	<0.001
		[-0.178:0.370]	[-0.177:0.639]	
	SVM	0.135±0.164	0.224±0.127	<0.001
		[-0.190:0.440]	[-0.090:0.444]	
	RF	0.077±0.023	0.084±0.017	<0.001
		[0.040:0.120]	[0.048:0.115]	
	DL	2.207±0.966	2.210±0.962	0.760
		[0.007:3.430]	[0.018:3.409]	
100,000 SNP	PRS	1.372±1.247	1.393±1.251	<0.001
		[-0.008:2.547]	[-0.002:2.574]	
	SVM	0.019±0.042	0.111±0.040	<0.001
		[-0.095:0.117]	[-0.013:0.158]	
	RF	0.027±0.004	0.036±0.004	<0.001
		[0.018:0.044]	[0.020:0.044]	
	DL	1.456±0.484	1.460±0.483	0.382
		[0.004:1.940]	[0.010:1.927]	

SNP: Single nucleotide polymorphism; PRS: Polygenic risk score; SVM: Support vector machine; RF: Random forest; DL: Deep learning; SD: Standard deviation

A statistically significant difference was observed between the case and control groups using the classical PRS calculation method and SVM ($p<0.001$); however, DL ($p=0.989$) and RF ($p=0.290$) did not produce significant differences. When the case and control groups are evaluated separately, the control group's mean of -0.214 for the classical PRS method falls within the [-0.246:0.963] simulation range. Similarly, the DL mean of 0.178 is within [-0.004:1.339]. For SVM, the simulation results closely match the mean values obtained from real data, whereas RF does not exhibit the same similarity.

In the case group, the classical, SVM, and DL methods all yield mean values that lie within the [minimum:maximum] ranges derived from the simulations (Table 5).

TABLE 5: Comparison of Classical Polygenic Risk Score (PRS) and Risk Scores Weighted by Other Models Using Real Data with Patient Group and Healthy Control Group

Method	251 Case- 232 Control Polygenic Risk Score Table		p value
	489805 SNP		
	Control	Case	
	X±SD	X±SD	
	[Minimum:maximum]	[Minimum:maximum]	
PRS	-0.214±0.049	0.216±0.051	<0.001
	[-0.289: -0.073]	[0.121:0.364]	
SVM	-0.073±0.046	0.077±0.043	<0.001
	[-0.101:0.026]	[-0.020:0.103]	
RF	0.481±0.154	0.495±0.137	0.290
	[0.005:0.550]	[0.005:0.553]	
DL	0.178±0.004	0.178±0.004	0.989
	[0.162:0.189]	[0.168:0.191]	

SNP: Single nucleotide polymorphism; PRS: Polygenic risk score; SVM: Support vector machine; RF: Random forest; DL: Deep learning; SD: Standard deviation

DISCUSSION

Overall, SVM exhibits lower standard errors and narrower confidence intervals, making it a powerful alternative for calculating PRS in both control-only and case-only groups. Similar studies have reported that risk scores could be defined for control groups alone.[30] Other methods (classical, RF, DL) remain viable options if one wishes to estimate PRS separately for case or control populations. It is important to note that genetic variation alone (captured by SNPs) does not fully explain disease risk. Clinical factors, environmental influences, and gene-environment interactions also play key roles. This is consistent with studies where clinical risk scores and PRS are used together to enhance the power of case-control discrimination.[31] The DL approach, despite its strengths with low SNP counts, tends to struggle with binary GWAS data as the number of SNPs increases. All PRS methods are effectively weight×SNP. Hence, differences in PRS primarily stem from how each method derives the weight vectors. In DL, weights were more homogeneously distributed, limiting case-control differentiation. In SVM, the algorithm naturally seeks to maximize the margin between classes, resulting in weight vectors that more distinctly separate cases and controls. RF uses variable importance measures to derive weights, which also can offer a robust discrimination, especially at higher SNP counts. Moreover, certain real-data comparisons illustrate how DL can match outcomes from smaller SNP sets. For instance, with 1,000 cases and 1,000 controls at 5,000 SNPs, DL results fit within the ranges previously reported in prostate cancer datasets from Johns Hopkins University and Ambry Genetics.[32] Another study using 289 SNPs for breast cancer risk also reached similar mean PRSs. It is critical to remember that PRS is not a diagnostic test; it does not definitively state whether an individual does or does not have a given disease. Instead, PRS quantifies the risk or probability of disease progression.[33] Consequently, ML methods (SVM, RF) may yield more stable results than DL and classical methods at higher SNP counts.[34,35] For the largest sample size and highest number of SNPs, classical and DL methods yielded PRS values above 1.00 for both cases and controls, whereas SVM and RF remained in the (0.00, 0.20) range. In our simulations with 2,000 cases and 2,000 controls and 5,000 SNPs, SVM aligned well with real-data findings.

CONCLUSION

SVM and RF generally perform better or at least comparably to the classical PRS method across a wide range of sample sizes and SNP counts. DL stands out for smaller SNP counts but loses discriminative power as SNP numbers increase. These insights can guide researchers in choosing the most appropriate PRS calculation strategy for specific study designs and sample constraints.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ragıp Onur Öztornacı, Cemil Çolak, Bahar Taşdelen;* ***Design:*** *Ragıp Onur Öztornacı, Cemil Çolak, Bahar Taşdelen;* ***Control/Supervision:*** *Ragıp Onur Öztornacı, Cemil Çolak, Bahar Taşdelen;* ***Data Collection and/or Processing:*** *Ragıp Onur Öztornacı;* ***Analysis and/or Interpretation:*** *Ragıp Onur Öztornacı;* ***Literature Review:*** *Ragıp Onur Öztornacı;* ***Writing the Article:*** *Ragıp Onur Öztornacı;* ***Critical Review:*** *Cemil Çolak, Bahar Taşdelen, Erdal Coşgun;* ***References and Fundings:*** *Ragıp Onur Öztornacı, Erdal Coşgun;* ***Materials:*** *Ragıp Onur Öztornacı, Erdal Coşgun.*

REFERENCES

1. Dorak MT. Genetic Association Studies: Background, Conduct, Analysis, and Interpretation. 1st ed. New York: Garland Science; 2016.
2. Konuma T, Okada Y. Statistical genetics and polygenic risk score for precision medicine. Inflamm Regen. 2021;41(1):18. PMID: 34140035; PMCID: PMC8212479.
3. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759-72. PMID: 32709988; PMCID: PMC7612115.
4. Alpaydın E. Introduction to Machine Learning. 3rd ed. Cambridge: The MIT Press; 2004.
5. Akpınar H. Data Veri Madenciliği Veri Analizi. 1. Baskı. İstanbul: Papatya Yayıncılık; 2013
6. Gönen M, Alpaydın E. Multiple kernel learning algorithms. Journal of Machine Learning Research. 2011;12:2211-68. https://jmlr.org/papers/volume12/gonen11a/gonen11a.pdf
7. Köse T, Özgür S, Coşgun E, Keskinoğlu A, Keskinoğlu P. Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. Biomed Res Int. 2020;2020:1895076. PMID: 32733929; PMCID: PMC7378600.
8. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. 3rd ed. San Francisco: Elsevier; 2012.
9. Pisner DA, Schnyer DM. Support vector machine. Machine Learning. 2020:101-21. DOI: 10.1016/B978-0-12-815739-8.00006-7
10. Temel, G. (2004). Clustering and Regression Trees (Yüksek Lisans Tezi). Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, Mersin.
11. Orekici Temel G, Camdeviren H, Akkus Z. Diagnosing restless legs syndrome (RLS) patients with help of classification tree. Annals of Medical Research. 2021;12(2):111-7. https://www.annalsmedres.org/articles/2005/volume12/issue2/111-117.pdf
12. Strobl C, Zeileis A. Danger: High power!-exploring the statistical properties of a test for random forest variable importance. In Brito P, eds. COMPSTAT 2008-Proceedings in Computational Statistics, Vol. II. Heidelberg: Physica-Verlag; 2008. p. 59-66. doi: 10.5282/ubm/epub.2111
13. Liu Y, Zhao H. Variable importance-weighted Random Forests. Quant Biol. 2017;5(4):338-51. PMID: 30034909; PMCID: PMC6051549.
14. Aminanto E, Kim K. Deep learning in intrusion detection systems: an overview. 2016 International Research Conference on Engineering and Technology (IRCET). Higher Education Forum. 2016.
15. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1724-34. doi: 10.3115/v1/D14-1179
16. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1412-21. doi:10.18653/v1/D15-1166
17. Hecht-Nielsen R. Theory of the Backpropagation Neural Network. Neural Networks for Perception Computation, Learning, and Architectures. 1992:65-93. https://doi.org/10.1016/B978-0-12-741252-8.50010-8
18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9. PMID: 16862161.
19. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet. 2007;80(4):605-15. PMID: 17357068; PMCID: PMC1852705.
20. Daly MJ, Altshuler D. Design of case-control studies for genome-wide association scans. Nature Genetics, 2005;37(7):671-7. DOI:10.1038/ng1580
21. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822. PMID: 23300413; PMCID: PMC3531285.
22. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440-5. PMID: 12883005; PMCID: PMC170937.
23. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. Int J Epidemiol. 2004;33(1):21-5. PMID: 15075141.
24. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005;1(3):e32. PMID: 16151517; PMCID: PMC1200427.
25. R Core Team. R language definition. Vienna, Austria: R Foundation for Statistical Computing; 2000.
26. Copeland M, Soh J, Puca A, Manning M, Gollob D. Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud. 1st ed. New York, USA: Apress; 2015. p. 3-26. doi: 10.1007/978-1-4842-1004-8
27. Python W. (2021). Python. Python Releases for Windows, 24.

28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. PMID: 17701901; PMCID: PMC1950838.
29. Sobell MG. A practical guide to Ubuntu Linux. 4th ed. Canada; Pearson Education: 2015.
30. Black MH, Li S, LaDuca H, Lo MT, Chen J, Hoiness R, et al. Validation of a prostate cancer polygenic risk score. Prostate. 2020;80(15):1314-21. PMID: 33258481; PMCID: PMC7590110.
31. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020;323(7):636-45. PMID: 32068818; PMCID: PMC7042853.
32. Placek K, Benatar M, Wuu J, Rampersaud E, Hennessy L, Van Deerlin VM, et al; CReATe Consortium; Chen W, Wu G, Paul Taylor J, McMillan CT. Machine learning suggests polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis. EMBO Mol Med. 2021;13(1):e12595. PMID: 33270986; PMCID: PMC7799365.
33. Huang S, Ji X, Cho M, Joo J, Moore J. DL-PRS: a novel deep learning approach to polygenic risk scores. Research Square. 2021. https://doi.org/10.21203/rs.3.rs-423764/v1
34. Mamani NM. Applications of machine learning techniques and polygenic risk scores to genetic disease prediction. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal. 2020;9(1):5-14. https://doi.org/10.14201/ADCAIJ202091514
35. Paré G, Mao S, Deng WQ. A machine-learning heuristic to improve gene score prediction of polygenic traits. Scientific Reports. 2017;7(1):1-11. https://doi.org/10.1038/s41598-017-13056-1