Optimizing clinico-genomic disease prediction across ancestries: a machine learning strategy with Pareto improvement

Yan Gao[1,2] and Yan Cui[1,2,3*]

Abstract

Background Accurate prediction of an individual's predisposition to diseases is vital for preventive medicine and early intervention. Various statistical and machine learning models have been developed for disease prediction using clinico-genomic data. However, the accuracy of clinico-genomic prediction of diseases may vary significantly across ancestry groups due to their unequal representation in clinical genomic datasets.

Methods We introduced a deep transfer learning approach to improve the performance of clinico-genomic prediction models for data-disadvantaged ancestry groups. We conducted machine learning experiments on multi-ancestral genomic datasets of lung cancer, prostate cancer, and Alzheimer's disease, as well as on synthetic datasets with built-in data inequality and distribution shifts across ancestry groups.

Results Deep transfer learning significantly improved disease prediction accuracy for data-disadvantaged populations in our multi-ancestral machine learning experiments. In contrast, transfer learning based on linear frameworks did not achieve comparable improvements for these data-disadvantaged populations.

Conclusions This study shows that deep transfer learning can enhance fairness in multi-ancestral machine learning by improving prediction accuracy for data-disadvantaged populations without compromising prediction accuracy for other populations, thus providing a Pareto improvement towards equitable clinico-genomic prediction of diseases.

Keywords Disease prediction, Transfer learning, Deep neural network, Health equity

Background

Clinico-genomic prediction of diseases is essential to precision medicine. Traditionally, disease prediction was primarily based on epidemiological risk factors such as lifestyle variables and family history. Recent advances in high-throughput genotyping and genome sequencing technologies have enabled genome-wide association studies (GWAS) in large cohorts, providing the foundation for genomic disease prediction. However, more than 80% of the existing GWAS data were acquired from individuals of European descent [1–7], and the ancestral (or ethnic) diversity in GWAS has not improved in recent years [1, 5]. The lack of adequate genomic data for non-European populations, who make up approximately 84% of the world's population, results in low-quality artificial intelligence (AI) models for these data-disadvantaged populations (DDPs). Genomic data

*Correspondence:
Yan Cui
ycui2@uthsc.edu
[1] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
[2] Center for Integrative and Translational Genomics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
[3] Center for Cancer Research, University of Tennessee Health Science Center, Memphis, TN 38163, USA

inequality is thus emerging as a significant health risk and a new source of health disparities [7, 8].

Genomic prediction models based on GWAS data from predominantly European ancestry have limited applicability to other ancestry groups [5, 9–15]. Recent studies indicate that cross-ancestry generalizability in polygenic models can be improved by calibrating parameters for genetic effect sizes or model sparsity (or shrinkage) patterns across ancestry groups [16–24]. However, the limitations of these models, including assumptions of linearity, additivity, and distribution normality, restrict their ability to learn and transfer complex representations across different ancestry groups. In contrast, deep neural networks have much higher model capacities to capture complex, non-linear relationships and are more adept at transfer learning [25].

We have developed a framework to address the impact of biomedical data inequality and distribution shift on multi-ancestral machine learning and demonstrated its effectiveness in cancer progression and survival prediction tasks [7, 8, 26]. This study extends that framework to optimize clinico-genomic disease prediction across ancestries. We compare the performance of different models within and across the multi-ancestral machine learning schemes and find that deep transfer learning can significantly improve disease prediction accuracy for data-disadvantaged ancestry groups without compromising the prediction accuracy for other ancestry groups. Therefore, our study shows that deep transfer learning may provide a Pareto improvement [27] for optimizing multi-ancestral genomic prediction of diseases.

Methods

Clinico-genomic datasets for multi-ancestral machine learning experiments

The genotype, phenotype, and clinical data for lung cancer, prostate cancer, and Alzheimer's disease were retrieved from the dbGaP datasets: OncoArray Consortium—Lung Cancer Studies (phs001273.v3.p2) [28–32], OncoArray: Prostate Cancer (phs001391.v1.p1) [31], Alzheimer's Disease Genetics Consortium Genome Wide Association Study -NIA Alzheimer's Disease Centers Cohort (phs000372.v1.p1) [33, 34], and Columbia University Study of Caribbean Hispanics with Familial and Sporadic Late Onset Alzheimer's disease (phs000496.v1.p1) [35–37]. Genetic ancestries of the GWAS participants determined using GRAF-pop [38] were also retrieved from the dbGaP. We implemented a standard quality-control process [39] and used the PLINK software [40] to identify disease-associated single nucleotide polymorphisms (SNPs). We excluded SNPs with a missing rate > 20%, a minor allele frequency (MAF) < 0.05, or a Hardy–Weinberg Equilibrium (HWE) p-value < 10^{-5}. We also removed the samples with sex discrepancy or a missing SNP rate > 20%. To reduce data redundancy from linkage disequilibrium (LD), we used a sliding window of 50 SNPs, with a step length of 5 SNPs and an LD cut-off coefficient of 0.2. The LD pruning procedure involved examining each window for pairs of variants with squared correlation exceeding a predefined threshold. Pairs meeting this criterion were identified, and a greedy algorithm was applied to prune variants from the window until no pairs with squared correlation above the threshold remained. For a pair of SNPs in high LD, we retained the SNP with the lower p-value and removed the other. The resulting datasets for the machine learning experiments are summarized in Table 1.

We assembled four machine learning datasets for multi-ancestral clinico-genomic prediction of diseases: a lung cancer dataset with European and East Asian populations, a prostate cancer dataset with European and African American populations, and two Alzheimer's disease datasets (one with European and Latin American populations, another with European and African American populations). Each dataset comprised two subpopulations: the European (EUR) population and a data-disadvantaged population (DDP).

We split each dataset into three parts: the training set, the validation set, and the testing set, each comprised of 80%, 10%, and 10% of the individuals, respectively, stratified by ancestry and class label (case/control). The training set was used to learn the underlying patterns in the data by fitting the machine learning model's parameters. The validation set was used to tune the hyperparameters during model training to prevent overfitting. The testing set was used to assess the performance of the model after the training process was complete.

Feature selection

We conducted the association analysis on the training set using the additive logistic regression model provided by the PLINK software. The feature mask, ANOVA F-value, and p-values were generated using the SelectKBest function from the scikit-learn machine learning software library in Python. The feature set for lung cancer comprises the top 500 (or 1000) SNPs identified in the association analysis and the clinical variables of age, sex, and smoking status. Similarly, the feature set for prostate cancer includes the top 500 (or 1000) SNPs from the association analysis and the clinical variables of age and family history.

We retrieved the lists of SNPs of 16 polygenic scores for Alzheimer's disease (AD) from the Polygenic Score Catalog [41] (with accession numbers PGS000025, PGS000026, PGS000334, PGS000779, PGS000811, PGS000812, PGS000823, PGS000876, PGS000898,

Table 1 Multi-ancestral datasets for clinico-genomic prediction of diseases

Lung cancer (dataset 1)		
European	Case	9675
	Control	8107
	Age range	15–90+
	Male	11,605
	Female	6177
East Asian	Case	1363
	Control	801
	Age range	2–90+
	Male	1924
	Female	240
Prostate cancer (dataset 2)		
European	Case	28,417
	Control	19,796
	Age range	19.0–93.0
	Male	48,213
	Female	0
African American	Case	1231
	Control	2017
	Age range	21.0–89.0
	Male	3248
	Female	0
Alzheimer's disease (dataset 3)		
European	Case	3493
	Control	1518
	Male	2211
	Female	2800
Latin American	Case	1096
	Control	1222
	Male	789
	Female	1529
Alzheimer's disease (dataset 4)		
European	Case	3493
	Control	1518
	Male	2211
	Female	2800
African American	Case	682
	Control	821
	Male	442
	Female	1061

PGS000945, PGS001348, PGS001349, PGS001775, PGS001828, PGS002280, and PGS002731), which were published in recent studies [13, 42–53]. To prevent information leakage in feature selection [54, 55], polygenic scores derived from the dbGaP datasets used in this study were excluded. We compiled a list of 22 SNPs present in any of the 16 polygenic scores and the dbGaP datasets for Alzheimer's disease used in this study. The AD feature set for our machine learning experiments includes the 22 SNPs, sex, and allele value for the Apolipoprotein E (APOE) gene. The APOE gene is crucial in AD risk and progression, primarily influencing amyloid-beta plaque accumulation, tau pathology, and lipid metabolism in the brain [56]. It has three common isoforms encoded by the ε2 (associated with reduced AD risk), ε3 (neutral), and ε4 (associated with increased AD risk) alleles.

Synthetic data

We generated synthetic datasets with built-in data inequality and distribution shifts across ancestry groups. Each synthetic dataset (D) contains data from two ancestry populations: $D = D_1 \bigcup D_2$, $D_1 = \{x_{ij}, y_i\}_{i=1}^n$, $D_2 = \{x'_{ij}, y'_i\}_{i=1}^{n'}$. Here, D_1 represents the EUR population, D_2 represents a data-disadvantaged population (DDP), n and n' are the numbers of individuals in the EUR and DDP respectively, x_{ij} is the jth feature of individual i in the European population, x'_{ij} is the jth feature of individual i in the DDP, y_i is the case/control status of individual i in the European population, y'_i is the case/control status of individual i in the DDP. The case/control status of the ith EUR individual was generated using $y_i = \begin{cases} 1 \text{ if } g_i > \text{thr} \\ 0 \text{ otherwise} \end{cases}$, where $g_i = h\sum_{j=1}^m w_j * x_{ij} + a\sqrt{1-h^2} * \zeta$, thr is the parameter to determine case-to-control ratio, w_j is the effect size of the jth SNP, ζ was sampled from a standard normal distribution over [-1,1], a is a scaling factor for variance normalization, m is the number of SNPs, and h^2 is the heritability. Similarly, the label of the ith DDP individual was generated using the function $y'_i = \begin{cases} 1 \text{ if } g'_i > \text{thr}' \\ 0 \text{ otherwise} \end{cases}$, where $g'_i = h\sum_{j=1}^m w'_j * x'_{ij} + a'\sqrt{1-h^2} * \zeta$, thr' is the parameter to determine case-to-control ratio, w'_j is the effect size of the jth SNP, a' is a scaling factor for variance normalization. The genetic effect vector $W = [w_1, w_2, w_3, \ldots w_m]$ for EUR was randomly sampled from a normal distribution over $[-1,1]$. $W' = [w'_1, w'_2, w'_3, \ldots w'_m]$ for the DDP was generated using $W' = \rho * W + \sqrt{1-\rho^2} * \zeta'$, where ρ is the correlation of the genetic effect sizes between EUR and the DDP, ζ' was randomly sampled from a standard normal distribution over [-1,1].

The feature matrices were assembled using the simulated genotype data from the Harvard Dataverse [24, 57]. The dataset includes five ancestry populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS), each consisting of 120,000 individuals. After applying a quality control process to remove SNPs with MAF < 0.05 or HWE p-value < 10^{-4}, we randomly selected 500 SNPs as features, and 10,000 EUR individuals and 2000 individuals from each of the four non-EUR populations to construct the synthetic datasets. We calculated the case/control status for various heritability

$(h^2 = 0.5$ and $h^2 = 0.25)$ and genetic correlation (ρ) values. The parameter ρ is determined using a function of genetic distance. The genetic distance between EUR and the kth DDP is $d_k = \sum_{j=1}^{500}|F_j^{\text{EUR}} - F_j^{\text{DDP}_k}|$, where F_j^{EUR} is the frequency of the minor allele of jth SNP in the EUR population, and $F_j^{\text{DDP}_k}$ is the frequency of the same allele in the kth DDP. The genetic distances were calculated as $d_0 = 42.7$ for EUR and AMR, $d_1 = 46.1$ for EUR and SAS, $d_2 = 80.6$ for EUR and EAS, and $d_3 = 94.1$ for EUR and AFR. The function $\rho_k = \rho_0 \left(\frac{d_0}{d_k}\right)^r$ was used to generate the ρ values, with $\rho_0 = 0.8$. We set $r = 0.5$ to generate the ρ values for synthetic datasets SD1-SD8 and SD1*-SD8*, and $r = 1.0$ to generate the ρ values for synthetic datasets SD9-SD16 and SD9*-SD16*. The case-to-control ratio is 1:1 for synthetic datasets SD1-SD16, and 1:4 for synthetic datasets SD1*-SD16*, respectively.

Multi-ancestral machine learning schemes and experiments

A multi-ancestral machine learning strategy should be optimized for both prediction accuracy and equity, encompassing disparity detection and mitigation. The primary challenge lies in handling datasets that contain multiple ancestry groups with data inequality and distribution shifts. Therefore, how to utilize data from different ancestry groups is crucial in a multi-ancestral machine learning strategy. We have categorized multi-ancestral (or multi-ethnic) machine learning schemes based on the way they utilize data from different subpopulations [7, 8] (Table 2). Mixture learning indistinctly uses data from all ancestral populations for model training and testing. Independent learning trains and tests a model for each ancestral group separately. In naïve transfer, the model trained on source domain (EUR) data was applied directly to the target domain (DDP) without adaptation. In transfer learning, a model is first trained on the data of the European population (source domain), then the knowledge (or representation) learned from the source domain is transferred to facilitate model development for a DDP (target domain). In this work, we assessed the performance of these multi-ancestral machine learning schemes in the context of clinico-genomic disease prediction. Our experiments are designed to address both bias detection and mitigation: experiments with mixture and independent learning schemes aim to detect model performance disparities, while those involving transfer learning focus on disparity mitigation.

Deep neural network (DNN) and DNN-based transfer learning

We constructed the deep neural network (DNN) models using the Keras (https://keras.io/) and Tensorflow

Table 2 Multi-ancestral machine learning schemes and experiments

Multi-ancestral machine learning scheme	Experiment	Ancestral composition	
		Training data	**Testing data**
Mixture learning	Mix0	EUR + DDP	EUR + DDP
	Mix1		EUR
	Mix2		DDP
Independent learning	Ind1	EUR	EUR
	Ind2	DDP	DDP
Naïve transfer	NT	EUR	DDP
Transfer learning	TL	EUR (source domain) DDP (target domain)	DDP

EUR European, *DDP* Data-disadvantaged population

(https://www.tensorflow.org/) software libraries. The DNN models were designed with a pyramid architecture [58] consisting of four layers: an input layer with K nodes for the input features of genetic and other risk factors, two hidden layers, including a fully connected layer with 100 nodes followed by a dropout layer [59], and a logistic regression output layer. We used the stochastic gradient descent (SGD) algorithm with a learning rate of 0.25 to minimize a loss function consisting of a binary cross-entropy term and two regularization terms: $l(W) = -\sum(y_i log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)) + \lambda_1|W| + \lambda_2\|W\|_2$, where y_i is the observed control/case status for individual i, $\widehat{y}_i$ is the predicted control/case status for individual i, and W represents the weights in the DNN model. We applied the ReLU activation function $f(x) = \max(0, x)$ to the hidden layer output to avoid the vanishing gradient problem. For each dropout layer, we set the dropout probability $p = 0.5$ to randomly omit half of the weights during the training to reduce the collinearity between feature detectors. We split the data into multiple mini-batches (batch size = 32) for training to speed up the computation and improve the model prediction performance. We set the maximum number of iterations at 200 and applied the Nesterov momentum [60] method (with momentum = 0.9) to prevent premature stopping. We set the learning rate decay factor at 0.003. We also used early stopping with a patience value of 200 iterations to monitor the validation accuracy during model fitting. The two regularization terms λ_1 and λ_2 were set at 0.001.

In transfer learning, knowledge and representation learned from the source domain are transferred to assist the learning task for the target domain [61–68]. In each task, we used the EUR as the source domain and the DDP as the target domain. We used a supervised fine-tuning algorithm for transfer learning. We first pretrained a DNN model using the source domain data: $M \sim f(Y_{\text{Source}}|X_{\text{Source}})$, where M represents the

pretrained model, X_{Source} and Y_{Source} represent the features and the class labels in the source domain, respectively. We trained the DNN model using the parameters described above. After the pretraining, we fine-tuned the model with the backpropagation method using the target domain data: $M' = \text{fine\_tuning}(M|Y_{\text{Target}}, X_{\text{Target}})$, where M' represents the final model, X_{Target} and Y_{Target} represent the features and the class labels in the target domain, respectively.

Logistic regression (LR) and LR-based transfer learning

Logistic regression models, capable of incorporating genetic and clinical factors, have been widely used in the clinico-genomic prediction of binary disease outcomes [69]. We used the logistic regression model with L2 regularization from the Python scikit-learn library [70]. For LR-based transfer learning, we adapted the TL_PRS [21] model, a linear polygenic model pretrained on EUR genomic data and fine-tuned on data from other ancestry groups to improve cross-population transferability. The adapted model, integrating additional terms for clinical variables, can be expressed as:

$$\widehat{Y}_i = \sum_{j=1}^{m} G_{ij}\beta_j + \sum_{k=1}^{M} C_{ij}\gamma_k + \epsilon = \sum_{j=1}^{m} G_{ij}(\beta_j^{\text{pre}} + \tau_j) + \sum_{k=1}^{M} C_{ij}(\gamma_k^{\text{pre}} + \delta_k) + \epsilon \quad ,$$

where $\widehat{Y}_i$ is the predicted phenotype of the ith sample in the target ancestry group, m is the number of SNPs, G_{ij} is the genotype of the jth SNP of individual i, β_j is the effect size of the jth SNP in the target population (a DDP), M is the number of clinical variables, C_{ik} is the kth clinical variable of individual i, γ_k is the effect size of the kth clinical variable in the target population, ϵ is the white noise from a standard normal distribution, β_j^{pre} refers to the estimated effect size of the jth SNP in EUR, τ_j is the difference between β_j^{pre} and β_j, γ_k^{pre} refers to the estimated effect size of the kth clinical variable in EUR, and δ_k is the difference between γ_k^{pre} and γ_k. For training the LR-based models, we utilized the stochastic average gradient solver which can provide efficient convergence for large datasets [71].

Application of PRS-CSx

PRS-CSx (Polygenic Risk Score with Cross-Study summary statistics) is a recently developed method that utilizes summary statistics from genome-wide association studies (GWAS) across multiple populations to construct a robust and generalizable polygenic score [16]. We evaluated the performance of PRS-CSx using our synthetic datasets. In our experiments, we followed the guidelines provided in the PRS-CSx GitHub repository (https://github.com/getian107/PRScsx). We used the software package Plink (version 1.9) to generate GWAS summary statistics for the two ancestry groups in each dataset. Subsequently, the GWAS summary statistics, combined with the LD references, were used as input for the PRS-CSx Python scripts to perform polygenic prediction.

Evaluation and comparison of machine learning model performance

We used five metrics to assess and compare the performance of machine learning models: the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPR), Tjur's R^2, positive predictive value (PPV), and negative predictive value (NPV). AUROC and AUPR are global metrics for assessing the performance of predictive models, summarizing the model's performance across all possible classification thresholds. Tjur's R^2, also known as the coefficient of discrimination, quantifies a model's ability to distinguish between binary outcomes [72]. Compared to other pseudo-R^2, this metric provides a more direct measure of the model's ability to differentiate between binary outcomes and is asymptotically equivalent to the traditional R^2 measures at large sample sizes [72]. PPV and NPV are threshold-dependent metrics that assess model performance at a single, specific classification threshold. We used the Youden's Index [73] to determine the optimal classification threshold for calculating PPV and NPV. The Youden's Index is defined as $J = \text{sensitivity} + \text{specificity} - 1$. The threshold that maximizes Youden's Index is considered optimal as it maximizes the overall correct classification rate while balancing sensitivity (true positive rate) and specificity (true negative rate). All these performance metrics have been used in recent studies to evaluate multi-ancestral clinico-genomic prediction of diseases [50, 74, 75].

The evaluation of multi-ancestral machine learning models often involves comparing model performance across ancestry groups with different disease prevalences. AUROC, which is independent of prevalence (or class distribution), is therefore used as the primary metric for assessing model performance in this study. While other metrics may vary in their sensitivity to prevalence, incorporating multiple metrics provides a more comprehensive view of model performance.

We conducted 20 independent runs for each experiment, calculated the mean values of the metrics of the 20 runs, and used a one-sided Wilcoxon rank sum test to calculate p-values to assess the statistical significance of the performance differences between the various experiments. Additionally, the one-sided Wilcoxon signed-rank test was used to evaluate the performance differences across multiple matched experiments and/or datasets.

Results
Multi-ancestral clinico-genomic prediction of diseases
We assembled four datasets for the multi-ancestral clinico-genomic prediction of lung cancer, prostate cancer, and Alzheimer's disease, utilizing data from dbGaP (Table 1). These datasets were used in machine learning

tasks to predict disease status (case/control). The specific multi-ancestral machine learning schemes and experiments are outlined in Table 2. In each experiment, we applied two machine learning models: a logistic regression (LR)-based model and a deep learning (DL)-based model (Table 3, Fig. 1). In both LR- and DL-based experiments, the mixture and independent learning schemes resulted in significant model performance disparity gaps between EUR and DDPs (Table 4). The performance disparity gap is defined as $G = \overline{\mathrm{AUROC}}_{\mathrm{EUR}} - \overline{\mathrm{AUROC}}_{\mathrm{DDP}}$, where $\overline{\mathrm{AUROC}}_{\mathrm{EUR}}$ and $\overline{\mathrm{AUROC}}_{\mathrm{DDP}}$ are the mean AUROC for the EUR and DDP in an experiment, respectively.

In mixture learning, the performance disparity gaps from the LR (Mix1_LR vs Mix2_LR) and DL (Mix1_DL vs Mix2_DL) models and the *p*-values for the statistical significance of the performance disparities are:

- Lung cancer (European and East Asian populations): $G_{\mathrm{Mix\_LR}} = 0.11 \, (p = 3.36 \times 10^{-8})$ and $G_{\mathrm{Mix\_DL}} = 0.14 \, (p = 3.36 \times 10^{-8})$
- Prostate cancer (European and African American populations): $G_{\mathrm{Mix\_LR}} = 0.10 \, (p = 3.32 \times 10^{-8})$ and $G_{\mathrm{Mix\_DL}} = 0.14 \, (p = 3.34 \times 10^{-8})$
- Alzheimer's disease (European and Latin American populations): $G_{\mathrm{Mix\_LR}} = 0.18 \, (p = 3.38 \times 10^{-8})$ and $G_{\mathrm{Mix\_DL}} = 0.10 \, (p = 3.35 \times 10^{-8})$
- Alzheimer's disease (European and African American populations): $G_{\mathrm{Mix\_LR}} = 0.20 \, (p = 3..39 \times 10^{-8})$ and $G_{\mathrm{Mix\_DL}} = 0.20 \, (p = 3.32 \times 10^{-8})$

Table 3 Multi-ancestral clinico-genomic prediction of diseases

Datasets		Area under ROC curve (LR, DL)						
Disease	DDP	Mix0	Mix1	Mix2	Ind1	Ind2	NT	TL
Lung cancer	East Asian	0.66, 0.71	0.66, 0.72	0.55, 0.58	0.64, 0.71	0.52, 0.54	0.55, 0.54	0.54, 0.64
Prostate cancer	African American	0.72, 0.74	0.73, 0.75	0.63, 0.60	0.73, 0.75	0.61, 0.67	0.58, 0.58	0.60, 0.74
Alzheimer's disease	Latin American	0.70, 0.69	0.72, 0.69	0.54, 0.58	0.72, 0.68	0.52, 0.58	0.53, 0.55	0.53, 0.62
Alzheimer's disease	African American	0.71, 0.71	0.72, 0.72	0.52, 0.52	0.73, 0.72	0.54, 0.53	0.53, 0.54	0.54, 0.60

LR Logistic regression, *DL* Deep learning, *DDP* Data-disadvantaged population. Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2
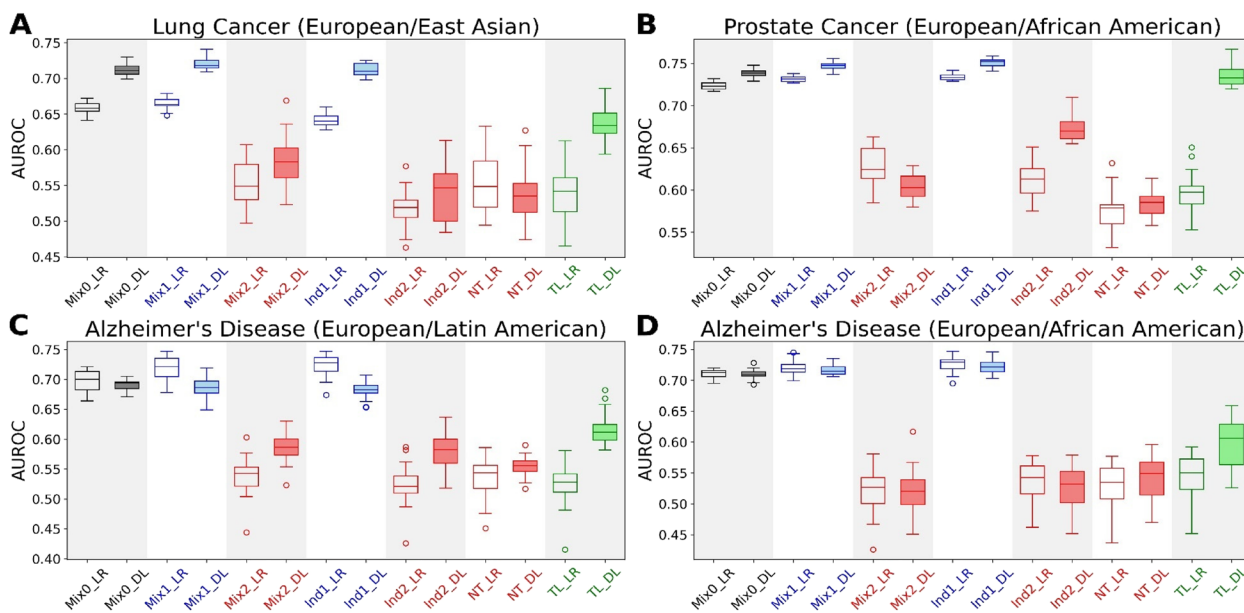

Fig. 1 Multi-ancestral clinico-genomic prediction of **A** lung cancer involving European and East Asian populations, **B** prostate cancer involving European and African American populations, **C** Alzheimer's disease involving European and Latin American populations, and **D** Alzheimer's disease involving European and African American populations. Each box plot represents the machine learning model performance (AUROC) of 20 independent runs. LR, logistic regression; DL, deep learning. Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2

Table 4 Disparity detection and mitigation in multi-ancestral clinico-genomic prediction

Key observations	Comparison	Mean difference in AUROC			
		LC	PC	AD1	AD2
Performance disparity gap between EUR and DDPs in mixture and independent learning	Mix1_LR vs Mix2_LR	**0.11**	**0.10**	**0.18**	**0.20**
	Mix1_DL vs Mix2_DL	**0.14**	**0.14**	**0.10**	**0.20**
	Ind1_LR vs Ind2_LR	**0.13**	**0.12**	**0.20**	**0.19**
	Ind1_DL vs Ind2_DL	**0.17**	**0.08**	**0.10**	**0.20**
Improvement from DL-based transfer learning	TL_DL vs Mix2_DL	**0.06**	**0.13**	**0.03**	**0.08**
	TL_DL vs Ind2_DL	**0.10**	**0.06**	**0.03**	**0.07**
	TL_DL vs NT_DL	**0.10**	**0.15**	**0.06**	**0.06**
Improvement from LR-based transfer learning	TL_LR vs Mix2_LR	-0.02	-0.03	-0.01	**0.02**
	TL_LR vs Ind2_LR	**0.02**	-0.02	0.00	0.01
	TL_LR vs NT_LR	-0.01	**0.02**	-0.01	0.02
Performance difference between DL- and LR-based transfer learning	TL_DL vs TL_LR	**0.10**	**0.14**	**0.09**	**0.06**

The statistically significant (p-value less than 0.05) performance differences are highlighted using bold font. LC, PC, AD1, and AD2 are the clinico-genomic datasets; *LC* Lung cancer (European and East Asian populations), *PC* Prostate cancer (European and African American populations), *AD1* Alzheimer's disease (European and Latin American populations), *AD2* Alzheimer's disease (European and African American populations), *LR* Logistic regression, *DL* Deep learning; Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2

In independent learning, the performance disparity gaps (Ind1_LR vs Ind2_LR, and Ind1_DL vs Ind2_DL) and p-values are:

- Lung cancer (European and East Asian populations): $G_{Ind\_LR} = 0.13$ ($p = 3.37 \times 10^{-8}$) and $G_{Ind\_DL} = 0.17$ ($p = 3.34 \times 10^{-8}$)
- Prostate cancer (European and African American populations): $G_{Ind\_LR} = 0.12$ ($p = 3.34 \times 10^{-8}$) and $G_{Ind\_DL} = 0.08$ ($p = 3.32 \times 10^{-8}$)
- Alzheimer's disease (European and Latin American populations): $G_{Ind\_LR} = 0.20$ ($p = 3.36 \times 10^{-8}$) and $G_{Ind\_DL} = 0.10$ ($p = 3.36 \times 10^{-8}$)
- Alzheimer's disease (European and African American populations): $G_{Ind\_LR} = 0.19$ ($p = 3.39 \times 10^{-8}$) and $G_{Ind\_DL} = 0.20$ ($p = 3.34 \times 10^{-8}$)

The naïve transfer approach, in which the model trained on source domain (EUR) data was applied directly to the target domain (DDP) without adaptation, also resulted in low performance for the DDPs. This is consistent with previous findings demonstrating the limited generalizability of models trained on EUR data to other ancestry groups [5, 9–15].

Using performance of mixture learning, independent learning, and naïve transfer for the DDP (Mix2, Ind2, and NT) as baselines, we quantified the improvement in model performance by DL- and LR-based transfer learning:

- $I_{Mix\_DL} = \overline{AUROC}_{TL\_DL} - \overline{AUROC}_{Mix2\_DL}$ is the performance improvement over Mix2 (TL_DL vs Mix2_DL).

- $I_{Ind\_DL} = \overline{AUROC}_{TL\_DL} - \overline{AUROC}_{Ind2\_DL}$ is the performance improvement over Ind2 (TL_DL vs Ind2_DL).
- $I_{NT\_DL} = \overline{AUROC}_{TL\_DL} - \overline{AUROC}_{NT\_DL}$ is the performance improvement over NT (TL_DL vs NT_DL).
- $I_{Mix\_LR} = \overline{AUROC}_{TL\_LR} - \overline{AUROC}_{Mix2\_LR}$ is the performance improvement over Mix2 (TL_LR vs Mix2_LR).
- $I_{Ind\_LR} = \overline{AUROC}_{TL\_LR} - \overline{AUROC}_{Ind2\_LR}$ is the performance improvement over Ind2 (TL_LR vs Ind2_LR).
- $I_{NT\_LR} = \overline{AUROC}_{TL\_LR} - \overline{AUROC}_{NT\_LR}$ is the performance improvement over NT (TL_LR vs NT_LR).

The improvements by DL-based transfer learning and the corresponding p-values are:

- Lung cancer (European and East Asian populations): $I_{Mix} = 0.06$ ($p = 1.02 \times 10^{-5}$), $I_{Ind} = 0.10$ ($p = 4.55 \times 10^{-8}$), and $I_{NT} = 0.10$ ($p = 1.28 \times 10^{-7}$)
- Prostate cancer (European and African American populations): $I_{Mix} = 0.13$ ($p = 3.37 \times 10^{-8}$), $I_{Ind} = 0.06$ ($p = 3.37 \times 10^{-8}$), and $I_{NT} = 0.15$ ($p = 3.38 \times 10^{-8}$)
- Alzheimer's disease (European and Latin American populations): $I_{Mix} = 0.03$ ($p = 2.66 \times 10^{-4}$), $I_{Ind} = 0.03$ ($p = 4.81 \times 10^{-4}$), and $I_{NT} = 0.06$ ($p = 4.57 \times 10^{-8}$)
- Alzheimer's disease (European and African American populations): $I_{Mix} = 0.08$ ($p = 4.29 \times 10^{-6}$), $I_{Ind} = 0.07$ ($p = 7.49 \times 10^{-6}$), and $I_{NT} = 0.06$ ($p = 1.87 \times 10^{-4}$)

While the DL-based transfer learning significantly improved the model performance for the DDPs, the improvements from LR-based transfer learning are largely insignificant, even with negative I_{Mix_LR}, I_{Ind_LR}, and I_{NT_LR} in many cases (Table 4).

The performance difference between DL-based and LR-based transfer learning is $D = \overline{AUROC_{TL\_DL}} - \overline{AUROC_{TL\_LR}}$ The performance differences between DL-based and LR-based transfer learning (TL_DL vs TL_LR) and the *p*-values are:

- Lung cancer (European and East Asian populations): $D = 0.10$ ($p = 5.31 \times 10^{-8}$)
- Prostate cancer (European and African American populations): $D = 0.14$ ($p = 3.38 \times 10^{-8}$)
- Alzheimer's disease (European and Latin American populations): $D = 0.09$ ($p = 3.93 \times 10^{-8}$)
- Alzheimer's disease (European and African American populations): $D = 0.06$ ($p = 2.95 \times 10^{-4}$)

In summary, the key observations from our experiments are (1) the mixture and independent learning schemes led to significant model performance disparity gaps between EUR and DDPs, regardless of the specific machine learning models (LR-based or DL-based); (2) DL-based transfer learning significantly improved model performance for the DDPs; (3) LR-based transfer learning did not achieve such substantial improvements; and (4) DL-based transfer learning significantly outperformed LR-based transfer learning (Table 4).

Four additional performance metrics, AUPR, Tjur's R^2, PPV, and NPV, were used to confirm these key observations. In the model performance assessment and comparison using AUPR and Tjur's R^2 (Table S1, Table S2, and Fig. S1), all the key observations are consistent with those obtained using AUROC as the performance metric, except in the case of lung cancer dataset, where the logistic regression (LR)-based model did not exhibit significant performance disparity gaps (measured using AUPR) in mixture and independent learning (Table S3).

PPV and NPV are dependent on disease prevalence. The sensitivity and specificity metrics can be used, along with disease prevalence in a population of interest, to calculate adjusted PPV and NPV [74]:

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)}$$

$$NPV = \frac{specificity \times (1 - prevalence)}{specificity \times (1 - prevalence) + (1 - sensitivity) \times prevalence}$$

This adjustment accounts for variations in disease prevalence across different populations, ensuring that the predictive values are more precisely aligned with the specific context. Prevalence is often reported for chronic diseases such as Alzheimer's disease but not commonly used for lung and prostate cancers. We calculated sensitivity and specificity (Table S4 and S5) and used these metrics along with ancestry-specific prevalence to compute adjusted PPV and NPV for Alzheimer's disease across all the experiments except Mix0 where the target population comprises individuals from two different ancestry groups (Table S6 and S7). The prevalence of Alzheimer's disease among individuals aged 65 and older in the USA was used: African American (13.8%), Latino (12.2%), and European (10.3%) [76]. All the key observations derived from the prevalence-adjusted PPV and NPV are consistent with those obtained using AUROC (Table S8).

Additionally, we conducted the same sets of experiments on the lung cancer and prostate cancer datasets that include the top 1000 SNPs and found that the key observations remained consistent when using more SNP features (Fig. S2 and Fig. S3).

Multi-ancestral machine learning experiments on synthetic data

To test the generalizability of these observations across diverse conditions, we conducted the same multi-ancestral machine learning experiments (as outlined in Table 2) on synthetic datasets encompassing individuals of five ancestry groups. We created two compendia of synthetic datasets with case-to-control ratio of 1:1 (SD) and 1:4 (SD*), respectively. Using synthetic data enables us to test the generalizability of our findings across a broad spectrum of conditions characterized by ancestry, heritability, and shift of genotype–phenotype relationship between ancestry groups (represented by the parameter ρ). We evaluated the machine learning performance on the synthetic datasets using AUROC, AUPR, Tjur's R^2, PPV, and NPV (Table 5 and Fig. 2, Table S9–S17 and Fig. S4–S6).

As expected, heritability (h^2) influences prediction performance in all experiments, with higher heritability linked to more accurate predictions. Nevertheless, our

Table 5 Multi-ancestral machine learning experiments on synthetic dataset compendium SD

Synthetic datasets				Area under ROC curve (LR, DL)						
ID	DDP	h^2	ρ	Mix0	Mix1	Mix2	Ind1	Ind2	NT	TL
SD1	AMR	0.50	0.80	0.77, 0.78	0.77, 0.79	0.73, 0.73	0.77, 0.79	0.69, 0.70	0.71, 0.71	0.70, 0.74
SD2	SAS	0.50	0.77	0.77, 0.79	0.78, 0.79	0.72, 0.73	0.78, 0.79	0.69, 0.70	0.71, 0.70	0.74, 0.74
SD3	EAS	0.50	0.58	0.75, 0.77	0.76, 0.78	0.67, 0.70	0.77, 0.77	0.69, 0.66	0.62, 0.56	0.67, 0.72
SD4	AFR	0.50	0.54	0.76, 0.77	0.77, 0.78	0.69, 0.72	0.77, 0.78	0.70, 0.69	0.65, 0.64	0.69, 0.74
SD5	AMR	0.25	0.80	0.65, 0.66	0.65, 0.66	0.63, 0.62	0.65, 0.65	0.57, 0.57	0.62, 0.62	0.63, 0.64
SD6	SAS	0.25	0.77	0.64, 0.66	0.65, 0.66	0.62, 0.62	0.64, 0.66	0.58, 0.59	0.61, 0.61	0.60, 0.64
SD7	EAS	0.25	0.58	0.62, 0.64	0.63, 0.64	0.59, 0.60	0.63, 0.65	0.59, 0.61	0.56, 0.55	0.54, 0.62
SD8	AFR	0.25	0.54	0.65, 0.65	0.65, 0.66	0.62, 0.61	0.65, 0.66	0.60, 0.61	0.60, 0.59	0.57, 0.64
SD9	AMR	0.50	0.80	0.77, 0.78	0.77, 0.79	0.73, 0.73	0.77, 0.79	0.69, 0.70	0.71, 0.71	0.70, 0.74
SD10	SAS	0.50	0.74	0.76, 0.77	0.77, 0.78	0.72, 0.73	0.77, 0.78	0.68, 0.69	0.70, 0.68	0.66, 0.75
SD11	EAS	0.50	0.42	0.75, 0.76	0.77, 0.78	0.64, 0.64	0.77, 0.78	0.67, 0.68	0.60, 0.60	0.63, 0.70
SD12	AFR	0.50	0.36	0.74, 0.76	0.76, 0.77	0.61, 0.66	0.77, 0.77	0.68, 0.67	0.56, 0.57	0.59, 0.71
SD13	AMR	0.25	0.80	0.65, 0.66	0.65, 0.66	0.63, 0.62	0.65, 0.65	0.57, 0.57	0.62, 0.62	0.63, 0.64
SD14	SAS	0.25	0.74	0.64, 0.64	0.65, 0.65	0.62, 0.62	0.64, 0.65	0.57, 0.57	0.60, 0.54	0.60, 0.62
SD15	EAS	0.25	0.42	0.64, 0.65	0.65, 0.66	0.57, 0.58	0.65, 0.65	0.58, 0.56	0.54, 0.52	0.55, 0.61
SD16	AFR	0.25	0.36	0.61, 0.62	0.62, 0.63	0.54, 0.55	0.63, 0.63	0.57, 0.58	0.52, 0.51	0.56, 0.61

LR Logistic regression, *DL* Deep learning, *SD* Synthetic dataset compendium with a case-to-control ratio of 1:1, *DDP* Data-disadvantaged population, *AFR* African, *AMR* Admixed American, *EAS* East Asian, *SAS* South Asian. Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2; SD9 and SD13 are the same as SD1 and SD5 and are included here for comparison

key observations remain consistent at different levels of heritability ($h^2 = 0.5$ and $h^2 = 0.25$). The parameter ρ, representing the cross-ancestry correlation of the genetic effects, is determined based on genetic distance between ancestry groups (detailed in the Methods section). Therefore, ρ indicates the degree of data distribution shifts between the EUR population and DDPs. In comparing machine learning model performance across synthetic datasets, we observed statistically significant performance disparity gaps between EUR and DDPs in both mixture and independent learning. There were also significant improvements with DL-based transfer learning, unlike with LR-based transfer learning, where improvements were generally absent. Moreover, DL-based transfer learning consistently outperformed LR-based transfer learning (Table 6, Table S18). Overall, the key observations are robust across the synthetic datasets representing various levels of heritability and data distribution shifts among ancestry groups.

We also compared the performance of DL and LR models in each of the seven types of experiments, as outlined in Table 2, using the two compendia of synthetic datasets (Table S19). In all mixture learning experiments (Mix0, Mix1, and Mix2) and one independent learning experiment (Ind1), DL outperformed LR, albeit by a small yet statistically significant (p-value < 0.05) margin, with a mean AUROC difference (DL-LR) of 0.01 to 0.02. No significant performance differences were observed in the

Ind2 and naïve transfer learning (NT) experiments. The advantage of DL over LR was largest in transfer learning, where DL-based transfer learning outperformed by a mean AUROC difference of 0.05 to 0.07. We also used AUPR, Tjur's R^2, PPV, and NPV as metrics to compare the performance of DL and LR models. While the relative performance of DL and LR vary in the mixture and independent learning scenarios, the advantage of DL over LR was consistently large (ranging from 0.04 to 0.10) and statistically significant (p-value < 0.05) in transfer learning (Table S19). Furthermore, we compared the performance of DL and LR models across all seven types of experiments. The performance differences in AUROC, AUPR, Tjur's R^2, PPV, and NPV are statistically significant (p-value < 0.05), except for the differences in the AUPR and Tjur's R^2 metrics for the SD compendium.

In a broader sense, PRS-CSx [16] can also be viewed as a form of transfer learning, as it applies knowledge from a source domain (EUR) to improve prediction accuracy in a target domain (DDP). As an extension of the PRS-CS (Polygenic Risk Score using Continuous Shrinkage) method [77], PRS-CSx retains the linear framework while accounting for cross-population differences in allele frequencies and linkage disequilibrium structures. Since the current form of PRS-CSx does not incorporate clinical features into the phenotype prediction model, it is not directly applicable to the real datasets used in this study. We compared the performance of PRS-CSx with LR- and
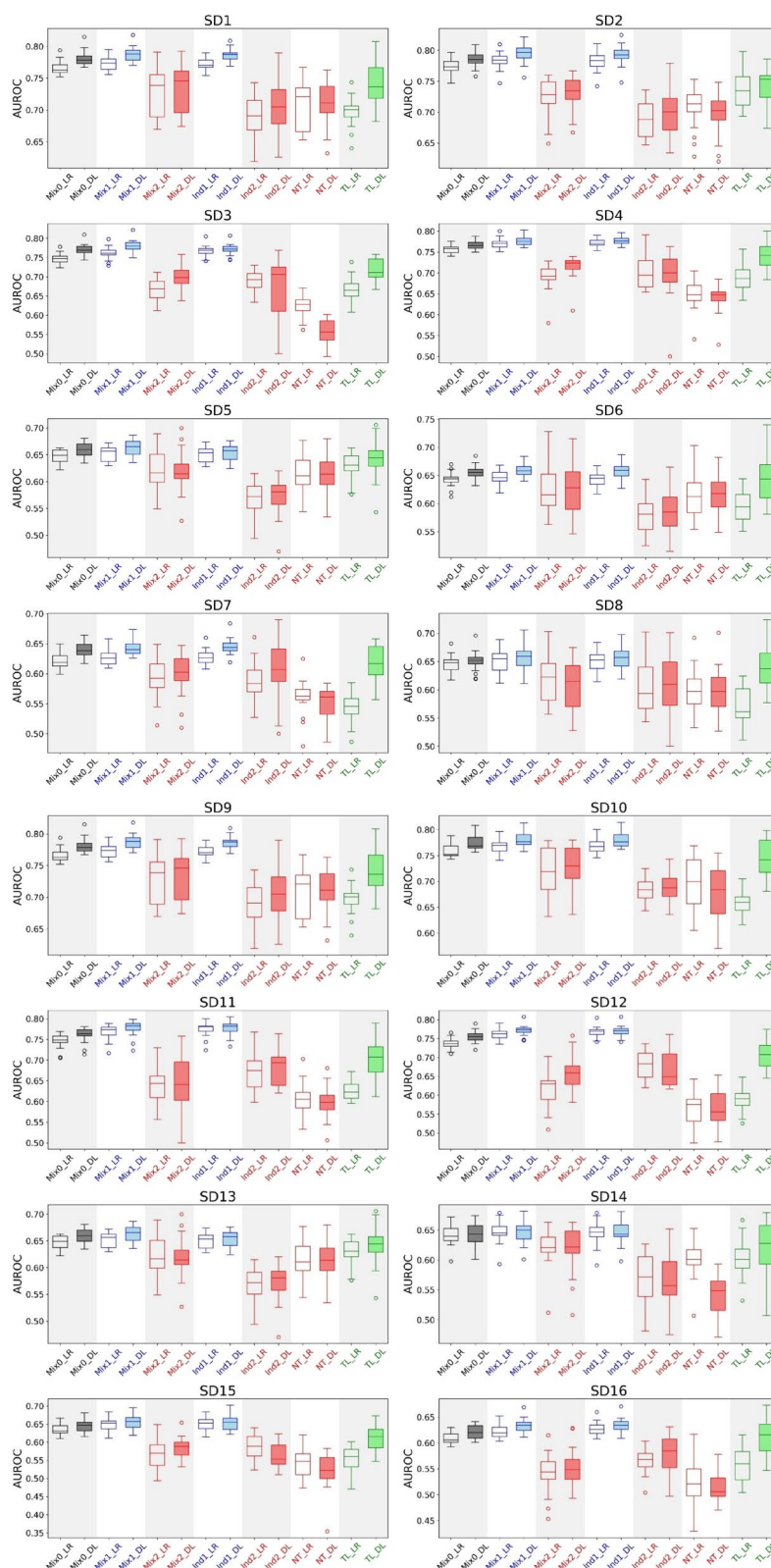
Fig. 2 Multi-ancestral machine learning experiments on synthetic dataset compendium SD. Each box plot represents the machine learning model performance (AUROC) of 20 independent runs. LR, logistic regression; DL, deep learning. Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2

Table 6 Disparity detection and mitigation in multi-ancestral machine learning using synthetic datasets

Key observations	Comparison	Mean difference in AUROC	
		SD	SD*
Performance disparity gap between EUR and DDPs in mixture and independent learning	Mix1_LR vs Mix2_LR	**0.06**	**0.07**
	Mix1_DL vs Mix2_DL	**0.06**	**0.06**
	Ind1_LR vs Ind2_LR	**0.08**	**0.07**
	Ind1_DL vs Ind2_DL	**0.08**	**0.07**
Improvement from DL-based transfer learning	TL_DL vs Mix2_DL	**0.02**	**0.03**
	TL_DL vs Ind2_DL	**0.04**	**0.04**
	TL_DL vs NT_DL	**0.07**	**0.07**
Improvement from LR-based transfer learning	TL_LR vs Mix2_LR	-0.02	-0.02
	TL_LR vs Ind2_LR	-0.01	-0.02
	TL_LR vs NT_LR	0.01	0.00
Performance difference between DL- and LR-based transfer learning	TL_DL vs TL_LR	**0.05**	**0.07**

The statistically significant (*p*-value less than 0.05) performance differences are highlighted using bold font. *LR* Logistic regression, *DL* Deep learning, *SD* Synthetic dataset compendium with a case-to-control ratio of 1:1; SD*: Synthetic dataset compendium with a case-to-control ratio of 1:4; Mix0, Mix1, Mix2, Ind1, Ind2, NT, and TL are the machine learning experiments outlined in Table 2

DL-based transfer learning on the two compendia of synthetic datasets and found that DL-based transfer learning showed the best performance on all of them (Fig. 3), confirming that DL models are more amenable to transfer learning in the context of multi-ancestral genomic prediction.

Discussion

Precision medicine increasingly relies on the predictive power of machine learning, with biomedical data forming the crucial foundation for developing high-quality models. Traditional polygenic models, primarily based on linear frameworks [78–85], are often inadequate for accurate, individualized disease prediction. This is mainly attributable to two factors: firstly, complex diseases result from an interplay of genetic, environmental, and lifestyle factors, not solely genetic determinants; and secondly, linear polygenic models lack the expressive power and model capacity to capture the non-linear, non-additive interactions inherent in the complex genotype–phenotype relationship. The inability to model non-additive genetic interactions significantly reduces the accuracy of polygenic prediction [86]. Recently, deep learning and other machine learning models that excel at handling complex nonlinearity have been used for genomic prediction of diseases [87, 88]. These models outperformed traditional polygenic prediction models in many applications [89–93]. In this study, we compared the performance of deep learning (DL) and logistic regression (LR) models across various multi-ancestral learning schemes. While DL outperformed LR in many, but not all, mixture and independent learning

experiments, DL-based transfer learning consistently outperformed LR-based transfer learning across all the real and synthetic datasets. This finding suggests that DL-based genomic prediction models are especially adept at transfer learning, even though they do not always outperform LR-based models under commonly used multi-ancestral machine learning schemes, such as mixture and independent learning.

Genomic data inequality significantly impedes the development of equitable machine learning in precision medicine, thereby posing substantial health risks to populations with limited data representation. Currently, fairness-aware machine learning [94] often relies on various ad hoc constraints or penalty terms within loss functions and training processes to enforce performance parity across different subpopulations. However, this approach leads to a fairness-accuracy tradeoff, a predicament where ensuring fairness may come at the cost of reduced accuracy for one or more subpopulations [95, 96]. A significant advantage of the deep transfer learning approach is that it is not subject to such fairness-accuracy tradeoffs. It can improve the performance of AI models for the data-disadvantaged populations by consulting but not affecting the model for the data-rich population.

Multi-ancestral machine learning can be viewed as a multi-objective optimization problem, with the goal of maximizing prediction accuracy for all ancestry groups involved in a learning task. Pareto improvement [27], a concept originated from economics, is important and relevant to multi-objective optimization because it inherently recognizes the presence of multiple, often
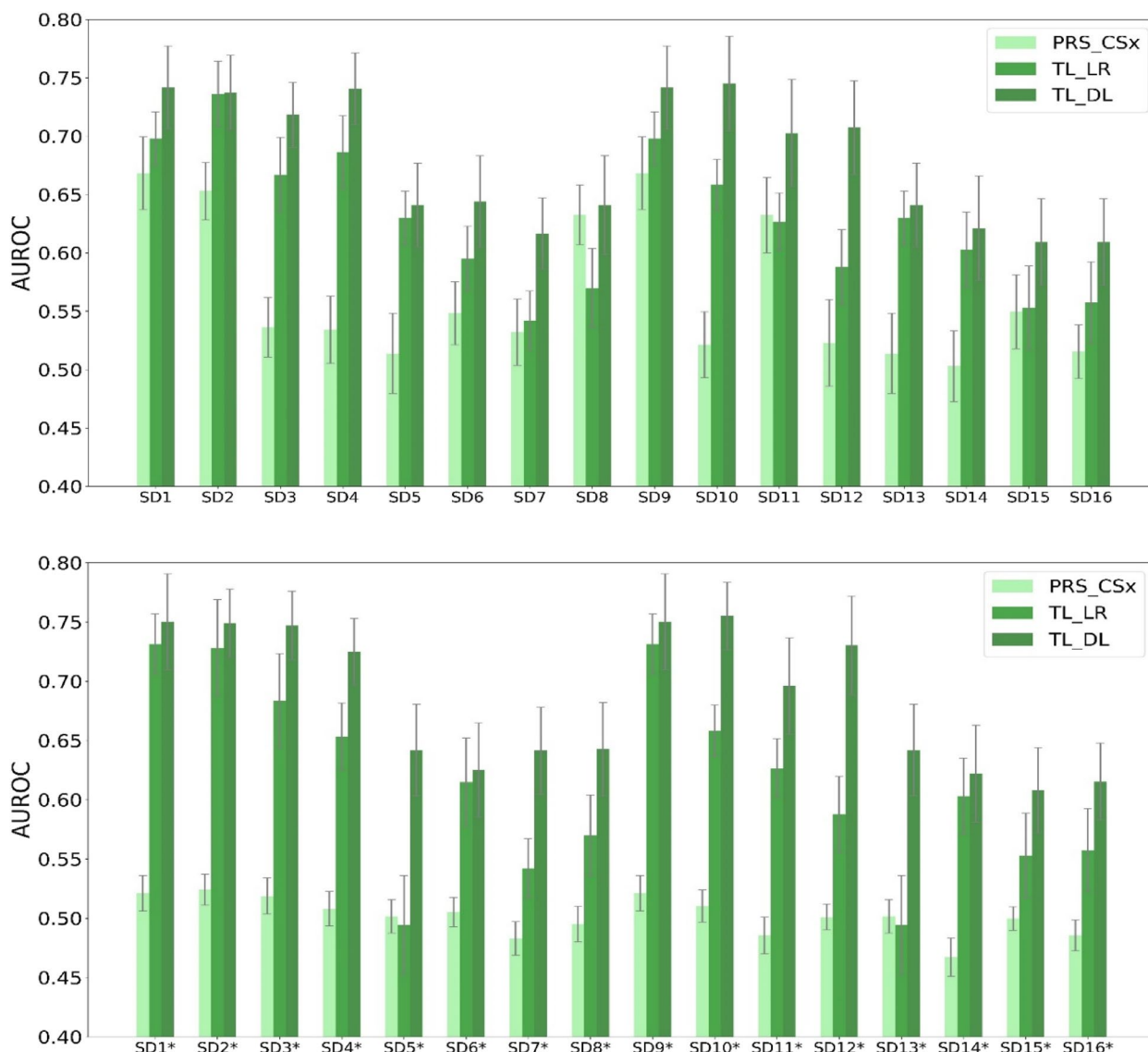
Fig. 3 Performance of PRS_CSx, TL_LR, and TL_DL across the synthetic datasets. The error bar of each column represents the standard deviation of the machine learning model performance (AUROC) of 20 independent runs. TL_LR, logistic regression-based transfer learning; TL_DL, deep learning-based transfer learning

competing, objectives and facilitates the identification of solutions that improve at least one objective without detriment to others [97, 98]. It is worth noting that Pareto improvement reflects the foundational principle in medical ethics of "first, do no harm (primum non nocere)", traditionally linked to the Hippocratic Oath taken by healthcare professionals. Recently, this principle has been extended to guide the development, deployment, and use of artificial intelligence technologies [99, 100]. Pareto improvement is particularly beneficial for advancing equitable genomic medicine, as it enhances disease prediction for data-disadvantaged populations without adversely affecting the outcomes for other populations.

Conclusions

This study shows that deep transfer learning provides a Pareto improvement towards equitable multi-ancestral machine learning for clinico-genomic prediction of diseases, as it improves prediction accuracy for data-disadvantaged populations without compromising accuracy for other populations. Machine learning experiments using synthetic data confirm that this improvement is consistent across various levels of heritability and data distribution shifts among ancestry groups.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-024-01345-0.

> Additional file 1: Supplementary tables Table S1-S19 and supplementary figures Fig.S1-Fig.S6.

Availability of data and materials
The genotype–phenotype datasets are available from the dbGaP database (https://www.ncbi.nlm.nih.gov/gap/) [101] under the study accession numbers phs001273.v3.p2 [102], phs001391.v1.p1 [103], phs000496.v1.p1 [104], and phs000372.v1.p1 [105]. The simulated genotype data are available from the Harvard Dataverse (https://doi.org/10.7910/DVN/COXHAP) [106]. The synthetic datasets are available from https://figshare.com/articles/media/TLGP_GM/25377532 [107]. The source code is available from https://github.com/ai4pm/TLGP [108].

Declarations

Ethics approval and consent to participate
All data used in this study were obtained from public datasets.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. Nat Genet. 2020;52(3):242–3.
2. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177(1):26–31.
3. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally diverse populations. Nat Rev Genet. 2019;20(9):520–35.
4. Guerrero S, López-Cortés A, Indacochea A, García-Cárdenas JM, Zambrano AK, Cabrera-Andrade A, Guevara-Ramírez P, González DA, Leone PE, Paz-y-Miño C. Analysis of racial/ethnic representation in select basic and applied cancer research studies. Sci Rep. 2018;8(1):13978.
5. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–91.
6. Bien SA, Wojcik GL, Hodonsky CJ, Gignoux CR, Cheng I, Matise TC, Peters U, Kenny EE, North KE. The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. Ann Rev Genom Human Genet. 2019;20:181–200.
7. Gao Y, Sharma T, Cui Y. Addressing the challenge of biomedical data inequality: an artificial intelligence perspective. Annu Rev Biomed Data Sci. 2023;6:153–71.
8. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat Commun. 2020;11(1):5131.
9. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. Human demographic history impacts genetic risk prediction across diverse populations. Am J Human Genet. 2017;100(4):635–49.
10. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.
11. Chen M-H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, Trivedi B, Jiang T, Akbari P, Vuckovic D, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell. 2020;182(5):1198-1213.e1114.
12. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020;11(1):3865.
13. Prive F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, Vilhjalmsson BJ. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022;109(1):12–23.
14. Zhou W, Kanai M, Wu K-HH, Rasheed H, Tsuo K, Hirbo JB, Wang Y, Bhattacharya A, Zhao H, Namba S, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genomics. 2022;2(10).
15. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, Kenny EE, Pasaniuc B. Polygenic Risk Methods in Diverse Populations Consortium Methods Working Group, Witte JS, Ge T: Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2024;25:8–25.
16. Ruan Y, Lin YF, Feng YA, Chen CY, Lam M, Guo Z, Stanley Global Asia I, He L, Sawa A, Martin AR, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022;54(5):573–580.
17. Cai M, Xiao J, Zhang S, Wan X, Zhao H, Chen G, Yang C. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am J Hum Genet. 2021;108(4):632–55.
18. Coram MA, Fang H, Candille SI, Assimes TL, Tang H. Leveraging multiethnic evidence for risk assessment of quantitative traits in minority populations. Am J Hum Genet. 2017;101(2):218–26.
19. Xiao J, Cai M, Hu X, Wan X, Chen G, Yang C. XPXP: improving polygenic prediction by cross-population and cross-phenotype analysis. Bioinformatics. 2022;38(7):1947–55.
20. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, Okada Y, Biobank Japan P, Martin AR, Finucane HK, Price AL. Leveraging finemapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022;54(4):450–8.
21. Zhao Z, Fritsche LG, Smith JA, Mukherjee B, Lee S. The construction of cross-population polygenic risk scores using transfer learning. Am J Human Genet. 2022;109(11):1998–2008.
22. Tian P, Chan TH, Wang Y-F, Yang W, Yin G, Zhang YD. Multiethnic polygenic risk prediction in diverse populations through transfer learning. Front Genet. 2022;13:906965.
23. Zhou G, Chen T, Zhao H. SDPRX: A statistical method for cross-population prediction of complex traits. Am J Human Genet. 2022;110:13–22.
24. Zhang H, Zhan J, Jin J, Zhang J, Lu W, Zhao R, Ahearn TU, Yu Z, O'Connell J, Jiang Y, et al. A new method for multiancestry polygenic prediction improves performance across diverse populations. Nat Genet. 2023;55(10):1757–68.
25. Huang X, Rymbekova A, Dolgova O, Lao O, Kuhlwilm M. Harnessing deep learning for population genetic inference. Nat Rev Genet. 2023;25:61–78
26. Gao Y, Cui Y: Multi-ethnic Survival Analysis: Transfer Learning with Cox Neural Networks. In: Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021. Edited by Russ G, Neeraj K, Thomas Alexander G, Mihaela van der S, vol. 146. Proceedings of Machine Learning Research: PMLR; 2021:252-257.
27. Black J, Hashimzade N, Myles G: A dictionary of economics: Oxford University Press, USA; 2012.

28. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46(7):736–41.

29. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeböller H, Risch A, McKay JD, Wang Y, Dai J, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet. 2012;21(22):4980–95.

30. Park SL, Fesinmeyer MD, Timofeeva M, Caberto CP, Kocarnik JM, Han Y, Love SA, Young A, Dumitrescu L, Lin Y, et al. Pleiotropic associations of risk variants identified for other cancers with lung cancer risk: the PAGE and TRICL consortia. J Natl Cancer Inst. 2014;106(4):dju061.

31. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, Casey G, Hunter DJ, Sellers TA, Gruber SB, et al. The oncoarray consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol Biomarkers Prev. 2017;26(1):126–35.

32. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126–32.

33. Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011;43(5):436–41.

34. Jun G, Naj AC, Beecham GW, Wang LS, Buros J, Gallins PJ, Buxbaum JD, Ertekin-Taner N, Fallin MD, Friedland R, et al. Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. Arch Neurol. 2010;67(12):1473–84.

35. Ghani M, Pinto D, Lee JH, Grinberg Y, Sato C, Moreno D, Scherer SW, Mayeux R, St George-Hyslop P, Rogaeva E: Genome-wide survey of large rare copy number variants in Alzheimer's disease among Caribbean hispanics. G3 (Bethesda). 2012;2(1):71–78.

36. Reitz C, Tang M-X, Schupf N, Manly JJ, Mayeux R, Luchsinger JA. A summary risk score for the prediction of alzheimer disease in elderly persons. Arch Neurol. 2010;67(7):835–41.

37. Lee JH, Cheng R, Barral S, Reitz C, Medrano M, Lantigua R, Jiménez-Velazquez IZ, Rogaeva E, St. George-Hyslop PH, Mayeux R. Identification of Novel Loci for Alzheimer Disease and Replication of CLU, PICALM, and BIN1 in Caribbean Hispanic Individuals. Arch Neurol. 2011;68(3):320–8.

38. Jin Y, Schaffer AA, Feolo M, Holmes JB, Kattman BL: GRAF-pop: a fast distance-based method to infer subject ancestry from multiple genotype datasets without principal components analysis. G3: Genes, Genomes, Genetics 2019, 9(8):2447–2461.

39. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2): e1608.

40. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Human Genet. 2007;81(3):559–75.

41. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, McMahon A, Abraham G, Chapman M, Parkinson H, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53(4):420–5.

42. Chouraki V, Reitz C, Maury F, Bis JC, Bellenguez C, Yu L, Jakobsdottir J, Mukherjee S, Adams HH, Choi SH, et al. Evaluation of a genetic risk score to improve risk prediction for Alzheimer's Disease. J Alzheimers Dis. 2016;53(3):921–32.

43. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, Thompson WK, Besser L, Kukull WA, Holland D, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 2017;14(3): e1002258.

44. Zhang Q, Sidorenko J, Couvy-Duchesne B, Marioni RE, Wright MJ, Goate AM, Marcora E, Huang KL, Porter T, Laws SM, et al. Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun. 2020;11(1):4799.

45. Zhou X, Chen Y, Ip FCF, Lai NCH, Li YYT, Jiang Y, Zhong H, Chen Y, Zhang Y, Ma S, et al. Genetic and polygenic risk score analysis for Alzheimer's

disease in the Chinese population. Alzheimers Dement (Amst). 2020;12(1): e12074.

46. Najar J, van der Lee SJ, Joas E, Wetterberg H, Hardy J, Guerreiro R, Bras J, Waern M, Kern S, Zetterberg H, et al. Polygenic risk scores for Alzheimer's disease are related to dementia risk in APOE ε4 negatives. Alzheimers Dement (Amst). 2021;13(1): e12142.

47. van der Lee SJ, Wolters FJ, Ikram MK, Hofman A, Ikram MA, Amin N, van Duijn CM. The effect of APOE and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study. Lancet Neurol. 2018;17(5):434–44.

48. Leonenko G, Sims R, Shoai M, Frizzati A, Bossù P, Spalletta G, Fox NC, Williams J, Hardy J, Escott-Price V. Polygenic risk and hazard scores for Alzheimer's disease prediction. Ann Clin Transl Neurol. 2019;6(3):456–65.

49. de Rojas I, Moreno-Grau S, Tesi N, Grenier-Boley B, Andrade V, Jansen IE, Pedersen NL, Stringa N, Zettergren A, Hernández I, et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. Nat Commun. 2021;12(1):3417.

50. Tanigawa Y, Qian J, Venkataraman G, Justesen JM, Li R, Tibshirani R, Hastie T, Rivas MA. Significant sparse polygenic risk scores across 813 traits in UK Biobank. PLoS Genet. 2022;18(3): e1010105.

51. Ebenau JL, van der Lee SJ, Hulsman M, Tesi N, Jansen IE, Verberk IMW, van Leeuwenstijn M, Teunissen CE, Barkhof F, Prins ND, et al. Risk of dementia in APOE ε4 carriers is mitigated by a polygenic risk score. Alzheimers Dement (Amst). 2021;13(1): e12229.

52. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, Naj AC, Campos-Martin R, Grenier-Boley B, Andrade V, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54(4):412–36.

53. Xicota L, Gyorgy B, Grenier-Boley B, Lecoeur A, Fontaine GL, Danjou F, Gonzalez JS, Colliot O, Amouyel P, Martin G, et al. Association of APOE-Independent Alzheimer disease polygenic risk score with brain amyloid deposition in asymptomatic older adults. Neurology. 2022;99(5):e462-475.

54. Refaeilzadeh P, Tang L, Liu H: On comparison of feature selection algorithms. In: Proceedings of AAAI workshop on evaluation methods for machine learning II: 2007: AAAI Press Vancouver; 2007: 5.

55. Molla M, Waddell M, Page D, Shavlik J. Using machine learning to design and interpret gene-expression microarrays. AI Mag. 2004;25(1):23.

56. Yamazaki Y, Zhao N, Caulfield TR, Liu C-C, Bu G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. Nat Rev Neurol. 2019;15(9):501–18.

57. Harvard Dataverse (https://doi.org/10.7910/DVN/COXHAP).

58. Phung SL, Bouzerdoum A. A pyramidal neural network for visual pattern recognition. IEEE Trans Neural Networks. 2007;18(2):329–43.

59. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learn Res. 2014;15(1):1929–58.

60. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: International conference on machine learning: 2013;2013:1139–47.

61. Yang Q, Zhang Y, Dai W, Pan SJ. Transfer learning. Cambridge University Press; 2020.

62. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.

63. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C: A survey on deep transfer learning. In: International Conference on Artificial Neural Networks: 2018: Springer; 2018:270–279.

64. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3(1):9.

65. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, Greene CS. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell Syst. 2019;8(5):380–94.

66. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR. Data denoising with transfer learning in single-cell transcriptomics. Nat Methods. 2019;16(9):875–8.

67. Sevakula RK, Singh V, Verma NK, Kumar C, Cui Y. Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks. IEEE/ACM Trans Comput Biol Bioinf. 2019;16(6):2089–100.

68. Ebbehoj A, Thunbo MØ, Andersen OE, Glindtvad MV, Hulman A. Transfer learning for non-image data in clinical research: A scoping review. PLOS Digital Health. 2022;1(2): e0000014.

69. Bazzoli C, Lambert-Lacroix S. Classification based on extensions of LS-PLS using logistic regression: application to clinical and multiple genomic data. BMC Bioinformatics. 2018;19(1):314.

70. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. J Machine Learn Res. 2011;12:2825–30.

71. Schmidt M, Le Roux N, Bach F. Minimizing finite sums with the stochastic average gradient. Math Program. 2017;162:83–112.

72. Tjur T. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. Am Stat. 2009;63(4):366–72.

73. Greenhouse SW, Cornfield J, Homburger F. The Youden index: Letters to the editor. Cancer. 1950;3(6):1097–100.

74. Ge T, Irvin MR, Patki A, Srinivasasainagendra V, Lin YF, Tiwari HK, Armstrong ND, Benoit B, Chen CY, Choi KW, et al. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. Genome Med. 2022;14(1):70.

75. Stuart PE, Tsoi LC, Nair RP, Ghosh M, Kabra M, Shaiq PA, Raja GK, Qamar R, Thelma BK, Patrick MT, et al. Transethnic analysis of psoriasis susceptibility in South Asians and Europeans enhances fine-mapping in the MHC and genomewide. HGG Adv. 2022;3(1):100069.

76. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, McGuire LC. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged≥ 65 years. Alzheimers Dement. 2019;15(1):17–24.

77. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10(1):1776.

78. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–90.

79. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28(R2):R133–42.

80. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Medicine. 2020;12(1):44.

81. Choi SW. Mak TS-H, O'Reilly PF: Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759–72.

82. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, Visscher PM. From basic science to clinical application of polygenic risk scores: a primer. JAMA Psychiat. 2021;78(1):101–9.

83. Polygenic Risk Score Task Force of the International Common Disease A. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nat Med. 2021;27(11):1876–84.

84. Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. Polygenic scores in biomedical research. Nat Rev Genet. 2022;23(9):524–32.

85. Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. 2021;37(11):995–1011.

86. Dai Z, Long N, Huang W. Influence of Genetic Interactions on Polygenic Prediction. G3. 2020;10(1):109–15.

87. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):1–16.

88. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine Learning SNP Based Prediction for Precision Medicine. Front Genet. 2019;10(267):431037.

89. Badré A, Zhang L, Muchero W, Reynolds JC, Pan C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. J Human Genet. 2020;66:359–69.

90. Elgart M, Lyons G, Romero-Brufau S, Kurniansyah N, Brody JA, Guo X, Lin HJ, Raffield L, Gao Y, Chen H, et al. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. Commun Biol. 2022;5(1):856.

91. Leist AK, Klee M, Kim JH, Rehkopf DH, Bordas SPA, Muniz-Terrera G, Wade S. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Sci Adv. 2022;8(42):eabk1942.

92. Gao Y, Cui Y. Clinical time-to-event prediction enhanced by incorporating compatible related outcomes. PLOS Digital Health 2022;1(5):e0000038.

93. Zhou X, Chen Y, Ip FCF, Jiang Y, Cao H, Lv G, Zhong H, Chen J, Ye T, Chen Y, et al. Deep learning-based polygenic risk analysis for Alzheimer's disease prediction. Commun Med (Lond). 2023;3(1):49.

94. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1–35.

95. Zhao H, Gordon G. Inherent tradeoffs in learning fair representations. J Mach Learn Res. 2022;23:1–26.

96. Menon AK, Williamson RC. The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency: 2018: PMLR. 2018:107–118.

97. Censor Y. Pareto optimality in multiobjective problems. Appl Math Optim. 1977;4(1):41–59.

98. Cho J-H, Wang Y, Chen R, Chan KS, Swami A. A survey on modeling and optimizing multi-objective systems. IEEE Commun Surveys Tutor. 2017;19(3):1867–901.

99. Goldberg CB, Adams L, Blumenthal D, Brennan PF, Brown N, Butte AJ, Cheatham M, deBronkart D, Dixon J, Drazen J, et al. To Do No Harm — and the Most Good — with AI in Health Care. Nejm Ai. 2024;1(3):Alp2400036.

100. Goldberg CB, Adams L, Blumenthal D, Brennan PF, Brown N, Butte AJ, Cheatham M, deBronkart D, Dixon J, Drazen J, et al. To do no harm - and the most good - with AI in health care. Nat Med. 2024;30:623–7.

101. National Library of Medicine. The database of Genotypes and Phenotypes. dbGaP, https://www.ncbi.nlm.nih.gov/gap.

102. Oncoarray Consortium. Lung Cancer Studies. dbGaP, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v3.p2.

103. Oncoarray Consortium. OncoArray: Prostate Cancer. dbGaP, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001391.v1.p1.

104. Genetic Consortium for Late Onset Alzheimer's Disease. Columbia University Study of Caribbean Hispanics with Familial and Sporadic Late Onset Alzheimer's disease. dbGaP, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000496.v1.p1.

105. Alzheimer's Disease Genetics Consortium. ADGC Genome Wide Association Study -NIA Alzheimer's Disease Centers Cohort. dbGaP, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000372.v1.p1.

106. Zhang H. et.al. Simulated data for 600,000 subjects from five ancestries. Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP.

107. Gao Y, Cui Y. Synthetic Datasets for Genomic Prediction Project. Figshare; 2024. https://figshare.com/articles/media/TLGP_GM/25377532.

108. Gao Y, Cui Y. Source code for Genomic Prediction Project. GitHub; 2024. https://github.com/ai4pm/TLGP.

Publisher's Note