Convolutional Embedded Networks for Population Scale Clustering and Bio-Ancestry Inferencing

Md. Rezaul Karim ⓘ, Michael Cochez ⓘ, Achille Zappa ⓘ, Ratnesh Sahay ⓘ,
Dietrich Rebholz-Schuhmann ⓘ, Oya Beyan ⓘ, and Stefan Decker ⓘ

Abstract—The study of genetic variants (GVs) can help find correlating population groups and to identify cohorts that are predisposed to common diseases and explain differences in disease susceptibility and how patients react to drugs. Machine learning techniques are increasingly being applied to identify interacting GVs to understand their complex phenotypic traits. Since the performance of a learning algorithm not only depends on the size and nature of the data but also on the quality of underlying representation, deep neural networks (DNNs) can learn non-linear mappings that allow transforming GVs data into more clustering and classification friendly representations than manual feature selection. In this paper, we propose convolutional embedded networks (CEN) in which we combine two DNN architectures called convolutional embedded clustering (CEC) and convolutional autoencoder (CAE) classifier for clustering individuals and predicting geographic ethnicity based on GVs, respectively. We employed CAE-based representation learning to 95 million GVs from the '1000 genomes' (covering 2,504 individuals from 26 ethnic origins) and 'Simons genome diversity' (covering 279 individuals from 130 ethnic origins) projects. Quantitative and qualitative analyses with a focus on accuracy and scalability show that our approach outperforms state-of-the-art approaches such as VariantSpark and ADMIXTURE. In particular, CEC can cluster targeted population groups in 22 hours with an adjusted rand index (ARI) of 0.915, the normalized mutual information (NMI) of 0.92, and the clustering accuracy (ACC) of 89 percent. Contrarily, the CAE classifier can predict the geographic ethnicity of unknown samples with an F1 and Mathews correlation coefficient (MCC) score of 0.9004 and 0.8245, respectively. Further, to provide interpretations of the predictions, we identify significant biomarkers using gradient boosted trees (GBT) and SHapley Additive exPlanations (SHAP). Overall, our approach is transparent and faster than the baseline methods, and scalable for 5 to 100 percent of the full human genome.

Index Terms—Population genomics, genotype clustering, bio-ancestry inference, deep neural networks, representation learning

✦

1 INTRODUCTION

GENETIC ariations (GVs) are structural variation in the DNA sequence in human genomes, which makes us all unique in terms of phenotype, e.g., genetic polymorphism is implicated in numerous diseases and constitute the majority of varying nucleotides among human genomes [1]. Genetic research has played a significant role in the discovery of new biological pathways underpinning complex human disease and the evaluation of new targets for therapeutic development [2]. In particular, biological understanding of the relationship between GVs can provide insights into the disease status of a patient (e.g., interacting GVs contributing to breast cancer risk) [3], which is the prerequisite to enable personalized treatment. Further benefits of studying GVs lies in the discovery and description of the genetic contribution to many human diseases based on their haplogroup and ethnicity [4]. Subsequently, finding similar population groups and identifying patients who are predisposed to common or rare diseases at early stages is increasingly important in understanding the effects of biomarkers on the development of certain disease [5].

Despite strides in characterizing human history from GVs data, progress in identifying genetic signatures of recent demography has been limited. Predicting haplogroup and ethnicity accurately (called bio-ancestry inferencing (BAI)[1]) is very challenging in which one of the most critical tasks is the analysis of genomic profiles to attribute individuals to specific ethnic populations or the interpretation of nucleotide haplotypes for diseases susceptibility [6]. Consequently, BAI is a frequently studied problem, with the main goal of identifying an individual's population of origin based on our knowledge of natural populations [7]. Accordingly, BAI has numerous applications in forensic analyses, genetic association studies, and personal genomics. Further, BAI is used as a checksum method to verify a sample's integrity, e.g., case-control studies [4], where BAI is key to understand

- M.R. Karim, O. Beyan, and S. Decker are with the Fraunhofer Institute for Applied Information Technology, 53754 Sankt Augustin, Germany, and also with RWTH Aachen University, 52062 Aachen, Germany. E-mail: {rezaul. karim, stefan.decker}@fit.fraunhofer.de, beyan@dbis.rwth-aachen.de.
- M. Cochez is with the Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands. E-mail: m.cochez@vu.nl.
- A. Zappa and R. Sahay are with the Insight Centre for Data Analytics, National University of Ireland Galway, H91 TK33 Galway, Ireland. E-mail: achille.zappa@insight-centre.org, ratnesh.sahay@gmail.com.
- D. Rebholz-Schuhmann is with the German National Library of Medicine, University of Cologne, 50923 Köln, Germany. E-mail: Rebholz@zbmed.de.

1. Geographic origin of an individual based on genetic variants

population stratification across the cohorts to help avoid probable spurious associations with even subtle ancestry differences [8]. Identification of ancestry is an important research challenge in which any direct assessment of disease-related GVs will yield more insights [9].

Since the race of an individual depends on ancestry, grouping each into a cluster is expected to correlate with the traditional concepts of race. However, this correlation is not perfect since GVs occurs according to probabilistic principles, which often does not follow any continuous distribution across races but slightly overlaps across diverse populations, e.g., research [10] has exposed that population groups from Asia, Europe, Africa, and America can be separated based on their genomic data based on the fact that 'Y' chromosome lineage can be geographically localized, forming an evidence for clustering of human alleles. In this study, we try to understand: i) if individual genetic profiles can be used to attribute into specific populations, ii) if individual GVs are suitable for predicting it's ethnicity, iii) how individuals are distributed geographically, e.g, among population groups. However, grouping individuals or BAI based on GVs requires access to genomics data and efficient analytic methodologies to cope with millions of GVs from thousands of individuals [11], [12], [13], [14]. However, extracting significant features from a large-scale GVs is not only computationally expensive but also a critical bottleneck.

Previous approaches [15], [16] employed machine learning (ML) based approaches to answer these questions. In particular, ML-based clustering algorithms such as DBSCAN, OPTICS, Gaussian Mixture (GMM), Agglomerative clustering (AC), and K-means have long been used in literature to address issues with higher-dimensional input spaces. However, they are fundamentally limited to linear embedding [17]. Often ML-based approaches failed to exploit non-linear relationships from high-dimensional and cannot exhibit good accuracy at BAI and clustering tasks. The reason lies in the fact that non-linear embedding is often necessary to deal with high-dimensionality before initializing the clustering operation [18]. In contrast, approaches based on neural networks (DNNs) can be more effective at RL and feature extraction [18]. In particular, DNN architectures (e.g., autoencoder (AE)) with multiple hidden layers and non-linear activation functions, can capture more complex and higher-level features and contextual information from the input [19], [20], [21], [22]. Further, non-linear mappings allows transforming input data into more clustering-friendly representations in which the data is mapped into a lower-dimensional feature space that helps fine-tune clustering [23].

Although DNNs have shown tremendous success to deal with such complex tasks, they are mostly perceived as 'black box' methods because of lack of understanding of their functionalities [24], which is a serious drawback. To recommend more accurate treatments and drug repositioning, interpretability is essential to provide insights on why and how a certain prediction is made by the algorithm outlining important biomarkers. Nevertheless, according to article 22 of the European Union (EU) General Data Protection Regulation (GDPR) states that individuals have the right not to be subject to a decision based solely on automated processing and whenever human subjects have their lives significantly impacted by an automatic decision-making machine, the human subject has the right to know why the decision is made, i.e., "right to explanation"[25].

We try to address the challenges and requirements in a scalable and efficient ways: first, we use Spark and ADAM for processing large-scale GVs to convert them into genotype objects. Then convolutional autoencoder (CAE) is employed representation learning on GVs data. Learned features are then used to: i) train the convolutional embedding clustering (CEC) for clustering individual to determine inter and intra-population groups, ii) train the CAE classifier to predict ethnicity of unknown samples. Finally, to provide interpretations of the predictions and to identify relevant features that contributed most, we identify significant biomarkers using gradient boosted trees (GBT) and SHapley Additive exPlanations (SHAP) [26]. The rest of the paper is structured as follows: Section 2 discusses related works and analyze their potential limitations. Section 3 chronicles our proposed approach in detail with materials and methods. Section 4 demonstrates some experiment results, discuss the findings, and highlights potential limitations of the study. Section 5 provides some explanations of the importance and relevance of the research reported and discussed some future works before concluding the paper.

2 RELATED WORK

Although 1000GP consortium has developed a global reference for human genetic variation for exome and genome sequencing and despite strides in characterizing human history from genetic polymorphism data, progress in identifying genetic signatures of recent demography has been limited [27]. Lek M. *et al.* [28], describe aggregation and analysis of high-quality protein-coding region and DNA sequence data for 60,706 individuals of diverse ancestries in which the objective metrics of pathogenicity are calculated for sequence variants against various classes of mutations. Their approach: i) identify as much as 3,230 genes with near-complete depletion of predicted protein-truncating variants, while 72 percent of these genes have no currently established human disease phenotype, ii) demonstrate that GVs can be used for efficient filtering of candidate disease-causing variants, which helps in the discovery of human 'knockout' variants in protein-coding genes.

Studies on population structure clustering include [27], which identify fine-scale population structure clustering of 770,000 genomes in North America, which reveals post-colonial population structure. Another used approach is ADMIXTURE [15], which performs maximum likelihood estimation (MLE) of individual ancestries from multilocus SNP genotype data. However, this approach cannot cluster GVs comfortably giving an ARI of only 0.25. Further, ADMIXTURE requires a preprocessing step from VCF to PED format, which takes a significant amount of time. To address the shortcomings of ADMIXTURE, VariantSpark is proposed [16], which provides an interface from MLlib that offers a seamless genome-wide sampling of variants and provides a pipeline for visualizing results from the 1000GP and PGP. However, overall clustering accuracy is low, and VariantSpark does not provide support for classifying individuals based on genotypic information.

Research also focused on genomic inferring, and ethnicity prediction, e.g., literature [29] proposed approximate Bayesian computation (ABC), which is a likelihood-free inference method based on simulating datasets and comparing their summary statistics. Although ABC's main advantages lie in its simplicity and ability to output a posterior distribution, it suffers from the 'curse of dimensionality' with decreasing accuracy and stability as the number of summary statistics grows [8]. Byun *et al.* [4], proposed a distance-based approach for BAI using principal component analysis and spatial analysis to assign individuals to population memberships.

In a previous work [30], [31], we applied K-means for the population scale clustering and achieved better accuracy than ADMIXTURE and VariantSpark. For predicting ethnicity, we trained an MLP classifier, which achieved a state-of-the-art result with high confidence. However, two limitations remained: i) the feature extraction process based on SPARQL query and converting genotype data into Resource Description Format (RDF) [32] take non-trivial time for all the chromosomes. Excellent performance was obtained for the genotype dataset for a single chromosome due to a low number of latent variables, which shows inferior results for all the chromosomes because of the overfitting and lack of generalization while training MLP model.

3 MATERIALS AND METHODS

In this section, we describe materials and methods of our approach: first, we describe our feature engineering step we followed to prepare the training data consisting of GVs features and labels. Then we chronicle network constructions for clustering and classifying population groups. Finally, we describe the training process and hyper-parameter tuning.

3.1 Problem Statement

Clustering individual's based on GVs is correlated with geographic ethnicity and bio-ancestry, where the main objective is grouping populations into clusters based on similarity, density, intervals, or particular statistical distributions measures of the data space [18]. Given GVs of n samples, $X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, where $X \in \mathbb{R}^D$. We consider clustering individuals into k-categories (i.e., k super-population or sub-population groups), each represented by a centroid $\mu_j, j = 1, \ldots, k$. On the other hand, predicting the ethnicity of an individual x_i is classifying a data point into a specific into sub-population groups based on its GVs.

However, instead of clustering or classifying samples directly in the original data space X, we first transform the data with a nonlinear mapping $f_\theta : X \to Z$, where θ are learnable parameters and $Z \in \mathbb{R}^K$ is the learned or embedded feature space, where $K \ll D$. In our approach, to parametrize f_θ, CAE architecture is employed due to it's function approximation properties and feature learning capabilities [33] by capturing local relationship values [22] in GVs with convolutional (conv) filters.

3.2 Datasets

Various genomics projects based on next Generation Sequencing technologies have emerged, including The Cancer Genome Atlas (TCGA),[2] International Cancer Genome Consortium (ICGC),[3] 1,000 genomes project[4] (1000GP), Human Genome Project (HGP),[5] Simons Genome Diversity Project (SGDP),[6] and Personal Genome Project(PGP).[7] The HGP showed that significant genetic differences exist between individuals, whereas, inspired by HGP, 1000GP seeks to measure those differences to help biomedical researchers understand the roles of GVs in health and illness. In previous studies, the 1000GP [1], [34] serves as one of the prime sources to analyze genome-wide single nucleotide polymorphisms (SNPs) at scale for predicting individual's ancestry with regards to continental and regional origins [35].

3.2.1 Data Selection

Data used in this study from the 1000GP (phase 3) and SGDP act large catalog of human GVs, where the phase 3 of 1000GP provides GVs data of 2,504 individuals from 26 populations (i.e., ethnicity) in which samples are grouped into five super-population groups according to their predominant ancestry: Europe, Africa, America, and Asia in which each of the 26 populations has about 60-100 individuals [36].

However, deletions and substitutions of less important variants (single nucleotide polymorphisms, indels, and other structural variant classes) in quality control leaves a total of 88 million variants: 84.7 million were SNPs, 3.6 million short indels, and 60,000 structural variants) identified as high-quality haplotypes [1], [34], [37]. Each genotype comprises all 23 chromosomes and a separate panel file containing samples and population information. For multi-allelic variants (e.g., Listing 1), each alternative allele frequency (AF) is calculated as the quotient of allele count and allele number (AN), where AF in the five super-population groups is calculated from the AN(range=[0,1]). Table 1 shows the distribution of the population groups in 1000GP.

Listing 1. an Example of Multi Allelic Variants in 1000GP

```
1    1 15211 rs78601809 T G 100 PASS AC=3050;
     AF=0.609026;AN=5008;NS=2504;DP=32245;
     EAS_AF=0.504;AMR_AF=0.6772;AFR_AF=0.5371;
     EUR_AF=0.7316;SAS_AF=0.6401;AA=t|||;VT=SNP
```

However, one downside of 1000GP is that it's sequencing study focused on demographically large populations, which, unfortunately, tend to ignore smaller populations that are also important for understanding human diversity [38]. To include such smaller populations, GVs from the SGDP are used, which contains deep genome sequences of 279 individuals from 130 populations chosen to span much of human genetic, linguistic, and cultural variation, covering: 44 Africans, 22 Native Americans, 27 Central Asians or Siberians, 47 East Asians, 25 Oceanians, 39 South Asians, and 75 West Eurasians. The 1000GP[8] and SGDP data[9] are publicly available in

---

2. *TCGA*: https://cancergenome.nih.gov/
3. *ICGC*: https://dcc.icgc.org/
4. *1000GP*: https://www.internationalgenome.org/
5. *HGP*: http://humangenes.org/
6. *SGDP*: http://reichdata.hms.harvard.edu/sgdp/
7. *PGP*: http://www.personalgenomes.org/
8. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
9. https://reichdata.hms.harvard.edu/pub/datasets/sgdp/

TABLE 1
Distribution of the Population Groups in 1000GP

Super population group	Acronym	Covering ethnicity	Sample count
East Asia	EAS	CHB, JPT, CHS, CDX, KHV	617
European	EUR	CEU, TSI, FIN, GBR, IBS	669
African	AFR	YRI, LWK, GWD, MSL, ESN, ASW, ACB	1,018
American	AMR	MXL, PUR, CLM, PEL	535
South Asian	SAS	GIH, PJL, BEB, STU, ITU	661

TABLE 2
A Snapshot From the Panel File From 1000GP

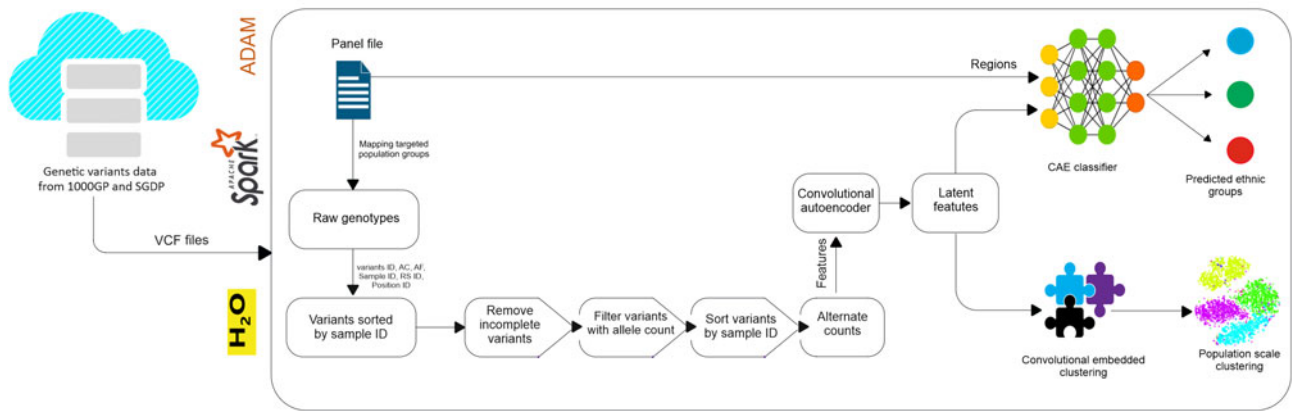| Sample ID	Population group	Ethnicity	Super population group	Gender
HG00096	GBR	British in England and Scotland	EUR	Male
HG00171	FIN	Finnish in Finland	EUR	Female
HG00472	CHS	Southern Han Chinese	EAS	Male
HG00551	PUR	Puerto Ricans from Puerto Rico	AMR	Female

Fig. 1. Processing pipeline of the proposed approach for for population-scale clustering and bio-ancestry inferencing.

variant call format (VCF). Additionally, population region for each sample is provided.

3.2.2 Population Stratification

Since majority of the variants are SNPs and INDELs [8], computing likelihood of complex population genetic models is often infeasible from the multiple individuals. Further, population stratification is necessary to identify the presence of a systematic difference in AF between sub-populations in a population, possibly due to different ancestry. First, we process the panel file containing sample IDs of the individuals, population group, ethnicity, super population groups, and the gender info (Table 2) and extract only the targeted population data which identify the population groups. Then we convert the GVs to common genotype object, followed by another round of filtering to extract data for the relevant individuals and super population groups only. Genotype objects are then converted into a sample variant object containing the following genotypic information:

- *Sample ID:* to uniquely identify a sample
- *Variant ID:* to uniquely identify a genetic variant
- *Alternate allele count:* count of alternate alleles (AA) in which the sample differs from the reference genome.

Furthermore, since ADMIXTURE's underlying statistical model does not take linkage disequilibrium (LD) into account, we remove variants with high LD and incomplete variants, assuming they are outliers [39]. Moreover, since 1000GP phase 3 contains overlapping and duplicate sites, we ignored duplicate sites in any analysis. The total number of sample (i.e., individual) is then determined, before grouping them using their variant IDs, and filtering out variants without support by the samples. Then we group variants by sample ID and sort them for each sample consistently using variant IDs, which gives us a sparse training data consisting of sample ID, variant ID, position ID, RS ID, and AA count, where a row represents an individual, a column represents a specific variant, and the "Region" column signifies labels.

3.3 Network Constructions and Training

In our approach, we model two tasks into a single pipeline, which has three stages as shown in Fig. 1: i) representation learning (RL) based on CAE, ii) population scale clustering using CEC in which we jointly optimize *clustering* and *non-clustering* losses, iii) ethnicity prediction using a CAE classifier. Non-clustering loss (NCL) is independent of the clustering algorithm and usually enforces a desired constraint
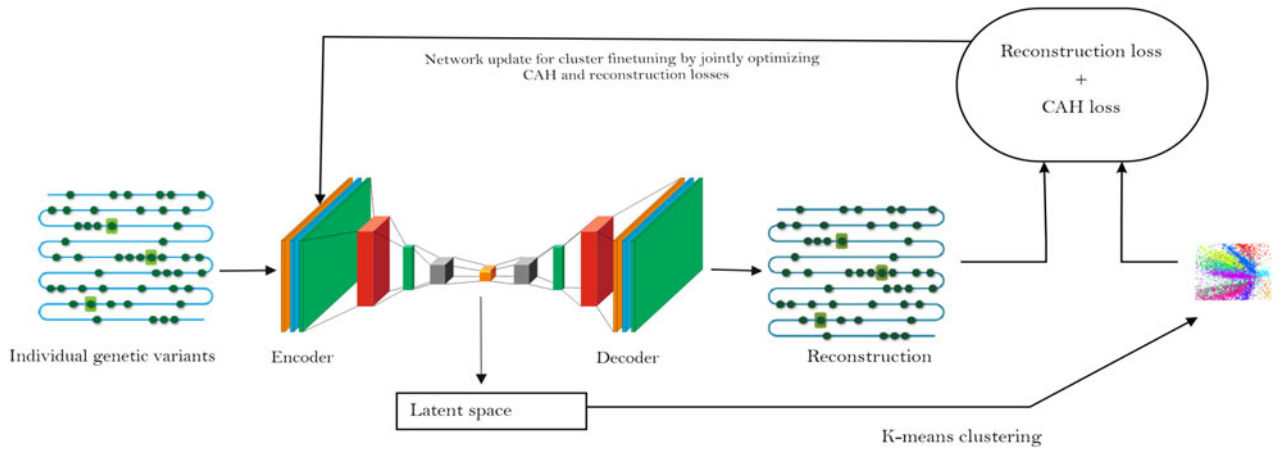
Fig. 2. Improving deep embedding clustering based on CAE and optimizing both reconstruction and CAH losses jointly. Less bright GVs (right side) means reconstruction errors exist.

on the learned model [40]. NCL also guarantees the learned representation to preserve spatial relationships between GVs so the original input can be reconstructed in the decoding phase [23], while the clustering loss is specific to the clustering method and clustering-friendliness of the learned representations [40].

3.3.1 Construction of Convolutional Autoencoder

Inspired by literature [18], we learn the cluster and classification friendly representations of samples by employing CAE: first, to apply conv operations, we embed extracted GVs of each sample into a 2D image inspired by literature [20] in which each sample is reshaped from a 4,238x1 array into a 66 x 66 image by adding zero padding around the edges and normalized each image to [0,255]. Instead of manually engineered conv filters in a CNN, we constructed the CAE by adding conv and pooling layers. The CAE consists of an encoder that performs convolution and pooling operations and a decoder that performs unpooling and deconvolution (deconv) operations. From the given GVs, the first conv layer in the encoder calculates jth feature map (FM) as follows [41]:

$$h^j = \sigma\Big(x_i * W_{ij}^j + b^j\Big), \qquad (1)$$

where x_i is the input sample, W_{ij}^j is the jth filter from input channel i and filter j, b^j is the bias for the jth filter, i.e., single bias per latent map,[10] σ is rectified linear unit (ReLu) activation function, and $*$ denotes the conv operation. To obtain the translation-invariant representations, max-pooling is performed by downsampling conv layer's output and taking the maximum value in each $m \times n$ non-overlapping subregion [41]. In the decoding phase, unpooling and deconv operations are performed to preserve the positional-invariance information during the pooling operations. The deconv operation is performed to reconstruct x_i as follows [41]:

$$x_i = \sigma\Big(o^j * W_{oj}^j + c^j\Big), \qquad (2)$$

where o^j is jth FM and W_{oj}^j is jth filter of unpooling layer o; j and c^j are filter and bias of jth output layer. Hence, CAE

10. One bias per GV would introduce many degrees of freedom

learns optimal filters by minimizing the RL1, which is the distance measure d_{CAE} between input x_i and its corresponding reconstruction $f(x_i)$ [18]

$$L_{CAE} = d_{CAE}(x_i, f(x_i)) = \sum_i ||x_i - f(x_i)||^2. \qquad (3)$$

Eventually, CAE-based RL results more abstract features that help to stabilize training and network converges faster, avoid corruption in feature space, and improve clustering quality [18]. The architecture of CAE consist of a 20-layer network in which batch normalization layer is used after every conv layer and the ReLu activation is used in every layer except for the last layer where softmax activation is used. In particular, the CAE part has the following structure:

- *Input layer*: genetic variants of each sample reduced from 4,238 × 1 to 66 × 66 × 1
- *Convolutional layer*: of size 32 × 32 × 32
- *Batch normalization layer*: of size 32 × 32 × 32
- *Convolutional layer*: of size 16 × 16 × 64
- *Batch normalization layer*: of size 16 × 16 × 64
- *Max-pooling layer*: of size 2 × 2
- *Convolutional layer*: of size 8 × 8 × 128
- *Batch normalization layer*: of size 8 × 8 × 128
- *Convolutional layer*: of size 4 × 4 × 256
- *Batch normalization layer*: of size 4 × 4 × 256
- *Max-pooling layer*: of size 2 × 2
- *Upsampling layer*: of size 2 × 2
- *Deconvolutional layer*: of size 4 × 4 × 256
- *Batch normalization layer*: of size 4 × 4 × 256
- *Deconvolutional layer*: of size 8 × 8 × 128
- *Batch normalization layer*: of size 8 × 8 × 128
- *Deconvolutional layer*: of size 16 × 16 × 64
- *Batch normalization layer*: of size 16 × 16 × 64
- *Upsampling layer*: of size 2 × 2
- *Deconvolutional layer*: of size 32 × 32 × 32.

3.3.2 Construction and Training of CEC Network

As shown in Fig. 2, architecturally, CEC is an improved variant of DEC in which405we employed CAE instead of vanilla AE and denoising AE. The CEC is trained in two phases: i) parameter initialization with a CAE (see

Section 3.3.1) and trained by optimizing the standard RL1, ii) parameter optimization by iterating between computing an auxiliary target distribution and minimizing the Kullback-Leibler divergence (KLD) [42] and *cluster assignment hardening loss* (CAHL) loss in which the cluster assignment is formulated, followed by the centroid updated with backpropagation.

The RL1 guarantees the learned representation to preserve important information (e.g., spatial relationships between features) so that the original input can be reconstructed [23]. Once the RL1 is optimized, latent features (LF) Z are extracted from the encoder, followed by normalizing them such that $\frac{1}{d}||z_i||_2^2$ is approximately 1, where d is dimension of the feature space $\{z_i \in Z\}$. Next, from an initial estimate of the non-linear mapping f_θ and centroids $\{\mu_j \in Z\}_{j=1}^K$ (as trainable weights Z), we improve the clustering by alternating between two steps which we repeat until a convergence criterion is met [43]:

- *Step 1:* soft assignment of Z to the cluster centroids.
- *Step 2:* updating the mapping f_θ and refining cluster centroids by learning from initial assignments using an auxiliary target distribution.

Initializing clustering on LF generates second type of loss called CAHL, which is specific to clustering method and clustering-friendliness of the learned representations [40]. Similar to literature [43], we consider normalized similarities between data points and centroids as the soft cluster assignments in which the Student t-distribution [44] is used as the kernel to measure the similarity between embedded point z_j and centroid μ_j, where $z_i = f_\theta(x_i) \in Z$ corresponds to $x_i \in X$ after embedding, α is the degree of freedom, and q_{ij} is the probability of assigning sample i to cluster j [33].

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + ||z_i - \mu_{j'}||^2/\alpha)^{-\frac{\alpha+1}{2}}}. \tag{4}$$

However, cross-validation of α in the unsupervised setting is not a viable option. Moreover, learning α is superfluous, similar to literature [43], we set α to 1. In step 2, similarity between the distributions is evaluated using KLD w.r.t. by decreasing the distance between soft assignments (q_{ij}) and the auxiliary distribution (p_{ij}) as follows [33]:

$$L_{KLD} = \text{KL}(P||Q) = \sum_i \sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}, \tag{5}$$

where $q_{ij} \in Q$ and $p_{ij} \in P$ are optimized through backpropagation. Minimizing this loss w.r.t. network parameters leads to smaller distances between the data points and their assigned cluster center for a better CQ, where the loss is computed by favoring a situation where points of a cluster are close to the mean of the cluster. Conversely, points that are close to the mean of another cluster will adversely affect the loss. However, since setting P is crucial to increase the CQ, similar to [43], the soft assignment q_{ij} is computed by raising auxiliary distribution p_{ij} to the second power and normalizing by frequency per cluster as follows:

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}}, \tag{6}$$

where $f_j = \sum_i q_{ij}$ are soft cluster frequencies and P forces the assignments to have stricter probabilities between [0–1]. On the other hand, since the constraints enforced by the RL1 can be lost after training the network longer, using only clustering loss may lead to worse clustering results [40]. To tackle this issue and similar to literature [23], [45], [46], we performed joint training by setting α such that the network training is affected by both clustering and non-clustering loss functions simultaneously in which combining them with a linear combination of individual loss is priory

$$L(\delta) = \sigma L_{KLD}(\delta) + (1 - \sigma)L_{CAE}(\sigma), \tag{7}$$

where $L_{KLD}(\delta)$ is the clustering loss, $L_{CAE}(\sigma)$ is the non-clustering loss, and $\sigma \in [0, 1]$ is a constant hyperparameter to specify the weighting between both functions. We optimize $L(\delta)$ using the first-order gradient-based AdaGrad with varying learning rates and different batch size, where gradients of L (w.r.t Z) for each data point z_i and cluster centroid μ_j are computed as follows [43]:

$$\frac{\partial L}{\partial z_i} = \frac{\alpha+1}{\alpha}\sum_j \left(1 + \frac{||z_i - \mu_j||^2}{\alpha}\right)^{-1} \tag{8}$$

$$\times (p_{ij} - q_{ij})(z_i - \mu_j)$$

$$\frac{\partial L}{\partial \mu_j} = -\frac{\alpha+1}{\alpha}\sum_i \left(1 + \frac{||z_i - \mu_j||^2}{\alpha}\right)^{-1} \tag{9}$$

$$\times (p_{ij} - q_{ij})(z_i - \mu_j).$$

where the gradients $\partial L/\partial z_i$ are used in standard backpropagation to compute network's parameter gradient $\partial L/\partial \theta$. This iterative process continues until less than $tol\%$ of points change cluster assignment between two consecutive iterations for the cluster assignments [31].

3.3.3 Training of CAE Classifier

As shown in Fig. 3, the CAE classifier has two module: autoencoder and classifier. As discussed in the previous section, after training the CAE, we remove the decoder components by making the first 20 layers trainable false, since the encoder part is already trained. 80 percent of the LF is used for the training the CAE classifier and 30 percent as the held-out test set. The training LF vector is fed into a flattening layer, followed by a dense, dropout, and Softmax layers (with output unit of 26 for 1000GP and 7 for SGDP) for the probability distribution over the classes.

Gaussian noise layer was also added followed by each conv, dense, and dropout (with a high probability value) to improve the model generalization by reducing overfitting. Moreover, to improve the classification results, we applied batch normalization and kept the adaptive rate. Weights are then updated using backpropagation, whereas the AdaGrad optimizer is used to optimize the categorical cross-entropy loss between predicted (P) versus true sub-population groups (T) in a random-search and 5-fold cross validation setting for finding optimal hyperparameters
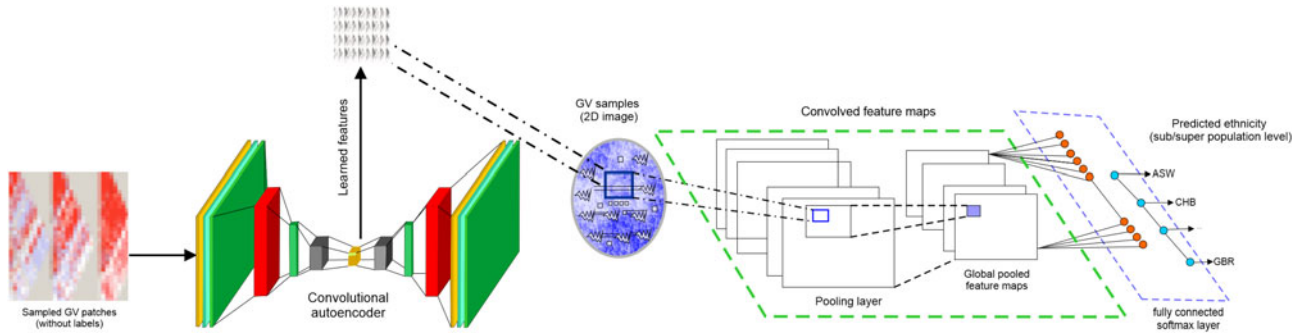
Fig. 3. Schematic representation of CAE classifier starts from taking input GVs and passing to CAE before getting the latent representation from the encoder to pass to dense, dropout, and softmax layer for the super (or sub) population prediction.

$$E = \sum_{i,j} T_{i,j} \log P_{i,j} + \left(1 - T_{i,j}\right) \log \left(1 - P_{i,j}\right). \tag{10}$$

When the training is completed, the model is used to score against the test set to measure predicted population groups versus GVs producing in a multiclass setting.

3.4 Formulating Performance Metrics

We used three empirical measures: Elbow [47], generalizability G, and normalized mutual information (NMI) [48]. In Elbow, we calculate the cost using within-cluster sum of squares (WCSS) as a function of the number of clusters, K. Since Elbow performs better in a classical clustering setting [31], NMI is used for evaluating clustering results with different cluster numbers [33], which tells us the reduction in the entropy of class labels, computed as follows:

$$NMI(y, c) = \frac{I(y, c)}{\frac{1}{2}[H(y) + H(c)]}, \tag{11}$$

where y signifies the ground-truth labels, c is the cluster assignment, I is the mutual information between y and c, and $H(.)$ is the entropy. On the other hand, G is the ratio between training and validation loss [33], in which G is small when training loss is lower than the validation loss, an indication of high degree of overfitting.

$$G = \frac{L_{train}}{L_{validation}}. \tag{12}$$

Since a good clustering accuracy also characterized by high intra-cluster similarity and low inter-cluster similarity for the data points. Accordingly, rand index (RI) is calculated based on the permutation model (PM) as follows [18]:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{13}$$

where TP, TN, FP, and FN signify true positive, true negative, false positive, and false-negative rates. RI has a value between 0 and 1, where 0 indicates the disagreement between two data clusters on any pair of points, and 1 signifies the perfect agreement (i.e., the same cluster). In our approach, normalized RI (ARI), which ranges between -1 (independent labeling) and 1 (perfect match) is used [49]. Further, to evaluate the CQ, unsupervised clustering accuracy (ACC) [33]

metric is used, which takes a cluster assignment from a base clustering algorithm, assigns the ground truths, and computes the best match between them. From the ground ground-truth label y_i and the cluster assignment c_i, ACC is computed as follows:

$$ACC = max_m \frac{\sum_{i=1}^{n} 1\left\{y_i = m(c_i)\right\}}{n}, \tag{14}$$

where m ranges overall possible one-to-one mappings between clusters and labels using Hungarian algorithm [50]. Further, since ground truths are available, homogeneity and completeness are formulated to desirable objectives for the cluster assignment using conditional entropy analysis [51]. While, the former signifies if each cluster contains only members of a single class, the latter tells us if all members of a given class are assigned to the same cluster.

4 EXPERIMENT RESULTS

We discuss the evaluation results, both quantitatively and qualitatively. Besides, a comparative analysis with state-of-the-art approaches is provided.

4.1 Experiment Setup

All the programs[11] were written in Python and experimented on a computer having 32 cores, 256 GB of RAM, and Debian 9.9, while the software stack consisted of Keras and scikit-learn with the TensorFlow backend. First, to achieve massive scalability while processing genotype data across all the chromosomes, we used ADAM,[12] while sparkling water[13] is used to transform data between ADAM and Spark. Network training is then carried out on an Nvidia Titan Xp GPU with CUDA and cuDNN enabled.

Results based on best hyperparameters produced through random search are reported empirically, where we verified whether the network converges to the optimal number of clusters by setting $K = 2$ and increasing it slowly. We also focused on investigating how the network training converged during the cluster assignments and updates by utilizing the Elbow method in which WCSS is calculated. Besides,

---

11. https://github.com/rezacsedu/Convolutional-embedded-networks
12. https://github.com/bigdatagenomics/adam
13. https://www.h2o.ai/sparkling-water/

TABLE 3
Confusion Matrix of CAE Classifier

Sub population	ASW	CHB	CLM	FIN	GBR	Support
ASW	56	0	3	2	0	5/61
CHB	2	98	0	3	0	5/103
CLM	1	0	88	4	1	6/94
FIN	1	2	91	3	2	8/99
GBR	2	1	1	4	83	8/91
Total	62	103	174	16	85	32/448

other metrics such as ARI, NMI, ACC, completeness, and homogeneity. On the other hand, macro-averaged precision, recall, F1-score, and Matthias correlation coefficient (MCC) are reported in the multiclass setting.

4.2 Ethnicity Prediction and Inferencing Analysis

We evaluated performance of the trained models in 3 folds: first, we extracted GVs of selected sub-population groups (i.e., 'ASW', 'CHB', 'CLM', 'FIN', 'GBR') from chromosome 22 data, which gives 494,328 alleles allowing 5 class and compare the actual labels to the same number of predicted ethnicity labels. This random sample selection provides a good classification accuracy (i.e., 92.86 percent), as shown in the confusion matrix in Table 3. The full test set is then used to evaluate the model by measuring the prediction performance for sub-populations at the boundaries of the super population, giving precision, recall, F1,

TABLE 4
Class-Specific Performance of Ethnicity Prediction

| Sub-population	Precision	Recall	F1	MCC
CHB	0.8153	0.8015	0.8083	0.7235
JPT	0.9037	0.8976	0.90.07	0.8256
CHS	0.8233	0.8175	0.8204	0.7421
CDX	0.8025	0.7945	0.7985	0.7043
KHV	0.8643	0.8525	0.8583	0.8046
CEU	0.8583	0.8311	0.8445	0.7967
TSI	0.8826	0.8782	0.8804	0.8024
FIN	0.9235	0.9169	0.9202	0.8369
GBR	0.8943	0.9029	0.8886	0.8162
IBS	0.8224	0.8173	0.8242	0.7121
YRI	0.8381	0.8456	0.8419	0.7235
LWK	0.8967	0.9123	0.9044	0.8531
GWD	0.8194	0.8085	0.8140	0.7891
MSL	0.8368	0.8245	0.8306	0.7951
ESN	0.8785	0.8643	0.8714	0.8097
ASW	0.8954	0.8832	0.8893	0.8430
ACB	0.8753	0.8671	0.8712	0.8421
MXL	0.8825	0.8733	0.8779	0.8262
PUR	0.8913	0.8719	0.8815	0.8076
CLM	0.8537	0.8611	0.8574	0.8735
PEL	0.9629	0.9567	0.9598	0.8525
GIH	0.8736	0.8722	0.8729	0.8134
PJL	0.8952	0.8845	0.8898	0.8236
BEB	0.9255	0.9123	0.9188	0.8374
STU	0.8795	0.8857	0.8826	0.8212
ITU	0.8697	0.8567	0.8632	0.8513
Average	**0.9025**	**0.8983**	**0.9004**	**0.8245**
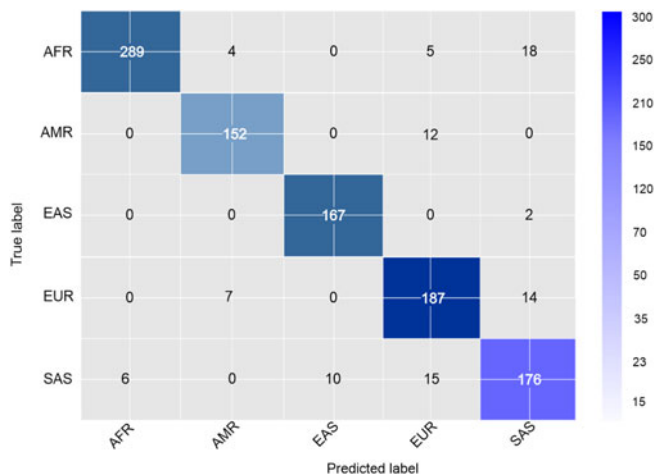
Fig. 4. Confusion matrix where predictions are for sub populations at the boundaries of the super population.

and MCC scores of 0.9025, 0.8983, 0.9004, and 0.8245, respectively.

However, since classes are imbalanced, accuracy gives a distorted estimation of the sub-populations. Hence, class-specific classification reports and MCC scores are reported in Table 4. As shown, precision, recall, and f1 for the majority of sub-populations groups are high. In particular, the CAE classifier classifies JPT, KHV, CEU, TSI, FIN, GBR, IBS, LWK, ASW, ACB, MXL, PUR, CLM, PEL, GIH, PJL, BEB, STU, and ITU samples mostly correctly. However, for the Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), and Chinese Dai in Xishuangbanna (CDX) populations, the CAE classifier made considerable misclassification, giving lower scores, probably because of similar GVs across those samples. Although African populations were mostly classified correctly, in the case of YRI (Yoruba in Nigeria), GWD (Gambian in Gambia), MSL (Mende in Sierra Leone), and ESN (Esan in Nigeria), a fairly high misclassification error is observed.

Even though, CAE classifier confused between AFR, SAS, AMR, and EUR samples (as shown in Fig. 4), this is still considerably low compared to literature [30]. The reason for the improvement is that all the variants with high LD and incomplete variants were removed in our preprocessing step, which has contributed towards non-corrupted latent features. The reason is simple with that minor factor, and we removed some impurities giving the network more quality features, which eventually helped in separating data points. On the other hand, class-specific MCC scores of the CAE classifier suggests that predictions were strongly correlated with the ground truth, yielding a Pearson product-moment correlation coefficient higher than 0.70 for the majority sub-population groups. Besides, the ROC curves in Fig. 5 show consistent AUC scores across folds for both datasets, which signifies that the predictions by the CAE classifier in both cases are much better than random guessing.

Finally, we trained the CAE classifier on the SGDP dataset for predicting unknown ethinicity in which similar data processing steps were followed. We experienced moderately lower predictive accuracy: out of 44 Africans, only 35 samples were predicted correctly, and the rest of the samples were predicted as South Asian. However, the CAE model
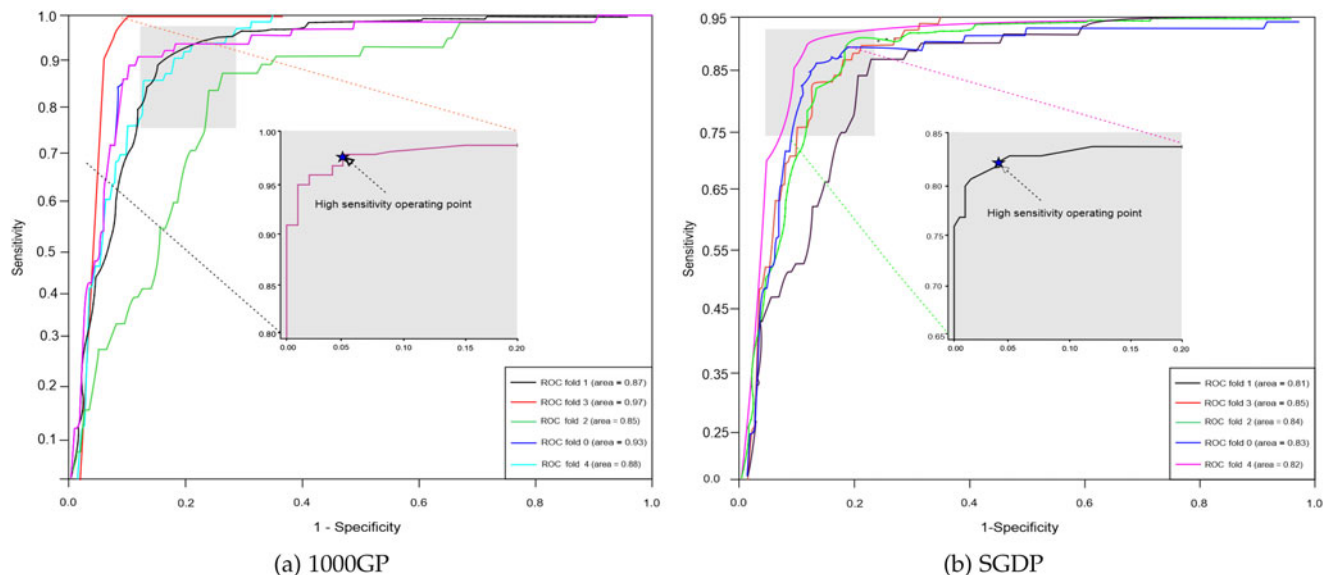
Fig. 5. ROC curves for the CAE classifier on 100GP and SGDP datasets.

shows high confidence at predicting Native Americans samples in which 20 samples out 22 were correctly predicted, and only 2 samples were predicted as West Eurasians; in the case of Central Asians or Siberians, 20 samples were correctly predicted, and 7 samples were predicted as West Eurasians, while out of 47 East Asians samples, 39 were correctly predicted, 6 samples were predicted as Africans, and 2 samples were predicted as South Asians.

In the case of Oceanians samples, 22 samples were correctly predicted out of 25, and only 3 samples were predicted as Native Americans. Besides, out of 39 South Asians samples, 29 were correctly predicted, and 10 samples were predicted as Africans, showing considerably high misclassification rates. Out of 75 West Eurasians samples, only 65 were correctly predicted, 8 were predicted as Central Asians or Siberians, and 2 were predicted as Native Americans. Overall, 211 samples out of 279 were correctly classified giving an F1-score of 0.83 in which CAE classifier was more confused between South Asian and Africans, Central Asians or Siberians and West Eurasians, and between East Asian and Africans, which is probably because GVs from these two groups are mostly mixed and share common alleles. A depth genomic analysis is further required to explain these.

4.3 Population Scale Cluster Analysis

Clustering results are reported in Table 5 with different metrics in which the AC algorithm performed the best clustering with optimal hyperparameters on CAE-based LF (highlighted

TABLE 5
Clustering With Autoencoders + Base Clustering
Algorithms (Hom*==Homogeneity, Com*==Completeness)

Clustering algorithm	ARI	NMI	ACC	Hom*	Com*
K-means	0.853	0.863	0.838	0.827	0.814
AC	**0.915**	**0.927**	**0.896**	**0.867**	**0.853**
GMM	0.814	0.821	0.792	0.775	0.776
DBSCAN	0.885	0.891	0.872	0.846	0.836
OPTICS	0.831	0.843	0.815	0.813	0.805

in cyan): we observed an ARI, NMI, and ACC of 0.915, 0.927, and 0.896, respectively in which each cluster contains only members of a single class in 86.7 percent of the cases (homogeneity) and in 85.3 percent of the cases all members of a given class are assigned to the same cluster (completeness). The reason is that CAE learned LF that are more quality ones than raw GVs data, as shown in Fig. 6, which eventually tends to better separability of data points.

Besides, DBSCAN and K-means (on CAE-based on LF) also performed moderately well compared GMM and OPTICS algorithms in which DBSCAN turns out to be the second-best clustering algorithm. Contrarily, both OPTICS and GMM did not perform well, making the GMM the worst performer. On one hand, OPTICS is inherently better for sequences hence couldn't provide better separability of the data points. On the other hand, GMM doesn't have any uncertainty measure or probability w.r.t how much a data point is associated with a specific cluster.

Predicted clusters by CEC can be observed in Fig. 6d, where GVs of African (AFR), East Asian (EAS), and American (AMR) super-population groups are highly separated and grouped into distinct clusters. On the other hand, although European individuals can be mostly separated into persons of Finnish and non-Finnish ancestry [7], overall EUR is more mixed with AMR and consists predominantly of individuals, so didn't cluster well. This is probably the effect of several mixtures of GVs, e.g., migrational backgrounds and a potential reason for such low clustering accuracy. However, in cases where individuals were not clearly clustered due to diverse migrational backgrounds, the ancestry itself influences the treatment hence accurate population association may not be known for the patients even from the human leukocyte antigen (HLA) allele genotyping from SNP information [39]. Albeit, literature [39] observed an increase in cluster accuracy by removing individuals with a mixed background and operating at super-population level, and it was not the cases in the sub-population level sine a recent research [28] has shown that the apparent separation between East Asian and other samples reflects a
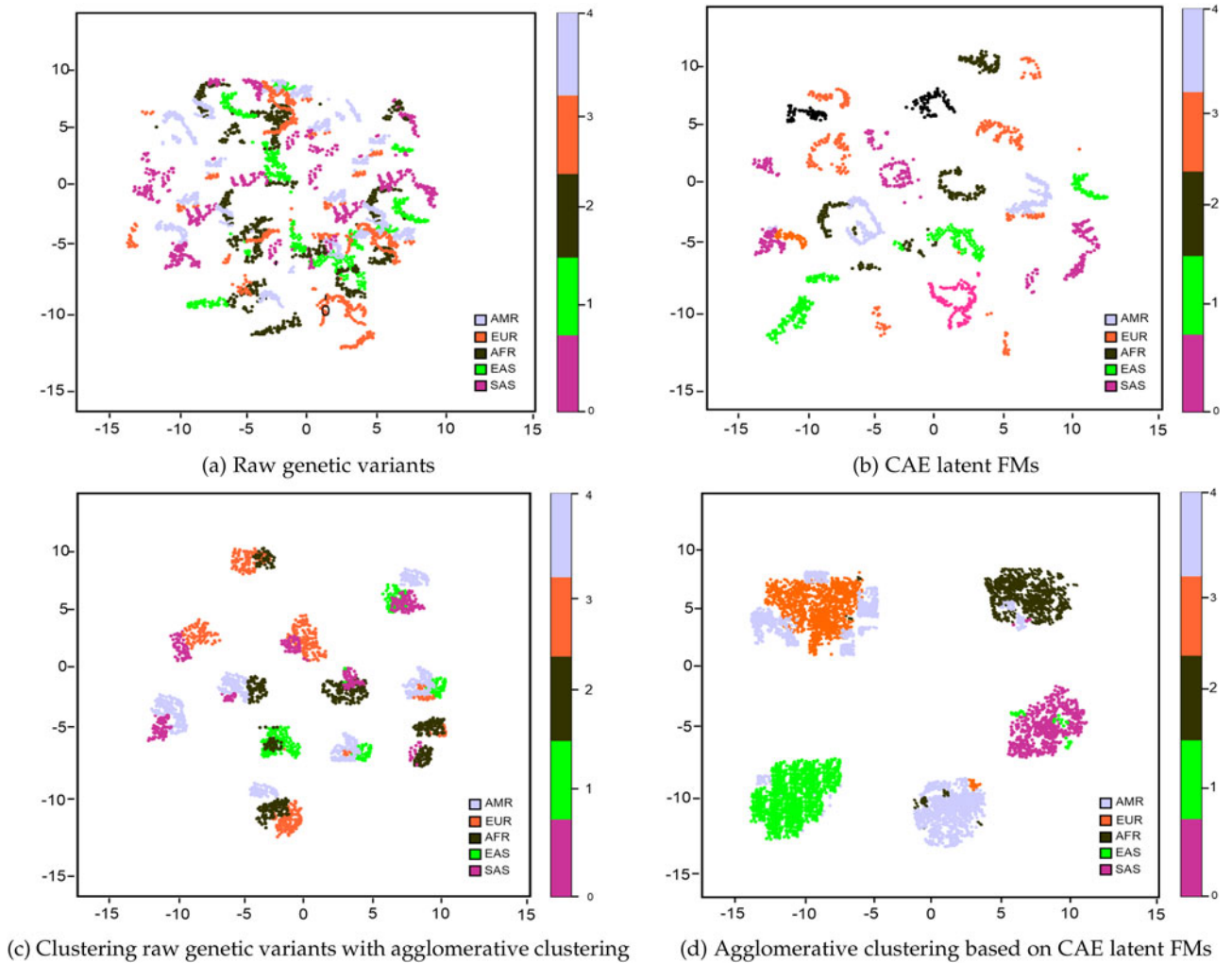
Fig. 6. t-SNE plots of different clustering stages: a) no distinctive patterns visible between super-population groups on raw GVs, b) learned latent genetic properties from GVs profiles with CAE, showing moderately high distinctive patterns between super-population groups, c) clustering super-population groups on raw GVs, d) clustering super-population groups on learned feature maps with CAE.

deficiency of Middle Eastern and Central Asian samples in the dataset.

Further, we investigate how well CEC converges to the optimal number of population groups: we started the clustering by setting K=2 (where applicable) and increased up to 35 and observe the clustering performance. We plotted NMI and G versus several clusters and found a sharp drop of generalizability for K = 26 and 27, which means that 26 is the optimal number of clusters. To support this argument, the graph also shows that for 26, we observed the highest NMI of about 0.92. Subsequently, we utilize the Elbow method in which we calculated WCSS as a function of the number of clusters (i.e., K), for which we observed a drastic fall of WCSS when several clusters were around 25 and 26 for 1000GP and 7 and 8 for the SGDP.

To compare with VariationSpark, we further analyze the clustering of five super-population groups (i.e., 'EUR', 'AMR', 'AFR', 'EAS', 'SAS') for each to the label assigned by CEC. This experiment results in an ARI of 0.87, an ACC of 0.86, and an NMI of 0.88, showing higher confidence, at least in terms of ARI (albeit, this is still low compared to our overall ARI). Finally, we investigate which super-population group contains what percentage of human GVs in which

CEC reveals an interesting statistics showing majority of the genetic variants were clustered into EUR (28.32 percent) and lowest into AMR (12.68 percent), while the distributions of samples from EAS, AFR, and SAS super-population groups were 22.25, 18.65, and 18.10 percent, respectively.

4.4 Qualitative Study of the Learned Representations

Since GVs data are high-dimensional, learning the association between each feature was fairly considered. Inspired by literature [20], [21], [52] and to qualitatively study whether the learned representation can express the biological characteristics of the individuals, t-SNE of the CAE encoder's output, i.e., latent FM and raw GVs are plotted in Fig. 6. From t-SNE plots, we can observe moderately high distinctive patterns between 4 super-population groups. However, not all these patterns clearly visible in the t-SNE plot of raw GVs, which signifies how CAE learned latent genetic properties better from the GVs profiles.

4.5 Shapley Value-Based Explanations

Since CAE learns the representations are not easily interpretable, interpretable and disentangled representations are

Fig. 7. GVs feature contribution for the first prediction: red pushes the predictions higher, blue ones to lower.

essential to provide insight into what features did the representations capture and what attributes of the samples are the clusters based on. To provide interpretations of the predictions, we identify significant GV biomarkers using GBT in which SHAP generates explanations. Shapley values are used to calculate the importance of a feature by comparing what a model predicts with and without a feature from all possible combinations of n features in the dataset S, where the prediction p of the model for a given feature $i \in S$ is generated w.r.t i, for which the Shapely value ϕ is calculated as follows [53]:

$$\phi_i(p) = \sum_{S \subseteq N/i} \frac{|S|!(n - |S| - 1)!}{n!} (p(S \cup i) - p(S)). \quad (15)$$

However, since the order in which a model sees features can affect the predictions, this computation is repeated in all possible orders to compare the features fairly [24]. The feature that does not modify the predicted value is expected to produce a Shapley value of 0. However, if two features contribute equally to the prediction, Shapley values should be the same [53]. The base value indicating the directions of the first prediction made by the GBT model is Fig. 7 in which how much each feature is pushing model's output from the base value[14] to the predicted output is shown. Features pushing the prediction higher are shown in red, those pushing to lower are in blue.

In Fig. 8, top-15 common biomarkers are sorted by the sum of SHAP value magnitudes over all the samples and are ordered according to their importance, where the color represents the value of the feature from low to high, i.e., red represents high feature values, and blue represents low feature values. Overlapping points are jittered in the y-axis signifying the distribution of the Shapley values per feature, i.e., delivery of the impact of each feature on the model output. However, since all effects describe the overall behavior of the model and are not necessarily causal, we provide additional annotations for these biomarkers in Table 6. As seen, the majority of the variants are multi-allele SNPs.

4.6 Discussion and Comparative Analysis

Based on optimal K and other hyperparameters, CEC completes clustering in 22 hours with an ARI of 0.915, an NMI of 0.927 and ACC of 0.896 as shown in Table 7, while VariationSpark requires 30h to finish the overall computation, leveraging an ARI of 0.82 only.[15]

ADMIXTURE performs clustering based on the maximum likelihood estimation of individual ancestries and

multi-locus SNP genotypes. Overall processing time is considerably high (takes 35h), giving an ARI of only 0.25 [16]. Following are some potential reasons behind low clustering accuracy of VariationSpark and ADMIXTURE:

- Being a K-means-based approach, VariationSpark has several limitations. K-means is based on the assumption that each cluster is equal-sized, where clusters to have hyper-sphere shapes [18], which is not the case for both 1000GP and SGDP datasets. Further, since K-means is sensitive to noise and outliers thus was probably trapped in a local optimum during the clustering operations. Nevertheless, we observed that the clustering results were slightly different for a different initial value of K since it's the only hyperparameter.
- Being an MLE-based approach, ADMIXTURE is limited to accurately estimate population mean and standard deviation [54] in case of multi-locus SNP genotypes.

Our study also investigated what percentage of the cases for each cluster contains members of a single class and what percentage of the cases for all members of a given class are assigned to the same cluster, giving 86.7 of homogeneity and 85.3 percent of the completeness of the clustering. Further, the in-memory caching mechanism of ADAM and
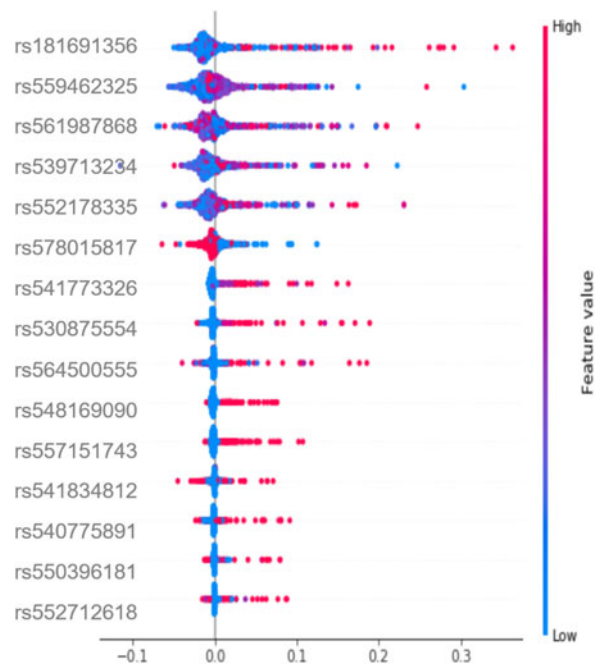

Fig. 8. Clinical features ordered by ascending importance on the y-axis (dots represent SHAP values of specific features).

---

14. The average model output over the training dataset passed
15. VariationSpark and ADMIXTURE did not report NMI, ACC, homogeneity, and completeness

TABLE 6
Annotations of Top-15 Variants

Variant ID	Variant type	Alleles	Allele frequency
rs181691356	SNP	C > A,T	A = 0.0008/4
rs559462325	SNP	G > A	A = 0.0002/1
rs561987868	SNP	G > A,T	T = 0.0002/1
rs539713234	SNP	G > C	C = 0.0046/23
rs552178335	SNP	G > A	A = 0.0002/1
rs578015817	SNP	G > A	A = 0.0002/1
rs541773326	SNP	C > G,T	T = 0.0002/1
rs530875554	SNP	C > A,T	A = 0.001/5
rs564500555	SNP	A > C,G	G = 0.0074/37
rs548169090	SNP	G > A	A = 0.0002/1
rs557151743	SNP	T > C	C = 0.0004/2
rs541834812	SNP	G > A	A = 0.0002/1
rs540775891	SNP	C > A,T	C = 0.0002/1
rs550396181	SNP	T > C	C = 0.0004/2
rs552712618	SNP	A > G,T	T = 0.0006/3

TABLE 7
Clustering Result Comparison Between Different
Approaches (Hom*==Homogeneity, Com*==Completeness)

Approach	ARI	NMI	ACC	Hom*	Com*	Time
CEC	0.915	0.927	0.896	0.867	0.853	22h
VariantSpark	0.82	-	-	-	-	30h
ADMIXTURE	0.25	-	-	-	-	35h

natural language. In the future, we intend to extend this work by providing: i) a more detailed analysis of intra-super-population groups and discuss the homogeneity and heterogeneity among different groups, ii) considering other datasets and factors like predicting population groups within larger geographic continents, iii) exploring if we can make share representations of the features out of both 1000 genomes and PGP datasets and cluster them simultaneously using CEC, iv) by generating decision rules to provide more human-interpretable explanations using neuro-symbolic reasoning.

Spark while processing VCF files made our processing pipeline 32 and 90 percent faster compared to VariantSpark and ADMIXTURE, respectively. Nevertheless, VariationSpark and ADMIXTURE are both black-box methods, whereas we explain the prediction made by models showing class and cluster discriminating features.

5 CONCLUSION AND OUTLOOK

In this paper, we implemented two powerful architecture called CEC and CAE classifiers for the clustering population and predicting genomic ancestry based on GVs of about 3,000 individuals from the 1000GP and SGDP. Our Spark and ADAM based data processing is particularly suitable for handling large-scale genomic data. Experiment results with a focus on accuracy and scalability show that our approach outperforms state-of-the-art approaches such as ADMIXTURE and VariantSpark. Our approach can perform clustering on VCF files from 2,504 individuals consist of 84 million GVs in just 22h, allowing faster clustering for well-characterized cohorts, where 20 percent of the genome is sufficient for the training. CEC can cluster the whole population by jointly optimized feature space with an ACC of 89 percent, which can be viewed as an unsupervised extension of semi-supervised self-training. Similar to [33], CEC has linear complexity concerning several data points, which allowed us to scale to large datasets, whereas the CAE classifier can predict the ethnicity of unknown samples with an F1-score of 93 percent, which is consistent with all the genotypic dataset from 23 chromosomes, giving high-level of confidence.

We have seen that explaining the predictions with plots and charts are useful for exploration and discovery but interpreting them for the first time may be tricky, e.g., suppose, the CAE classifier predicts (or groups) a selected sample is of FIN population (or into EUR super-population group) in which model's average response to the dataset is 0.6 and the model predicts that the selected sample's bio-ancestry with probability 0.75 by showing all the variables that have contributed to that prediction but still interpreting these for the first time may be difficult and need more human-interpretable decision rules in

REFERENCES

[1] D. M. Altshuler et al., "An integrated map of genetic variation from 1,092 human genomes," Nature, vol. 491, no. 7422, pp. 56–65, 2012.

[2] L. Bomba, K. Walter, and N. Soranzo, "The impact of rare and low-frequency genetic variants in common disease," Genome Biol., vol. 18, no. 1, 2017, Art. no. 77.

[3] H. Behravan et al., "Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls," Sci. Rep., vol. 8, no. 1, 2018, Art. no. 13149.

[4] J. Byun, Y. Han, and C. I. Amos, "Ancestry inference using principal component analysis and spatial analysis: A distance-based analysis to account for population substructure," BMC Genomics, vol. 18, no. 1, 2017, Art. no. 789.

[5] M. P. Miller and S. Kumar, "Understanding human disease mutations through the use of interspecific genetic variation," Hum. Mol. Genet., vol. 10, no. 21, pp. 2319–2328, 2001.

[6] Y. Laitman, B.-J. Feng, I. M. Zamir, J. N. Weitzel, E. Duncan, and A. Latif, "Haplotype analysis of the 185delAG BRCA1 mutation in ethnically diverse populations," Eur. J. Hum. Genet., vol. 21, no. 2, 2013, Art. no. 212.

[7] B. Padhukasahasram, "Inferring ancestry from population genomic data and its applications," Front. Genet., vol. 5, 2014, Art. no. 204.

[8] S. Sheehan and Y. S. Song, "Deep learning for population genetic inference," PLoS Comput. Biol., vol. 12, no. 3, 2016, Art. no. e1004845.

[9] L. B. Jorde and S. P. Wooding, "Genetic variation, classification and race," Nat. Genet., vol. 36, pp. S28–S33, 2004.

[10] Z. H. Rosser et al., "Y-chromosomal diversity in europe is clinal and influenced primarily by geography, rather than by language," The Amer. J. Hum. Genet., vol. 67, no. 6, pp. 1526–1543, 2000.

[11] D. H. Alexander, J. Novembre, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," Genome Res., vol. 19, no. 9, pp. 1655–1664, 2009.

[12] X. Gao and J. Starmer, "Human population structure detection via multil-genotype clustering," BMC Genetics, vol. 8, no. 1, 2007, Art. no. 34.

[13] B. Padhukasahasram, "Inferring ancestry from population genomic data and its applications," Front. Genet., vol. 5, 2014, Art. no. 204.

[14] J. M. Kidd, S. Gravel, S. Musharoff, and R. Chen, "Population genetic inference from personal genome data: Impact of ancestry and admixture on human genomic variation," The Amer. J. Hum. Genet., vol. 91, no. 4, pp. 660–671, 2012.

[15] E. J. Parra et al., "Estimating african american admixture proportions by use of population-specific alleles," The Amer. J. Hum. Genet., vol. 63, no. 6, pp. 1839–1851, 1998.

[16] A. R. O'Brien, N. F. Saunders, Y. Guo, F. A. Buske, R. J. Scott, and D. C. Bauer, "VariantSpark: Population scale clustering of genotype information," BMC Genomics, vol. 16, no. 1, 2015, Art. no. 1052.

[17] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2014.

[18] M. R. Karim *et al.*, "Deep learning-based clustering approaches for bioinformatics," *Brief. Bioinformatics*, 2020.

[19] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 202–208.

[20] M. R. Karim, A. Rahman, J. B. Jares, S. Decker, and O. Beyan, "A snapshot neural ensemble method for cancer prediction using copy number variations," *Neural Comput. Appl.*, pp. 1–19, 2019.

[21] M. R. Karim, G. Wicaksono, and I. G. Costa, S. Decker, and O. Beyan, "Prognostically relevant subtypes and survival prediction for breast cancer based on multimodal genomics data," *IEEE Access*, vol. 7, pp. 133850–133864, 2019.

[22] M. R. Karim, M. Cochez, J. B. Jares, M. Uddin, O. Beyan, and S. Decker, "Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network," in *Proc. 10th ACM Int. Conf. Bioinf. Comput. Biol.*, 2019, pp. 113–123.

[23] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39 501–39 514, 2018.

[24] M. Karim, M. Cochez, and O. E. A. Beyan, "OncoNetExplainer: Explainable predictions of cancer types based on gene expression data," in *Proc. 19th IEEE Int. Conf. Bioinf. Bioeng. (BIBE'2019)*, Athens, Greece, Oct. 28–30, 2019.

[25] M. E. Kaminski, "The right to explanation, explained," *Berkeley Technol. Law J.*, vol. 34, 2019, Art. no. 189.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[27] E. Han, P. Carbonetto, R. E. Curtis, Y. Wang, J. M. Granka, and M. J. Barber, "Clustering of 770,000 genomes reveals post-colonial population structure of north america," *Nat. Commun.*, vol. 8, 2017, Art. no. 14238.

[28] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, and E. A. Fennell, "Analysis of protein-coding genetic variation in 60,706 humans," 2016, *BioRxiv*.

[29] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun , and M. W. Feldman, "Population growth of human y chromosomes: A study of y chromosome microsatellites," *Mol. Biol. Evol.*, vol. 16, no. 12, pp. 1791–1798, 1999.

[30] M. R. Karim, A. Zappa, R. Sahay, and D. Rebholz-Schuhmann , "A deep learning approach to genomics data for population scale clustering and ethnicity prediction," in *Proc. Extended Semantic Web Conf.*, 2017, pp. 1–15.

[31] M. Karim *et al.*, "Recurrent deep embedding networks for genotype clustering and ethnicity prediction," 2018, *arXiv: 1805.12218*.

[32] M. R. Karim, A. Michel, A. Zappa, P. Baranov, R. Sahay, and D. Rebholz-Schuhmann, "Improving data workflow systems with cloud services and use of open data for bioinformatics research," *Brief. Bioinformatics*, vol. 19, pp. 1035–1050, 2018.

[33] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[34] G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, 2015, Art. no. 68.

[35] S. Belsare *et al.*, "Evaluating the quality of the 1000 genomes project data," *BMC Genomics*, vol. 20, no. 1, pp. 1–14, 2019.

[36] E. Birney and N. Soranzo, "Human genomics: End of start for population sequencing," *Nature*, vol. 526, no. 7571, pp. 52–53, 2015.

[37] G. Zaccone and M. R. Karim, "Deep learning with tensorflow second edition," Packt Publishing Ltd., Feb. 2018.

[38] A. Chakravarti, "Perspectives on human variation through the lens of diversity and race," *Cold Spring Harbor Perspectives Biol.*, vol. 7, no. 9, 2015, Art. no. a023358.

[39] A. R. O'Brien, N. F. Saunders, Y. Guo, F. A. Buske, R. J. Scott, and D. C. Bauer, "VariantSpark: Population scale clustering of genotype information," *BMC Genomics*, vol. 16, no. 1, 2015, Art. no. 1052.

[40] E. Aljalbout, V. Golkov, Y. Siddiqui, and D. Cremers, "Clustering with deep learning: Taxonomy and new methods," 2018, *arXiv: 1801.07648*.

[41] M. Alirezaie, M. Längkvist, M. Sioutis, and A. Loutfi, "Semantic referee: A neural-symbolic framework for enhancing geospatial semantic segmentation," 2019, *arXiv: 1904.13196*.

[42] J. M. Joyce, "Kullback-Leibler divergence," *Int. Encyclopedia Statist. Sci.*, pp. 720–722, 2011.

[43] D. H. Alexander, J. Novembre, and K. Lange, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn. (ICML'2016)*, New York City, NY, USA, June 19–24, 2016, pp. 478–487.

[44] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[45] P. Huang, Y. Huang, W. Wang, and L. Wang, "Deep embedding network for clustering," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 1532–1537.

[46] O. Kilinc and I. Uysal, "Learning latent representations in neural networks for clustering through pseudo supervision and graph-based activity regularization," 2018, *arXiv: 1802.03063*.

[47] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Statist. Soc. Series B (Statist. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.

[48] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. Dec, pp. 583–617, 2002.

[49] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[50] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Res. Logistics*, vol. 2, no. 1/2, pp. 83–97, 1955.

[51] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empir. Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 410–420.

[52] S. Rhee, S. Seo, and S. Kim, "Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification," 2017, *arXiv: 1711.05859*.

[53] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[54] R. B. Jain and R. Y. Wang, "Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection," *Analytical Chem.*, vol. 80, no. 12, pp. 4767–4772, 2008.

Md. Rezaul Karim received the MSc degree in computer science from Kyung Hee University, Korea. He is currently working toward the PhD degree at RWTH Aachen University, Germany. He is a researcher at Fraunhofer FIT, Germany. Previously, he worked as a researcher at Insight Centre for Data Analytics, Ireland and as a lead software engineer at Samsung Electronics. His research interests include machine learning, bioinformatics, and explainable artificial intelligence (XAI).

Michael cochez received the BSc degree in information technology from the University of Antwerp, Belgium, and the MSc and PhD degrees in mathematical information technology from the University of Jyväskylä, Finland. He is an assistant professor of artificial intelligence at Vrije Universiteit Amsterdam, the Netherlands. Previously, he was a postdoctoral researcher at Fraunhofer FIT, Germany. His research interests include knowledge representation, machine learning, and data mining.

Achille Zappa received the PhD degree in bioengineering from the University of Genoa, Genoa, Italy. He is a senior postdoctoral researcher at Insight Centre for Data Analytics and Data Science Institute at NUI Galway, Ireland. His research interests include semantic Web, semantic data mashup, bioinformatics, data management, and knowledge engineering.

Ratnesh Sahay is an associate director and senior information architect for Clinical Data Management at AstraZeneca, Cambridge, United Kingdom. Previously, he was the head of Semantics in eHealth and Life Sciences (SeLS) research unit at Insight Centre for Data Analytics, NUI, Galway, Ireland. He has worked on several European and national (Irish) R&D projects with an emphasis on using semantics for solving key integration/interoperability challenges in the e-health, clinical trial, and biomedical domains. The work present in this article is conducted while his employment at the the Insight Centre for Data Analytics, NUI Galway, Ireland.

Dietrich Rebholz-Schuhmann is a professor of medicine and the director of German National Library of Medicine, University of Cologne, Germany. He was formerly professor at NUI Galway, Ireland and executive director of Insight Centre for Data Analytics and previously, group leader at European Bioinformatics Institute. His research interests include semantic Web and linked data, biomedical texts mining, and bioinformatics.

Oya Beyan is a researcher at Fraunhofer FIT and RWTH Aachen University, Germany. Her research focuses on methods of data reusability and FAIR data, data-driven transformation, and distributed analytics. Her area of expertise is in the semantic Web technologies and their application in healthcare and life sciences.

Stefan Decker is the managing director of Fraunhofer FIT and professor of computer science at RWTH Aachen University, Germany. He was formerly professor and director of Insight Centre for Data Analytics, NUI Galway, Ireland. His research interests include semantic Web and linked data, and knowledge representation.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.