Neural Networks 165 (2023) 562–595

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Genetic data visualization using literature text-based neural networks: Examples associated with myocardial infarction

Jihye Moon *, Hugo F. Posada-Quintero, Ki H. Chon

Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 17 August 2022
Received in revised form 11 April 2023
Accepted 9 May 2023
Available online 19 May 2023

Keywords:
Explainable Artificial Intelligence
Natural language processing
Unsupervised learning
Cross-modal representation
Data visualization
Cardiovascular Disease risk prediction

ABSTRACT

Data visualization is critical to unraveling hidden information from complex and high-dimensional data. Interpretable visualization methods are critical, especially in the biology and medical fields, however, there are limited effective visualization methods for large genetic data. Current visualization methods are limited to lower-dimensional data and their performance suffers if there is missing data. In this study, we propose a literature-based visualization method to reduce high-dimensional data without compromising the dynamics of the single nucleotide polymorphisms (SNP) and textual interpretability. Our method is innovative because it is shown to (1) preserves both global and local structures of SNP while reducing the dimension of the data using literature text representations, and (2) enables interpretable visualizations using textual information. For performance evaluations, we examined the proposed approach to classify various classification categories including race, myocardial infarction event age groups, and sex using several machine learning models on the literature-derived SNP data. We used visualization approaches to examine clustering of data as well as quantitative performance metrics for the classification of the risk factors examined above. Our method outperformed all popular dimensionality reduction and visualization methods for both classification and visualization, and it is robust against missing and higher-dimensional data. Moreover, we found it feasible to incorporate both genetic and other risk information obtained from literature with our method.

1. Introduction

As genomic data are high-dimensional, consisting of millions of genotype markers, despite recent efforts, many of the existing data have not yet been fully elucidated (Al-Husain & Hafez, 2015). Understanding the dynamics of genotype data is critical, which leads to potential breakthroughs, especially in biology and the medical fields (Diaz-Papkovich et al., 2019). To foster better discrimination of genetic variants among vast amounts of genomic data, one such approach to reducing complexity of the dynamics is data visualization (Li et al., 2014; Yang et al., 2018). For data visualization to be informative, it is critical to capture both the overall shape (global structure) and the fine granular shapes (local structure) of the data (Thioulouse et al., 1995). Global structures are those that separate two spatial groups such as population patches or geographic clines whereas local structures represent genetic differences among neighbors of entities (i.e., disease risk) (Diaz-Papkovich et al., 2019; Jombart et al., 2008; Sakaue et al., 2020).

In recent years, genetic data visualization tools have been widely used for various applications including genetic data quality control (QC) procedures (de Bakker et al., 2008; Morris et al., 2010), population structure estimations (Diaz-Papkovich et al., 2019), and disease risk analysis (Reisberg et al., 2017; Sakaue et al., 2020). The visualization has been conducted on single nucleotide polymorphisms (SNP) using the principal component analysis (PCA) (Jolliffe & Cadima, 2016), t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008), and uniform manifold approximation and projection (UMAP) (Diaz-Papkovich et al., 2021; McInnes et al., 2020). However, these visualization methods are sensitive to the data size and noise (Diaz-Papkovich et al., 2019; Dorrity et al., 2020; Li et al., 2017) which can lead to misleading visualization results (Huang et al., 2022). Hence, given these popular methods' shortcomings, better visualization methods that preserve and unravel the dynamics of both global and local structures with interpretability are critically needed.

In this paper, we propose literature text-based visualization approaches using natural language processing techniques for

* Correspondence to: Biomedical Engineering Department, Engineering and Science Building (ESB), Room 407, University of Connecticut Storrs, CT 06269, USA.
E-mail addresses: jihye.moon@uconn.edu (J. Moon), hugo.posada-quintero@uconn.edu (H.F. Posada-Quintero), ki.chon@uconn.edu (K.H. Chon).

https://doi.org/10.1016/j.neunet.2023.05.015

better genetic data visualization with textual interpretability. Natural language processing techniques have been gaining increasing popularity since they can perform multi-dimensional gathering of information and processing to provide effective question answers and summaries (Locke et al., 2021; Zhou et al., 2022). One such successful approach is the use of neural networks for a word embedding model, which has been applied for word distribution representations (Wang et al., 2020), natural language inference (Conneau et al., 2017), text classification (Martinez-Rico et al., 2019), information retrieval (Krämer et al., 2022; Roy et al., 2018), knowledge mining (Yao et al., 2017), and ChatBot assistant systems (Suhaili et al., 2021). The word embedding models are attractive, as they capture semantically similar words related to a specific word in a document from a large-scale text dataset using co-appearance of words. Words with a similar meaning are mapped to a similar location in the vector space (Mikolov et al., 2013) (e.g. "King"-"Man" + "Woman" = "Queen"), which preserves correlations between words systemically (Sang et al., 2020). Since gene symbols (e.g., NUP413) or identified SNP symbols (e.g., rs147843333) have literature data and these genetic symbols can all be encoded as text data, the embedding vectors of words in the related literature may provide significant correlations between the genetic symbols. However, the use of text data for SNP visualization has not been performed.

In general, well-trained neural networks can represent correlations between data points to map higher-dimensional data to lower-dimensional embedding vector space by preserving global and local structures of the data (Fuhrman et al., 2022). Consequently, word embeddings that are trained using genetic symbols and their associated literature-words have feasibility to capture local and global structures of high-dimensional genetic data (e.g., SNP) even though genetic and text data have different modalities. Moreover, an embedding model can capture semantic correlations between words. A literature-based embedding model representation has the potential to incorporate not only genetic and SNP information but other important factors contributing to diseases and provide textual explanations for each genetic symbol. Since semantic text information improves model interpretability (Dong et al., 2017), the literature information encompassing explanations of genetic symbols can provide better understanding of visualization results and may reveal hidden disease risks. Hence, it is critical to develop robust literature-based visualization methods. visualization approach to uncover the dynamics of genetic data using literature text data with textual interpretability. For this purpose, we developed an unsupervised literature text-based neural network-based distribution projection (NNDP) for visualization of genetic data. Our NNDP approach was designed so that: (1) it preserves both global and local structures of the SNP data, and (2) it can provide interpretable and explainable visualization results using textual information. The textual information also enables the feasibility of incorporating other risk factors obtained from literature to visualization results.

For validation of the proposed NNDP, we designed a neural network model called the literature embedding model, and compared the literature embedding model-based NNDP to PCA (Jolliffe & Cadima, 2016), UMAP (McInnes et al., 2020), and random projection (RP) (Bingham & Mannila, 2001). We also compared the literature-model-based NNDP against other NNDP variations of the embedding models—Word2Vec (Mikolov et al., 2013), Global Vectors (GloVe) (Pennington et al., 2014), FastText (Bojanowski et al., 2017; Joulin et al., 2017), Embeddings from Language Models (ELMo) (Peters et al., 2018), Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019), and A Light Bidirectional Encoder Representation from Transformers (ALBERT) (Lan et al., 2020). For this purpose, we used a large SNP data set consisting of (a) dbGaP accession phs000279.v2.p1; and (b) dbGaP accession phs000883.v1.p1 with race, myocardial infarction (MI) event age groups, and sex categories as a genetic dataset. We also collected literature data using SNP-linked gene names from PubMed.

2. Related work

Visualization techniques are paramount to better understanding of high-dimensional complex data (Wang et al., 2021). Ideally, they should tackle high-dimensional data by reducing the dimension without compromising the dynamic structures of the data.

One benefit of reducing the dimension should be reduction of undesired noise as well (Cheng et al., 2022; Dong et al., 2022; Spencer et al., 2020). A study (Dong et al., 2022) used PCA to remove noise so that a graph neural network could be optimized. Another study (Allaoui et al., 2020) used a different dimension-reducing visualization approach, UMAP, to increase clustering performance on image data. This study showed that the use of UMAP increased the accuracy significantly (∼60%) when compared to without UMAP. Data visualization techniques are also widely used to better understand deep neural network structures (Allen et al., 2021; De et al., 2015; Diaz-Papkovich et al., 2019; Moon et al., 2019; O'Donoghue et al., 2018; Rauber et al., 2017). PCA, t-SNE, and UMAP have been widely used to analyze the variation of features generated by convolutional neural networks (Fuhrman et al., 2022; Rauber et al., 2017). However, even though visualization methods can provide an insightful understanding of data, many visualization results are still not fully trustable (Huang et al., 2022). Currently, knowledge graph methods have been investigated for interpretable visualizations (Rožanec et al., 2022; Shimizu et al., 2022). The embedded knowledge graph improved the reliability of visualization results by integrating interpretable information into data features (Deagen et al., 2022; Shimizu et al., 2022; Tiddi & Schlobach, 2022; Zhang & Yao, 2022).

3. Proposed method

Many data visualization approaches such as PCA (Jolliffe & Cadima, 2016) aim to estimate reduced dimension $X'_{n\times M}$ when the original high-dimensional matrix $X_{n\times d}$ is projected onto an M-dimensional subspace using a matrix $E_{d\times M}$, as shown in Eq. (1):

$$X'_{n\times M} = X_{n\times d}E_{d\times M} \tag{1}$$

In this paper, we propose to estimate $E_{d\times M}$ using literature data to obtain reduced M-dimension $X'_{n\times M}$. Genetic entities such as gene names and their associated diseases can appear with other words (their co-apparent words). We hypothesize that the correlations between genetic data points can be obtained using word-to-word correlations between genetic entities represented as text. Hence, we collected literature documents with gene names from PubMed and trained a neural network model to generate optimized $E_{d\times M}$. The trained neural network embedding model has the capacity to compute semantic correlations between not only genetic entities but also words (Roy et al., 2018; Yao et al., 2017); the embedding model was also used to validate and interpret visualization results using semantic correlation analysis. The details of the literature data collection, the literature embedding model design, the literature-represented SNP data creation, data visualization, and semantic correlation analysis are described as follows:

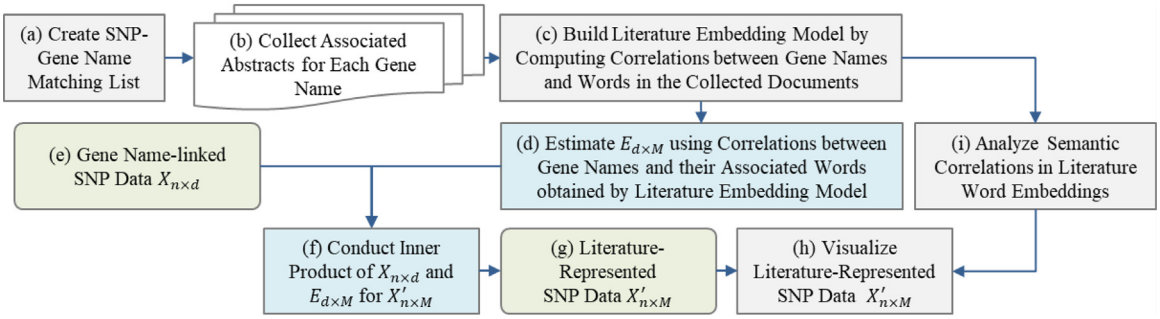


Fig. 1. Overall framework for genetic data visualization and analysis using literature.

Table 1
The number of published abstracts for each decade (1960s–2020s).

Years	Gene name keyword
1960s	3
1970s	35
1980s	156
1990s	1515
2000s	11,893
2010s	96,146
2020s	170,362
N/A	911
Total	281,111

3.1. Overall framework

As shown in Fig. 1, the first step (a) is to create SNP lists with their associated gene name matching lists. In this step we match SNPs and their associated gene names. Step (b) is to collect abstract documents associated with each of the gene names garnered in step (a). Using the obtained document, we build a literature embedding model by computing the correlations between gene names and associated words in the abstract documents in step (c). We estimate $E_{d\times M}$ from the pre-trained literature embedding model in step (d). In step (e) we determine SNP data associated with gene names. SNP arrays in d rows of $X_{n\times d}$ are matched with the SNP-associated gene name arrays in d columns of $E_{d\times M}$. The inner product between the SNP data $X_{n\times d}$ with $E_{d\times M}$ are computed in step (f), which results in a reduced M-dimension $X'_{n\times M}$ in step (g). Step (h) is to visualize the $X'_{n\times M}$ obtained in the (g). Using the pre-trained embedding model (c), the final step (i) is to investigate semantic correlations in word embeddings to validate SNP visualizations using given words' cosine similarity. Step (i) supports understanding and analysis of literature-represented SNP $X'_{n\times M}$-based visualization results in step (h). These processes are all further detailed in the proceeding sections.

3.2. Literature text data collection

We collected 281,111 published abstracts (published date—from 1965 to 2021 January) using 19,264 human gene names (e.g. gene name called TET3) from the PubMed database as shown in Table 1. The 19,264 gene names were associated with identified 920,314 single nucleotide polymorphisms (SNPs) using dbGaP accession phs000883.v1.p1. We extracted gene names from the SNP lists using BioPython library with Python.3.7.1. To preserve information concerning gene-to-gene correlations, we defined the gene names and their associated SNP variant names as unique entities, and they were labeled with upper cases with the "#" symbol appended. All text was converted to lower case. Unwanted words such as prepositions, subordinating conjunctions, determiners, personal pronouns, possessive pronouns, wh-adverbs (e.g., how, when, where, why), modals, comparative adverbs, superlative adverbs, coordinating conjunctions, and existential there is/are were removed. An example is shown in Table 2.

3.3. Literature embedding model

Using the literature data sets, we used a neural network model motivated by the continuous-bag-of-word (CBOW) structure of Word2Vec (Mikolov et al., 2013), which is one of the well-known neural network models in the natural language processing field. The CBOW is designed to predict the center word w_t as the output from a set of words using its context words $\{w_{t-k}, \ldots, w_{t+k}\}$ as the input where t is a word location in a given document, and k is a window size and a model input. The neural network structure optimizes weight matrices that are fully connected with the input and the output words, and the weight matrices learn semantic correlations between words when the network is trained. More details on Word2Vec are describe in Mikolov et al. (2013). We modified the CBOW structure in this study to produce different weight matrices to investigate better visualization approaches using literature-represented SNP. Our model's weight matrices are used as $E_{d\times M}$ in Eq. (1) to produce a dimension-reduced SNP, $X'_{n\times M}$.

Specifically, the model's weight matrices are obtained by training a neural network to predict a search keyword (e.g., a gene name w_j as shown in Table 2(a)) using a context word set $\{w_t, \ldots, w_{t+k}\}$ from collected documents (see Table 2(b)), where t is a word location in a given document and k is a window size. For example, consider a word set consisting of 16 words – obesity, associated, cardiovascular, diseases, body, weight, caloric, intake, organ, weight, lipid, profile, lipoprotein, lipase, #LPL, activity – captured using the search keyword "LPL". All words are tokenized into 16 word pieces. We start the training phases with a window size k=4, thus, by starting at t=0, the first four words (obesity, associated, cardiovascular, diseases) are used to predict the search keyword "#LPL". Subsequently, t increases by one until t=12 is reached with the window containing the words "lipoprotein, lipase, #LPL, activity".

Fig. 2 shows the structure of the neural network model with the modified CBOW structure during model training. The $\hat{x}$, h, and $\hat{y}$ are the input, hidden, and output layers, respectively. The input and output layers are fully connected with two weight matrixes—$U_{N\times M}$ and $V_{M\times N}$. N is the vocabulary size for unique words, M is the number of hidden layers, and K is the window size that defines the number of words that are used as the input vector for each training step. Before model training, we extract unique words from collected documents using all gene names and tokenize all unique words. When we have N number of unique words from documents, we obtain an N-sized

Table 2
Examples of collected abstracts using search keywords.

Status	(a) Search Keyword	(b) Collected Abstract
Original	TET3	**Some of the** environmental conditions **that** lead **to** obesity **are** physical activity, alcohol consumption, socioeconomic status, parent feeding behavior, **and** diet. Interestingly, **some of these** environmental conditions **are** shared **with** neurodegenerative **and** neurodevelopmental diseases.
Pre-processed	#TET3	environmental conditions lead obesity physical activity alcohol consumption socioeconomic status parent feeding behavior diet interestingly environmental conditions shared neurodegenerative neurodevelopmental diseases

Note: Changed words are bolded.

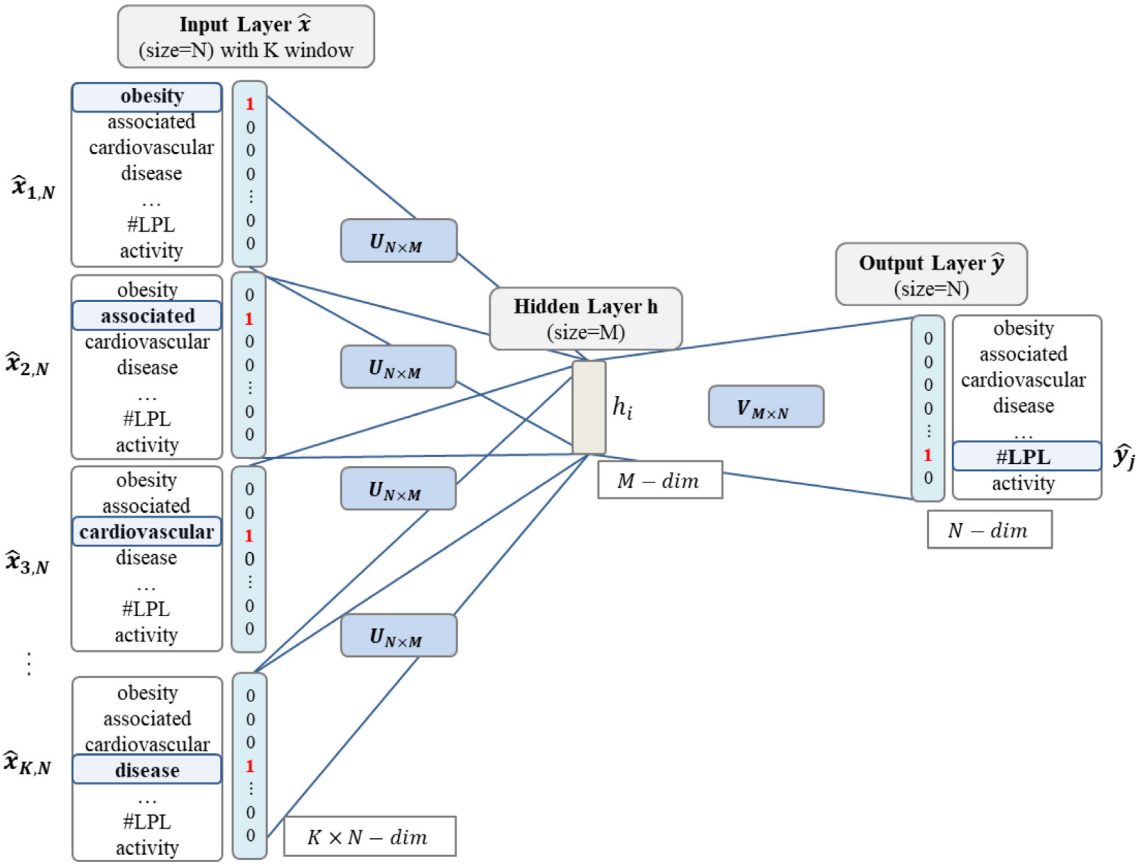


Fig. 2. Proposed literature model structure (modified CBOW is used as our literature embedding model). Related codes are available: https://github.com/JihyeMooon/NNDP_Visualization.

vocabulary of unique words $\{\widehat{w_1}, \ldots, \widehat{w_N}\}$. Each index of every unique word, one-hot encodes each unit of both the input layer $\{\widehat{x_1}, \ldots, \widehat{x_N}\}$ and the output layer $\{\widehat{y_1}, \ldots, \widehat{y_N}\}$ during the model training phase. For example, the unique word 'associated' in the above example of the sequence of words, is indexed with 2 in the input vocabulary $\{x_1, \ldots, x_N\}$, such that the 2nd column of the input is designated as 1 and the others are assigned values of zeros $\{\widehat{x_1}, \ldots, \widehat{x_N}\} = \{0, 1, \ldots, 0, 0\}$. Likewise, when the gene name '#LPL' is indexed with j in the output vocabulary, the jth column of the output layer is coded with 1 and the other columns with 0 (e.g., $\{\widehat{y_1}, \ldots, \widehat{y_j}, \widehat{y_N}\} = \{0, 0, \ldots, 1, 0\}$). For the case of four context words consisting of "obesity, associated, cardiovascular, disease", they are indexed as 0, 1, 2, and 3 with the gene name "#LPL" indexed with j, and the window size $K = 4$. Our model creates four one-hot encoded input layers with the corresponding word appearance index, as shown in Fig. 2. Note that only the output of this structure differs from the original CBOW, and further details for the CBOW structure without our gene-name-based output can be found in Mikolov et al. (2013). The model averages the vectors of the four input layers that represent the four context words, and computes an inner-product between the K-averaged input vector and the weight matrix U:

$$h = \frac{1}{K} U^T \left(\widehat{x_1}, +\widehat{x_2}, + \cdots + \widehat{x_K}\right) = \frac{1}{K}\left(u_{w_t} + u_{w_{t+1}} + \cdots + u_{w_{t+K}}\right)^T \quad (2)$$

In Eq. (2), K is the number of context words (window size), $\hat{x}$ is the input layer, $\{K1, \ldots, K4\}$ denote the index locations that are coded with either 1 or 0, $\{w_t, \ldots, w_{t+k}\}$ are the context words

(e.g., "obesity, associated, cardiovascular, disease"), t is a word location in a given document, u_w is the input vector of a word w, and h is the hidden layer. To predict the gene name "#LPL" that is indexed with j, we compute the inner product between the hidden layer and jth row of the matrix V:

$$z_j = v_{w_j}^T h \tag{3}$$

In Eq. (3), v_{w_j} is the output vector linking to the gene name "#LPL" that is located in the jth row of the matrix V of $M \times N$ size. Using these weights, we compute a score z_j for each gene name indexed as the jth in the vocabulary. Then we obtain the posterior distribution of words, a multinomial distribution using a softmax function, using Eq. (4):

$$p\left(w_j|w_t,\ldots,w_{t+k}\right) = \widehat{y_j} = \frac{\exp\left(z_j\right)}{\sum_{i'=1}^{N}\exp\left(z_{j'}\right)} \tag{4}$$

where $\widehat{y_j}$ is the output of the jth column unit in the output layer. By considering Eqs. (2)–(3), Eq. (4) is reformulated as:

$$p\left(w_j|w_t,\ldots,w_{t+k}\right) = \frac{\exp\left(v_{w_j}^T \cdot \frac{1}{\mathrm{K}}\left(u_{\mathrm{w_t}} + u_{\mathrm{w_{t+1}}} + \cdots + u_{\mathrm{w_{t+K}}}\right)^{\mathrm{T}}\right)}{\sum_{i'=1}^{N}\exp\left(v_{w_{j'}}^{T\ T} \cdot \frac{1}{\mathrm{K}}\left(u_{\mathrm{w_t}} + u_{\mathrm{w_{t+1}}} + \cdots + u_{\mathrm{w_{t+K}}}\right)^{\mathrm{T}}\right)} \tag{5}$$

In Eq. (5), u_w is the input vector for a word w_t, and v_j is the output vector for a gene name w_j that is indexed as j in the vocabulary. Since the input and output layers are fully connected with U and V, the model structure aims to optimize the U and V by predicting the output $\hat{y}_j$ (gene name w_j) using the average $\hat{x}$ (context words $\{w_t,\ldots,w_{t+k}\}$). The structure maximizes the average log probability of the word sets $\left(w_j|w_t,\ldots,w_{t+k}\right)$ to predict their associated gene name w_j, which can be defined as:

$$\frac{1}{\mathrm{T}}\sum_{t=n}^{T-k}\log p(w_j|w_t,\ldots,w_{t+k}) \tag{6}$$

In Eq. (6), T *and* t are the number of words and a word location in a given document, respectively, and k is a window size used for each training step. The model iterates until all words and search keywords are all accounted for as the input and output vectors, respectively. The process provides a vast distributed representation for each gene w_j using word correlations in the collected abstract documents without any prior knowledge of the genotype data. With the context words, word vectors that appear in similar contexts get aligned closer to each other by predicting the same output (e.g., a gene name) within the $U_{N\times M}$ matrix space. With our structure, $U_{N\times M}$ and $V_{M\times N}$ represent different distributions for each word—systemically, the $U_{N\times M}$ is formulated by all words, including gene names that are used as the input vector, whereas $V_{M\times N}$ is represented by only the gene names that are used as the output vector. Consequently, the $U_{N\times M}$ matrix provides better correlations between gene names that were obtained by the word–word appearances than does the $V_{M\times N}$ matrix. To validate our method, we used $U_{N\times M}$ and $V'_{N\times M}$ (a transpose of $V_{M\times N}$) as $E_{d\times M}$.

3.4. Literature-represented SNP data

To use the weight matrix ($U_{N\times M}$ or $V'_{N\times M}$) with $X_{n\times d}$, we extract the sub-weight matrices $U_{d\times M}$ and $V'_{d\times M}$ from $N \times M$ matrixes and their d-rows are matched with d-columns of $X_{n\times d}$, where d is the SNP gene name matched index and M is the dimension. The sub-weight matrix ($U_{d\times M}$ or $V'_{d\times M}$) is normalized by using the Frobenius norm (Golub & Van Loan, 2013). Using the normalized matrix, our NNDP takes the inner product between $X_{n\times d}$ with the sub-weight matrix (which can be either $U_{d\times M}$ or $V'_{d\times M}$) to produce $X'_{n\times M} = X_{n\times d}E_{d\times M}$ as shown in Eq. (1). $E_{d\times M}$ can be either $U_{d\times M}$ or $V'_{d\times M}$ and $E_{d\times M}$ projects the original matrix $X_{n\times d}$ onto an M-dimensional subspace, which leads to transformation from d-dimensional original data $X_{n\times d}$ to M-dimensional $X'_{n\times M}$. Note that the dimension M of $E_{d\times M}$ is defined as the embedding model's dimension and the dimension M determines the dimension of $X'_{n\times M}$. In this paper, we designate $X'_{n\times M}$ with $U_{d\times M}$ (input representation) as NNDP-IN and with $V'_{d\times M}$ (output representation) as NNDP-OUT.

3.5. Visualization using literature-represented SNP

We proposed a data visualization approach using our NNDP, and compared these to other popular visualization models such as the PCA and UMAP. Since NNDP provides a fixed dimension M, it can be combined with either PCA or UMAP on $X'_{n\times d}$ (literature-represented SNP data) for data visualization. We term these various combinations of using NNDP with either PCA or UMAP as the following: NNDP-IN-PCA, NNDP-IN-UMAP, NNDP-OUT-PCA, and NNDP-OUT-UMAP.

3.6. Evaluation of semantic correlations in word embeddings

Since our dimensionality reduction and visualization approaches are based on semantics correlations between words provided in $U_{N\times M}$ and $V'_{N\times M}$ (a transpose of $V_{M\times N}$) matrices, we validate and interpret our methods using textual semantic analysis involving the embedding matrices. These pre-trained matrices provide an N-sized vocabulary of unique words $\{\widehat{w_1},\ldots,\widehat{w_N}\}$ with M-dimensional unique embedding vectors. Cosine similarity between embedding vectors measures semantic correlations between them. The cosine similarity-based analysis is a common evaluation to assess the quality of model in the natural language processing field (Wang et al., 2019). We validate and interpret literature-based visualization using cosine similarity scores between a query word vector and other embedding vectors in $U_{N\times M}$ or $V'_{N\times M}$. Since our embedding vectors are normalized (see Section 3.4), the similarity is computed as follows:

$$Similarity\left(v_i,\ v_j\right) = |v_i|\,\big|\,v_j\big|\,cos\theta = v_i^T v_j \tag{7}$$

With Eq. (7), we calculate similarity scores between the two-word vectors v_i and v_j in the M-dimensional embedding space. The vector v is obtained from $U_{N\times M}$ or $V'_{N\times M}$. The cosine similarity metric ranges from -1 to 1, with -1 representing the most different, 0 as no correlation, and 1 as the most similar between data points. Using query words (i.e., "heart"), Eq. (7) calculates similarities of embedding vectors of the query and all words in the unique vocabulary of $U_{N\times M}$ or $V'_{N\times M}$, and sorts the words associated with the highest cosine similarity values. When the embedding model is well-trained, the captured words show semantic correlations to query word (i.e., a captured word "cardiac" for a query "heart"). The similarity analysis was used to understand and validate visualization results along with identifying semantic correlations between words.

4. NNDP variations using other embedding models

We also investigated our NNDP literature text-based visualization method using other embedding models. For the dimension-reduced $X'_{n\times M} = X_{n\times d}E_{d\times M}$, the $E_{d\times M}$ can be replaced with any word embedding models for the purpose of SNP data visualization. Hence, not only did we compare our NNDP method with traditional visualization methods – PCA, UMAP, and RP – but also, we compared literature-model-based NNDP (NNDP-IN and NNDP-OUT) against other NNDP variations of the embedding

models: the original Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT.

The word embedding models are categorized into two groups: **conventional word embedding** (Word2Vec, GloVe, and FastText) and **contextual word embedding** (ELMo, GPT-2, and ALBERT). The former generates context-independent word embeddings while the latter generates context-dependent word embeddings. For example, the word "watch" can have different meanings depending on the context, such as: "I like to **watch** television" vs. "I am wearing a smart **watch**". While the conventional embedding models – Word2Vec, GloVe, and FastText – consider different uses of "**watch**" to be the same, the contextual embedding models – ELMo, GPT-2, and ALBERT – capture the different meanings of "**watch**" in each sentence by considering its context. We investigated how the context-independent and context-dependent embedding representations work for this NNDP approach.

The Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT-based NNDP were defined as NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT. When these embedding models were used with either PCA or UMAP, the following naming notations were used: NNDP-Word2Vec-PCA, NNDP-Word2Vec-UMAP, NNDP-GloVe-PCA, NNDP-GloVe-UMAP, NNDP-FastText-PCA, NNDP-FastText-UMAP, NNDP-ELMo-PCA, NNDP-ELMo-UMAP, NNDP-GPT-2-PCA, NNDP-GPT-2-UMAP, NNDP-ALBERT-PCA, and NNDP-ALBERT-UMAP.

4.1. Literature data processing for NNDP variants

For the other embedding models, gene names and their associated sentences are merged into a sentence to be used as a training corpus to maximize the correlations between gene names and associated words. For example, a sentence "some of the environmental conditions that lead to obesity are physical activity and alcohol consumption" was obtained using a gene name "#TET3" so we created the merged: "#TET3 some of the environmental conditions that lead to obesity are physical activity and alcohol consumption". The merged sentences were used for the training corpus of the other embedding models—Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT.

4.2. Description of other embedding matrix extraction models

This subsection demonstrates how to extract embedding matrices from each of the six compared embedding models: NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT. All extracted embedding matrices were normalized via the Frobenius norm (Golub & Van Loan, 2013).

4.2.1. NNDP-Word2Vec

Word2Vec is the original version of our literature embedding model (Mikolov et al., 2013). It is a simple neural network consisting of an input layer, a hidden layer, and an output layer. The original Word2Vec predicts a center word using word contexts (CBOW structure) or word contexts using a center word (Skip-gram structure) while our embedding model predicts a gene name using a word context from an associated document. The Word2Vec structure allows participation of all words in the input and output while our literature embedding structure allows participation of words only in the input. More details of Word2Vec are described in Mikolov et al. (2013). Since our model was modified from the CBOW structure of Word2Vec, we used CBOW structure for NNDP-Word2Vec in this work. We used the d-dimensional sub-weight matrix of Word2Vec's input matrix as $E_{d\times M}$, where d is the matched SNP gene name index and M is the dimension.

4.2.2. NNDP-GloVe

GloVe is designed to address the local-context-biased representations of Word2Vec (Pennington et al., 2014). Word2Vec generates embedding vectors using only local contexts (Pennington et al., 2014; Wendlandt et al., 2018). Generally, GloVe and Word2Vec are similar since both models compute word-to-word co-occurrences using word contexts. However, GloVe also uses word-to-word global co-occurrence counts from the entire training corpus. GloVe's objective function specifies that the inner product between the center word embedding and the context word embedding should equal the logarithm of the words' probability of co-occurrence (Pennington et al., 2014). Since GloVe's structure involves both global and local word-to-word co-occurrence probabilities, its embedding space is more stable than that of Word2Vec (Mimno & Thompson, 2017; Wendlandt et al., 2018). We used the d-dimensional sub-matrix of the sum matrix of the center and context embedding matrices as $E_{d\times M}$. The sum of two embedding matrices is the in the GloVe library provided by (Pennington et al., 2014).

4.2.3. NNDP-FastText

Fasttext is designed to add morphological information into unique-word-based representations of Word2Vec to represent the relationships between characters per word. For example, words such as "genes" and "genetic" are different forms of the word "gene". The relationship of the two words "boy" and "friend" to create the word "boyfriend" are the same as the relationship of the two words "girl" and "friend" to create the word "girlfriend". Since Word2Vec and GloVe only consider the unique-word-based relationships as described in 4.2.1 and 4.2.2, they do not capture the internal information for each word. To obtain the morphological information, FastText suggests a sub-word n-gram approach to obtain the order relationship between characters in each word: for the word "subword", the FastText model adds two characters "<" and ">" to create <subword>. Then n-gram information is computed to capture the relationship between characters. If 3-gram information is selected, all sequential information for each sub-word is paired as three characters: "<su", "sub", "ubw", "bwo", "wor", "ord", "rd>". The correlations between characters are added into word embedding vectors while the FastText model trains. We used the d-dimensional sub-matrix of the FastText embedding matrix as $E_{d\times M}$.

4.2.4. NNDP-ELMo

ELMo generates contextual word embedding vectors by predicting next or prior words via left-to-right and right-to-left contexts for each given sentence, using a bidirectional long short-term memory network (LSTM) (Peters et al., 2018) with character convolutional neural networks (CNN). The character CNNs capture character n-gram information by applying multiple convolutional layers on characters in each word, and the bidirectional LSTM assigns each word a representation based on its context. The bidirectional LSTM looks at left-to-right and right-to-left contexts, which enables it to capture uses of words varying by different contexts. With these bi-directional representations, ELMo produces different word embedding representations for each word depending on its contexts. Since gene name is an entity term that does not change meaning with its context, we used the d-dimensional sub-matrix extracted from the pre-trained M-dimensional contextualized representation for each gene name as $E_{d\times M}$.

4.2.5. NNDP-GPT-2

GPT-2 with a causal language modeling generates contextual word embedding vectors by predicting the next word using prior words in a given sentence set with deep bidirectional transformer decoders (Radford et al., 2019). The special feature of GPT-2 is Byte Pair Encoding (BPE) tokenizer. BPE (Radford et al., 2019) replaces frequent word sequences as word level tokens and infrequent symbol sequences as character level tokens, which results in reduction of vocabulary size. GPT-2 is similar to ELMo in that both models capture contextual information for each word per its context. However, GPT-2 captures the contexts using unidirectional representations via the state-of-the-art neural network transformer decoder instead of bidirectional LSTMs. Since the bidirectional LSTM predicts prior/future information sequentially, it is biased to local contexts. However, GPT-2's self-attention mechanism enables capturing all contexts equally for each word in each given sentence. Since the GPT structure generates unidirectional representations by predicting the next word using the prior words, GPT structures have been widely used for text generation tasks. We used the d-dimensional sub-matrix extracted from the M-dimensional context-dependent representations (the last embedding matrix of GPT-2) for each gene name as $E_{d\times M}$.

4.2.6. NNDP-ALBERT

ALBERT with a masked language modeling produces contextual word embedding vectors by predicting sentence orders and randomly masked words with deep bidirectional transformer encoders (Lan et al., 2020). Even though both ALBERT and GPT-2 use the transformer structure, ALBERT differs from GPT-2 as it predicts randomly masked words using all nearby words in each sentence using **the transformer encoder** while GPT-2 predicts the next words using prior words autoregressively with **the transformer decoder**. One of the special features of ALBERT is the word-piece tokenizer: the tokenizer creates vocabulary using sub-word units and ALBERT generates embedding vectors based on the sub-word vocabulary (Song et al., 2021). For example, the word "colorless" is decomposed to "color" and "##less" by the word-piece tokenizer. Since the tokenizer could also decompose gene name into some pieces, we averaged the word-piece embedding vectors to create a complete d-dimensional gene-name sub-matrix $E_{d\times M}$. For example, a gene name #GYPC is decomposed to '#', '##gyp', and '##c', then we averaged all of the three-piece-embedding vectors to create a unique word embedding vector for '#GYPC'.

To examine the effect of different semantic representations on our NNDP approach, we computed two different d-dimensional gene-name sub-matrices $E_{d\times M}$ using a token embedding matrix either alone or as the sum of token and positional embedding matrices for the context-dependent and context-independent cases, respectively. The NNPD approach using a context-**d**ependent embedding matrix (using the sum of token and positional embedding matrices) of ALBERT we called NNDP-ALBERT[**d**], and the NNDP approach using a context-**i**ndependent embedding matrix (using only token embedding) of ALBERT we called NNDP-ALBERT[**i**].

5. SNP database

5.1. Data description and encoding

We used SNP data from two case-control study datasets as $X_{n\times d}$: (a) dbGaP accession phs000279.v2.p1 which contains MI events for two age groups (512 subjects of young MI event age group, 206 subjects of old MI event age group), and five race groups (552 European, 145 African American (Black), 14 Hispanic, 14 others, and 2 Unknown), and (b) dbGaP accession phs000883.v1.p1 for two sex group (1331 male, 723 female). The dataset (a) contains 727 subjects and identifies 401,454 SNPs for the five race groups and two MI event groups. The event age is defined as the subject's age when MI occurred. Since the statistical differences between the two age groups <50 and ≥ 50 at the first incident of MI are significant regarding the family history of MI (Ambroziak et al., 2020), we defined the young MI event group as age ranges from 20 to 50 and the old MI event group as age ranges from above 50 to 60. The dataset (b) contains data from 2054 subjects (1331 male, 723 female) with 920,314 SNPs, and MI case-control labels (1030 case, 1024 control). Those groups with and without MI shared the same risk factors for MI at the baseline examination. However, the study did not conduct a follow-up study to examine if the control group's subjects had MI events, hence, we only used sex category for the database (b).

The SNP can be encoded by three possible values consisting of homozygous for reference, heterozygous, and homozygous for alternate. The homozygous for reference is where the two base pairs of SNP are the same and found in the reference genome; the heterozygous is where the two base pairs are different; and homozygous for alternate is where the two base pairs are not found in the reference genome (Soumare et al., 2021). By following previous works' SNP labeling strategy for machine learning (Monk et al., 2021; Patel et al., 2015), we encoded the data components as ordinary values with four criteria: if the SNP variant is homozygous for the reference allele, it is labeled as 0. If the SNP variant is heterozygous for reference and alternative alleles, it is labeled as 1. If the genotype is homozygous for the alternate allele, it is labeled as 2. If there is missing data for a particular sample, it is labeled as -1.

5.2. SNP data selection

We excluded SNPs when (1) the percentage of missing samples was > 5%; (2) they have the same value for all subjects (i.e., all subjects have 0 value (homozygous for reference) for an SNP); (3) they have no matched gene names (some SNPs do not have gene references); (4) SNP-matched gene names have no published abstracts in PubMed. When these exclusions were applied, the number of SNPs reduced from 401,454 to 385,706 for the dataset (a) and from 920,314 to 239,027 for the dataset (b). The SNP data were segmented into 80% for training and 20% for testing. We selected race, MI event age group, and sex-related SNPs based on the training data set. The data selection is to evaluate our proposed model's performance in capturing the local and global structure estimation. The quality-control procedure is the most common approach to remove unwanted SNPs for specific topics such as population and disease risk (Gola et al., 2020). For SNP selections based on the training data, we selected a Q*128 dimensional SNP set where Q=50 using a random forest classification model developed by scikit-learn v. 0.23.2 (Pedregosa et al., 2011) with Python 3.7. We used a grid search algorithm provided by scikit-learn for the random forest model. We obtained 6400 race-related SNPs, 6400 MI event age-related SNPs from (a) dbGaP accession phs000279.v2.p1, and 6400 sex-related SNPs from (b) dbGaP accession phs000883.v1.p1 (Q*128 where Q=50) for training and testing datasets. The segmented training and testing datasets were used to validate visualization and classification performance. We also investigated classification and data visualization performance using Q=100, 150, 200, and 250 in the Appendix to assess performance depending on Q*128 dimension (Q times dimensionality reduction).

6. Method validation

In this paper, we aimed to unravel global and local structural dynamics of the genetic data together with evaluations of semantic correlations in literature word embeddings. We validated our NNDP-based methods using both qualitative and quantitative evaluation methods. Global structure represents common and general features that many individuals share, including their genotypes such as sex and race, while the local structure represents certain diseases. We used race (European, Black, Hispanic, other, and unknown), MI event age groups (young MI event age group, old MI event age group), and sex (male, female) information for the visualization analysis that was the qualitative evaluation, and classification tasks as the quantitative evaluation. We analyzed whether the model is interpretable and contain other information that could be visualized, and quantified using semantic analysis. The generalizability of our method by using missing data simulation was also evaluated.

For this purpose, we validated our approach using details provided in the following four subsections: 6.1. Data visualization, 6.2. Evaluation of semantic correlations in word embeddings, 6.3. Classification with dimensionality reduction, and 6.4. Missing data simulation.

6.1. Data visualization

We visually assessed SNP data separations between classes as a qualitative evaluation. We defined distinct separations between different classes for each category (race, MI event age, and sex) as a metric of good visualization quality. In order to investigate the effects of literature information on MI risks in details, we analyzed how MI event age group-related clusters are different for each race group using the dbGaP accession phs000279.v2.p1 dataset that was described in Section 5. To reduce the complexity of the analysis, we used five classes (Young European MI, old European MI, young Black MI, old Black MI, and others) by combining two labels: race (European, Black, Hispanic, other, and unknown) and MI event age groups (young group ranges from 20 to 50 and old group ranges from above 50 to 60).

6.2. Evaluation of semantic correlations in literature word embeddings

We investigated interpretable visualizations via semantic correlations in literature word embeddings. Using Eq. (7), we computed all similarities between each query and all vocabulary words. We sorted top-10 similar words from the query and analyzed semantic correlations between them. The query vectors that consisted of more than two words (i.e., 'female + male') were averaged. We used 'heart', 'african + american', 'european + american', and 'female + male' as queries to interpret the visualization results semantically.

6.3. Classification with dimensionality reduction

For quantitative evaluation, we used the literature-represented SNP data $X'_{n\times M}$ (also called dimension-reduced SNP data) as the input to machine learning (ML) models which consisted of support vector machine (SVM) with linear, poly, and radial basis functions (RBF), logistic regression (LR), and multilayer perceptron (MLP). For classification performance evaluations, we used accuracy, sensitivity, specificity, and geometric mean score (G-mean). The G-mean is the squared root of the product of sensitivity and specificity. We also conducted statistical hypothesis testing with a two-sided p-value (Foody, 2009) to determine a significant difference between dimension-reduced data and the original data. The difference between a pair of models' performance was considered significant if the p-value $\leq$ 0.05 (Devore, 2000) for two-sided 5% significance level.

6.4. Missing data simulation

To demonstrate generalizability of the proposed approach, we considered missing SNP data. The missing data environment was simulated by replacing 10% of the SNP data with -1. Specifically, for the race and MI event age group categories, 465,280 arrays were replaced with -1 from 4,652,800 arrays derived from 727 subjects and 6400 SNPs. For the sex category, 1,314,560 arrays were replaced with -1 from 13,145,600 arrays obtained from 2054 subjects and 6400 SNPs. We analyzed the generalizability of the model using both qualitative and quantitative evaluations that are described in Sections 6.1, 6.3.

6.5. Model comparison

An unsupervised learning approach involves reducing the dimension of the data and visualizing the data in two-dimensional space. To compare our data visualization approach with other popular dimensionality reduction methods, we examined PCA (Jolliffe & Cadima, 2016), UMAP (McInnes et al., 2020), and RP (Bingham & Mannila, 2001). RP projects the original high-dimensional data matrix $X_{n\times d}$ onto a low-dimensional subspace using a random distribution matrix $E_{d\times M}$ (e.g. Gaussian) and the inner product between the two matrices formulated as $X^{RP}_{n\times M} = X_{n\times d}E_{d\times M}$. The obtained $X^{RP}_{n\times k}$ preserves the distance between data points of $X_{n\times d}$ onto lower-dimensional space based on the Johnson–Lindenstrauss lemma theory (Johnson, 1984). We used the M-dimensional Gaussian distribution $E_{d\times M}$, where M is 128 for RP (Egecioglu et al., 2004). For classification tasks (quantitative evaluation) for all methods (NNDP-IN, NNDP-OUT, NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, NNDP-ALBERT, PCA, UMAP, and RP), the data dimension was reduced from 6400 to 128. The dimension-reduced SNP data (termed the literature-represented SNP for our method) were used as the input of ML models. For visualization of the data for all methods, we reduced the data dimension to 2-D. For example, PCA and UMAP visualized the entire 6400-dimensional SNP data to 2-D. Our NNDP method and RP have a fixed dimension M by $E_{d\times M}$. Hence, our NNDP and RP were visualized using PCA and UMAP after the data dimension was reduced to 128. These various combinations of using NNDP and RP with either PCA or UMAP were termed the following: NNDP-IN-PCA, NNDP-IN-UMAP, NNDP-OUT-PCA, NNDP-Word2Vec-PCA, NNDP-GloVe-PCA, NNDP-FastText-PCA, NNDP-ELMo-PCA, NNDP-GPT-2-PCA, NNDP-ALBERT-PCA, NNDP-OUT-UMAP, NNDP-Word2Vec-UMAP, NNDP-GloVe-UMAP, NNDP-FastText-UMAP, NNDP-ELMo-UMAP, NNDP-GPT-2-UMAP, NNDP-ALBERT-UMAP, RP-PCA, and RP-UMAP.

7. Model training and visualization

7.1. Literature embedding model training

We trained our embedding model with the dimension of 128 using 281,111 published abstracts. For the hyper-parameters, we used negative sampling of 64, which is an alternative to the hierarchical softmax function (Mikolov et al., 2013), a minimum word count of 4, a window size of 4 (to capture 4 context words), an epoch of 30, and a learning rate of 1.0 with a gradient descent optimizer. The above-noted epoch size was chosen since the embedding models remained stable until epoch 30 (Borah et al., 2021), and the learning rate was chosen since it is the most widely used for embedding model training (Amalia et al., 2020; Dürrschnabel et al., 2022; Kowsher et al., 2022; Kuyumcu et al., 2019). The embedding model was trained using Python 3.7 with Tensorflow ver. 1.18.3 (Abadi et al., 2016). The number

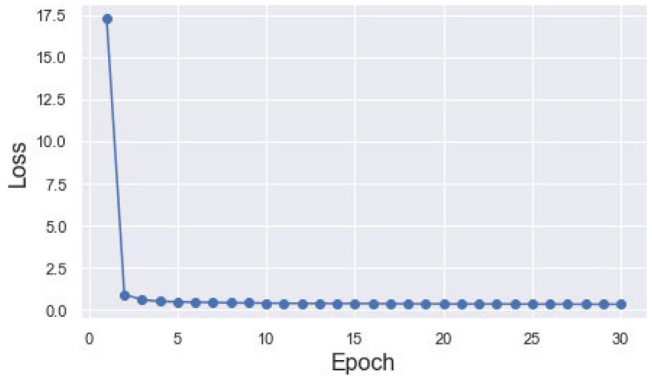


Fig. 3. Training losses at each epoch for the literature embedding model.

of unique vocabulary elements was 241,315 consisting of 221, 789 unique words and 19,526 gene names. Gene names were not excluded from the training corpus, regardless of their frequency count being less than 4. The training loss value for each epoch for the model is shown in Fig. 3.

7.2. Other embedding models' training

In this paper, we trained six embedding models—Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT. These models were trained to provide 128-dimensional embedding vectors.

Conventional word embeddings: For the hyper-parameters of Word2Vec, we used CBOW structure and a window size of 2 (to capture 4 context words) with the same hyper-parameters used in our embedding model, which are described in Section 7.1. Word2Vec was trained for 30 epochs using Python 3.7 with genism ver. 4.2.0. GloVe was trained using a window size of 2 and a minimum word count of 4 and the default hyper-parameters consisting of symmetric context, alpha of 0.75, and x_max of 100.0 with GloVe ver.1.2 library, C (Pennington et al., 2014). FastText was trained based on CBOW structure using a window size of 2, a minimum word count of 4, with 128 dimension and default hyper-parameters as noted in previous work (Bojanowski et al., 2017) using FastText library. The hyper-parameter was designed to account for character 3–5 gram. For all word embedding models, gene names were retained in the training corpus, irrespective of their frequency count being less than 4. Since these conventional word embedding models are trainable with CPUs, all conventional word embeddings were trained using Intel(R) Xeon(R) E- 2246G CPU @3.60 GHz, and 32 GB memory.

Contextual word embeddings: Since all conventional embedding models are designed to provide 128 embedding vectors in this work, the structures of ELMo and GPT-2 were adjusted to provide 128 embedding vectors. Since the original ELMo provides 1024 embedding vectors by concatenating 512-dimentional two-layer bidirectional LSTMs of 4096 units, our ELMo model was adjusted to provide 128 embedding vectors by concatenating 64-dimensional two-layer bidirectional LSTMs of 1028 units. Since the representation dimensions of the adjusted LSTMs are 8 times smaller than the dimensions of the original ELMo's LSTMs, the total number of character n-gram convolutional filters was also reduced 8 times (from 2048 to 256). ELMo was trained for 10 epochs using the default hyper-parameters noted in the previous work (Peters et al., 2018) with 256 batch size using a Tesla-T4 GPU with Tensorflow ver. 2.12.0, Python 3.9.

Since the default embedding size of GPT-2 is 768 (small version of GPT-2) as noted in the previous work (Radford et al., 2019), the GPT-2 structure was also adjusted to produce an embedding size of 128. The number of attention heads was 8 and each attention head size was 16 resulting in embedding size of 128. The GPT-2 model was trained using the following parameters: a maximum sequence length of 512, vocabulary of 52,256, embedding size of 128, the number of attention heads of 8, and the number of hidden layers of 12 with 18 batch size with the default hyper-parameters of small GPT-2 provided by HuggingFace library (Wolf et al., 2020) using a Tesla T-4 GPU with Pytorch 1.13.1, Python 3.9.

Since the original ALBERT is designed to provide 128 embedding vectors, our ALBERT model was trained using the default hyper-parameters of the original ALBERT-base structure: learning rate of 0.00176 (LAMB optimizer), number of training steps of 125,000, vocabulary of 30,000, maximum sequence length of 512, embedding size of 128, hidden size of 768, the number of attention heads of 12, and the number of hidden layers of 12 with 128 batch size. The details for the ALBERT-base's hyper-parameters are described in Lan et al. (2020). The ALBERT model was trained using eight Google Cloud TPU V3s with TensorFlow ver.1.15, Python 3.7 (Abadi et al., 2016).

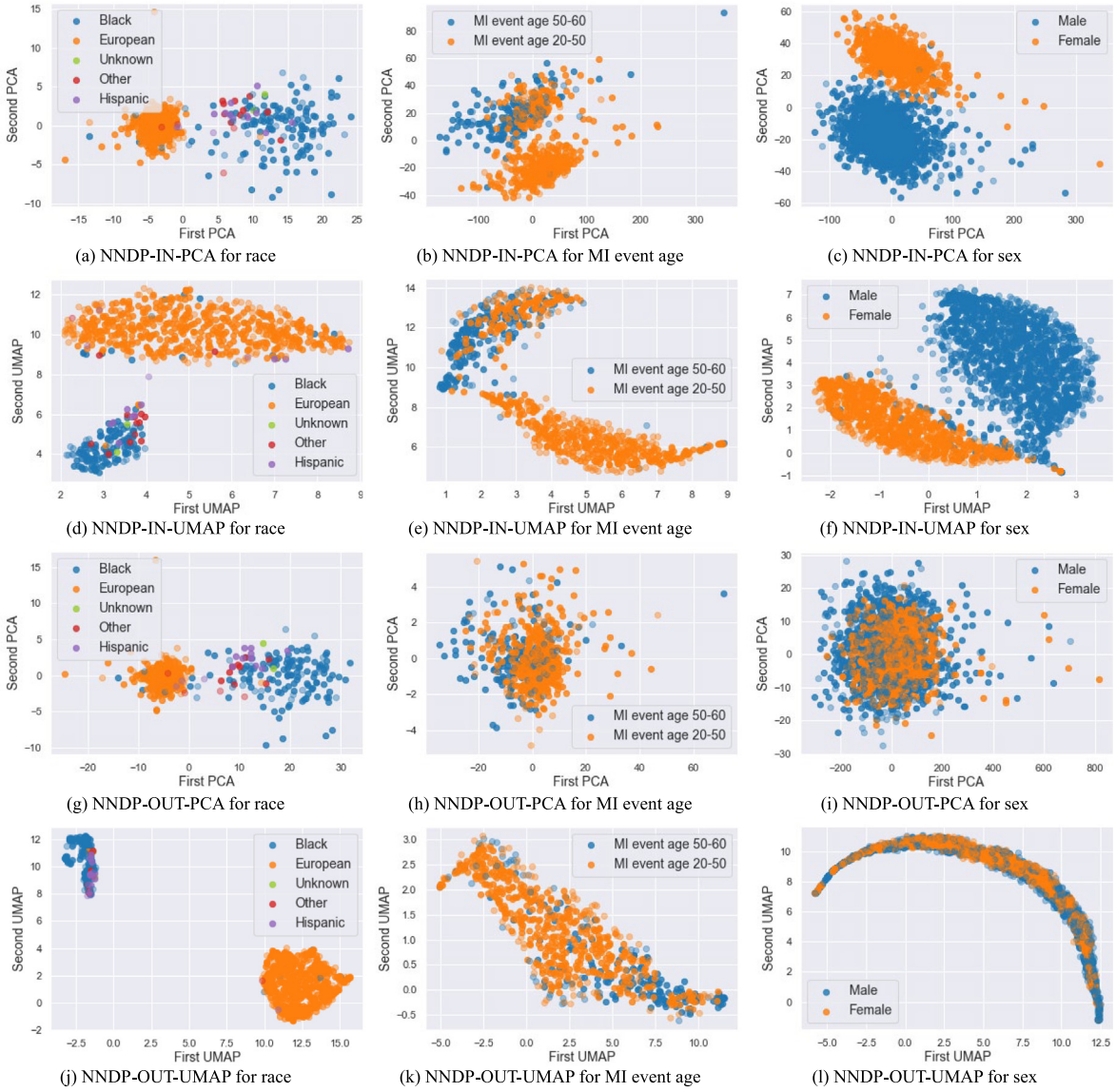
7.3. Classification model training and validation

The SNP data were segmented into 80% for training and 20% for testing. Dimension-reduced SNP data were created via NNDP-IN, NNDP-OUT, NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT, PCA, UMAP, and RP, and we trained all classifiers using the literature-represented /dimension-reduced SNP data for the race, MI event age groups, and sex classifications. We validated NNDP using the 5-fold cross validation for the classification task. For the 5-fold cross validation, the performance on the five testing data sets was averaged. We conducted data standardization before the dimensionality reduction tasks to prevent biases of the feature values.

Since our NNDP approach projects data onto a 128 -dimensional space, the other compared approaches – PCA, UMAP, and RP – also used the same dimension. We conducted the PCA and UMAP using scikit-learn v. 0.23.2 (Pedregosa et al., 2011) with Python 3.7. The hyper-parameters of those models were provided by the default values of scikit-learn. The mean and standard deviation of Gaussian distribution for RP were defined as 0 and 1, respectively. The five ML models: SVM (linear, poly, and RBF kernels), LR, and MLP were trained using scikit-learn v. 0.23.2 (Pedregosa et al., 2011) with Python 3.7 with grid search algorithms to select the best machine learning models. The final performance on the test data was selected based on the best accuracy during model training. For SVM and LR, the final hyper-parameters were selected from {1e−5, 1e−4, 1e−3} for 'tol' and {0.1, 1.0} for 'C.' For MLP, the hyper-parameters were selected from {(100), (100,100), (100, 100, 100)} for 'hidden_layer_sizes' and {0,01, 0,01, 0,001} for 'learning_rate_init.' Other hyper-parameters of these models were set by the default values provided by scikit-learn.

7.4. SNP visualization

We conducted data visualization on race, MI event age, and sex-related SNP datasets. NNDP-based SNP visualization was implemented by using PCA and UMAP after NNDP operation (NNDP-PCA, NNDP-UMAP) on the SNP data set. The dimension of the NNDP embedding model was 128, so the input size of each PCA and UMAP was 128. The output size of NNDP-PCA and NNDP-UMAP was 2, for 2-D visualization. To validate our NNDP-based data visualization, we also conducted PCA, UMAP, RP-PCA (PCA operation after RP), and RP-UMAP (UMAP operation after RP).



(a) NNDP-IN-PCA for race (b) NNDP-IN-PCA for MI event age (c) NNDP-IN-PCA for sex

(d) NNDP-IN-UMAP for race (e) NNDP-IN-UMAP for MI event age (f) NNDP-IN-UMAP for sex

(g) NNDP-OUT-PCA for race (h) NNDP-OUT-PCA for MI event age (i) NNDP-OUT-PCA for sex

(j) NNDP-OUT-UMAP for race (k) NNDP-OUT-UMAP for MI event age (l) NNDP-OUT-UMAP for sex

Fig. 4. Visualization result using NNDP methods (NNDP-IN-PCA, NNDP-IN-UMAP, NNDP-OUT-PCA, NNDP-OUT-UMAP for race (a, d, g, j), MI event age group (b, e, h, k), and sex (c, f, i, l) categories.

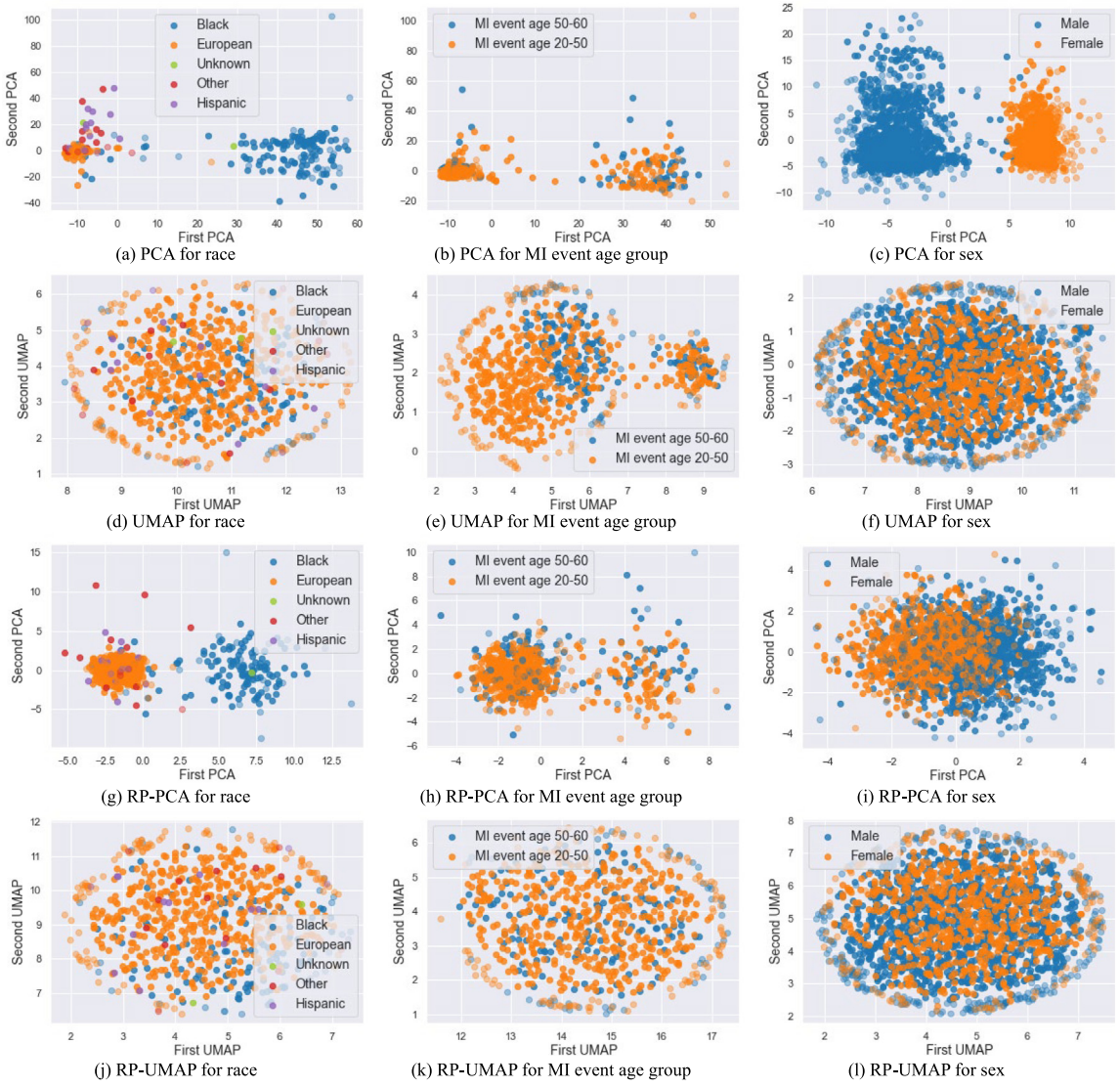
8. Results and analysis

To validate our NNDP-based methods objectively, we examined other popular dimensionality reduction models, namely, PCA, UMAP, and RP on the original 6400-dimensional SNP data for both visualization and classification tasks. We analyzed all visualization and classification results with the following comparison: **(1) NNDP-IN and -OUT vs. PCA, UMAP, and RP**, **(2) NNDP-IN and -OUT vs. NNDP-Word2Vec, -GloVe, -FastText, -ELMo, -GPT-2, and -ALBERT.**

Our NNDP-based methods (NNDP-IN, NNDP-OUT, NNDP-Word2Vec, NNDP-GLoVE, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2), and NNDP-ALBERT (ALBERT[d] and ALBERT[i]) and RP reduced the dimension of each 6400-dimensional SNP data set to 128 for the classification task (quantitative evaluation), then we applied PCA and UMAP on the dimension-reduced NNDP and RP outputs for visualization (qualitative evaluation). Note that distribution projection methods including NNDP and RP reduce the data dimension depending on the distribution $E_{d\times M}$'s size to form $X'_{n\times M} = X_{n\times d}E_{d\times M}$, thus, we combined NNDP and RP with other visualization models such as the PCA and UMAP. The NNDP and RP visualization methods are named NNDP-IN-PCA, NNDP-OUT-PCA, NNDP-Word2Vec-PCA, NNDP-GloVe-PCA, NNDP-FastText-PCA, NNDP-ELMo-PCA, NNDP-GPT-2-PCA, NNDP-ALBERT-PCA, NNDP-IN-UMAP, NNDP-OUT-UMAP, NNDP-Word2Vec-UMAP, NNDP-GloVe-UMAP, NNDP-FastText-UMAP, NNDP-ELMo-UMAP, NNDP-GPT-2-UMAP, NNDP-ALBERT-UMAP, RP-PCA, and RP-UMAP. PCA and UMAP can also be used to reduce the dimension of each 6400-dimensional SNP data set to 128 for classification task and 2 for visualization tasks.

8.1. Data visualization results

We show in Fig. 4 and Figs. 6–8 our approach for data visualization on $X'_{n\times M}$ estimated via the NNDP-based methods (Fig. 4 is for NNDP-IN and -OUT, Fig. 6 is for NNDP-Word2Vec, -GloVe, and -FastText, Fig. 7 is for NNDP-ELMo, -GPT-2, and Fig. 8 is for NNDP-ALBERT[d] and -ALBERT[i]), and the traditional approaches based on UMAP and PCA without data transformation (e.g., $X_{n\times d}$)) in Fig. 5.



(a) PCA for race (b) PCA for MI event age group (c) PCA for sex

(d) UMAP for race (e) UMAP for MI event age group (f) UMAP for sex

(g) RP-PCA for race (h) RP-PCA for MI event age group (i) RP-PCA for sex

(j) RP-UMAP for race (k) RP-UMAP for MI event age group (l) RP-UMAP for sex

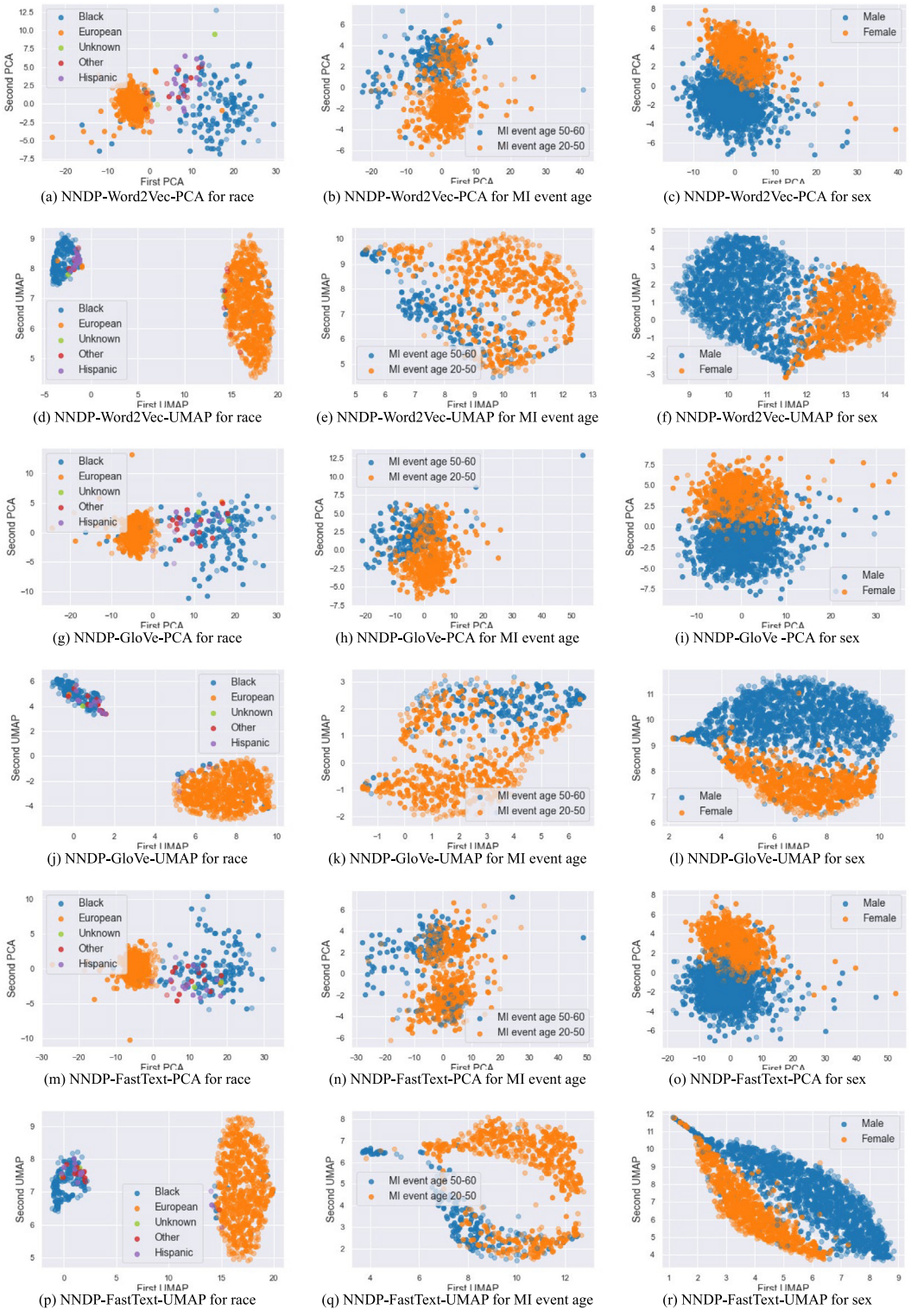
Fig. 5. Visualization results using other methods (basic PCA/UMAP, RP-PCA, RP-UMAP) for race, MI event age group, and sex categories.

8.1.1. NNDP-IN and -OUT vs. PCA, UMAP, and RP

In Fig. 4, we show 2-D visualization of clustering based on the application of both PCA and UMAP on the NNDP-IN data (the first two rows) consisting of three categories: race (left column), MI event age group (middle column), and sex (right column). The same procedures are also used for NNDP-OUT (the last two rows) in Fig. 4. In Fig. 5, we show 2-D visualization based on both PCA (1st row) and UMAP (second row) on the data itself (without NNDP-IN or NNDP-OUT) consisting of three categories: race (left column), MI event age group (middle column), and sex (right column). The third and fourth rows of Fig. 5 represent PCA and UMAP, respectively, on the data after they are transformed using RP. In these two figures, training/test data share the same color for each class but the training set's color is darker and the test data set's color is lighter. Our proposed NNDP-IN method (consisting of NNPD-IN-PCA and NNDP-IN-UMAP) outperformed all other methods including NNDP-OUT-PCA and NNDP-OUT-UMAP as well as PCA, UMAP, RP-PCA, and RP-UMAP as there are distinct clusters separating different races, MI event age groups, and sexes with our approach. For example, both NNDP-IN methods (NNDP-IN-PCA and NNDP-IN-UMAP) show clear clusters separating Blacks from Europeans for the race category, young MI events from old MI events, and males from females.

The NNDP-OUT which uses V matrix (the output weights) did not provide good results especially for the MI event age group and sex when compared to NNDP-IN. Unlike NNDP-IN, there are no distinct clusters separating different MI event age groups and the sex category. This is because NNDP-OUT has lower correlations between word representations; only gene name vectors are used to create the V matrix so there are fewer relationships between the gene name vectors and other word vectors in the V matrix. Therefore, this leads to worse performance of NNDP-OUT for the sex and MI event age group categories. Note, however, that good performance of NNDP-OUT for the race category can be seen.

As shown in Fig. 5, PCA transformation on the data itself shows good cluster separation between groups for two categories—race and sex (1st row). However, its UMAP transformation counterpart shown in the 2nd row of Fig. 5 was not able to provide good cluster separation when compared to PCA for race and sex categories. This is expected, as PCA is designed to work well in preserving global structures whereas UMAP works better for unraveling local structural dynamics (Sakaue et al., 2020). Transformation of the data with RP using Gaussian distribution followed by either PCA or UMAP for 2-D visualization are shown in the 3rd and 4th rows, respectively, of Fig. 5, for three categories. For RP with PCA, there is good clustering separation for race, but it does poorly with the



(a) NNDP-Word2Vec-PCA for race (b) NNDP-Word2Vec-PCA for MI event age (c) NNDP-Word2Vec-PCA for sex

(d) NNDP-Word2Vec-UMAP for race (e) NNDP-Word2Vec-UMAP for MI event age (f) NNDP-Word2Vec-UMAP for sex

(g) NNDP-GloVe-PCA for race (h) NNDP-GloVe-PCA for MI event age (i) NNDP-GloVe -PCA for sex

(j) NNDP-GloVe-UMAP for race (k) NNDP-GloVe-UMAP for MI event age (l) NNDP-GloVe-UMAP for sex

(m) NNDP-FastText-PCA for race (n) NNDP-FastText-PCA for MI event age (o) NNDP-FastText-PCA for sex

(p) NNDP-FastText-UMAP for race (q) NNDP-FastText-UMAP for MI event age (r) NNDP-FastText-UMAP for sex

Fig. 6. Visualization results using conventional embedding methods-NNDP-Word2Vec-PCA, NNDP-Word2Vec-UMAP, NNDP-GloVe-PCA, NNDP-GloVe-UMAP, NNDP-FastText-PCA, and NNDP-FastText-UMAP for race (a, d, g, j, m, p), MI event age group (b, e, h, k, n, q), and sex (c, f, i, l, o, r) categories.

MI event age and sex groups. For RP with UMAP, separation of clusters among race, MI event age group, and sex is nonexistent.

8.1.2. NNDP-IN and -OUT vs. NNDP-Word2Vec, -GloVe, -FastText, -ELMo, -GPT-2, and -ALBERT

We show a 2-D visualization of clustering based on the application of both PCA and UMAP on the NNDP-Word2Vec (1st and 2nd rows), NNDP-GloVe (3rd and 4th rows), and NNDP-FastText (5th and 6th rows) in Fig. 6, NNDP-ELMo (1st and 2nd rows) and NNDP-GPT-2 (3rd and 4th rows) in Fig. 7, and NNDP-ALBERT[d] (1st and 2nd rows) and NNDP-ALBERT[i] (3rd and 4th rows) in Fig. 8 for three categories: race (left column), MI event age group (middle column), and sex (right column). In Fig. 4, 6–8, all NNDP embedding variants provided good separation between Black and

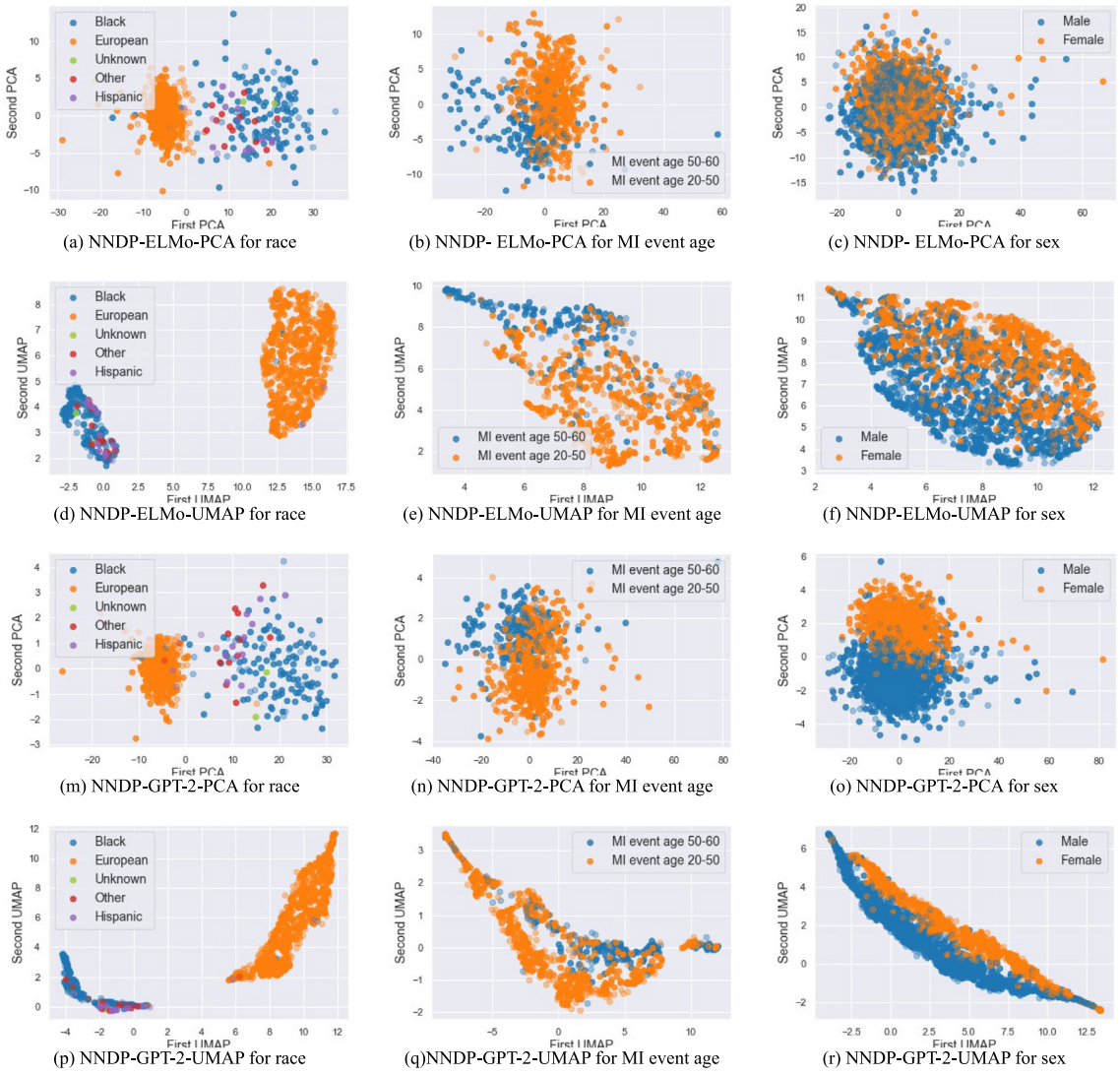


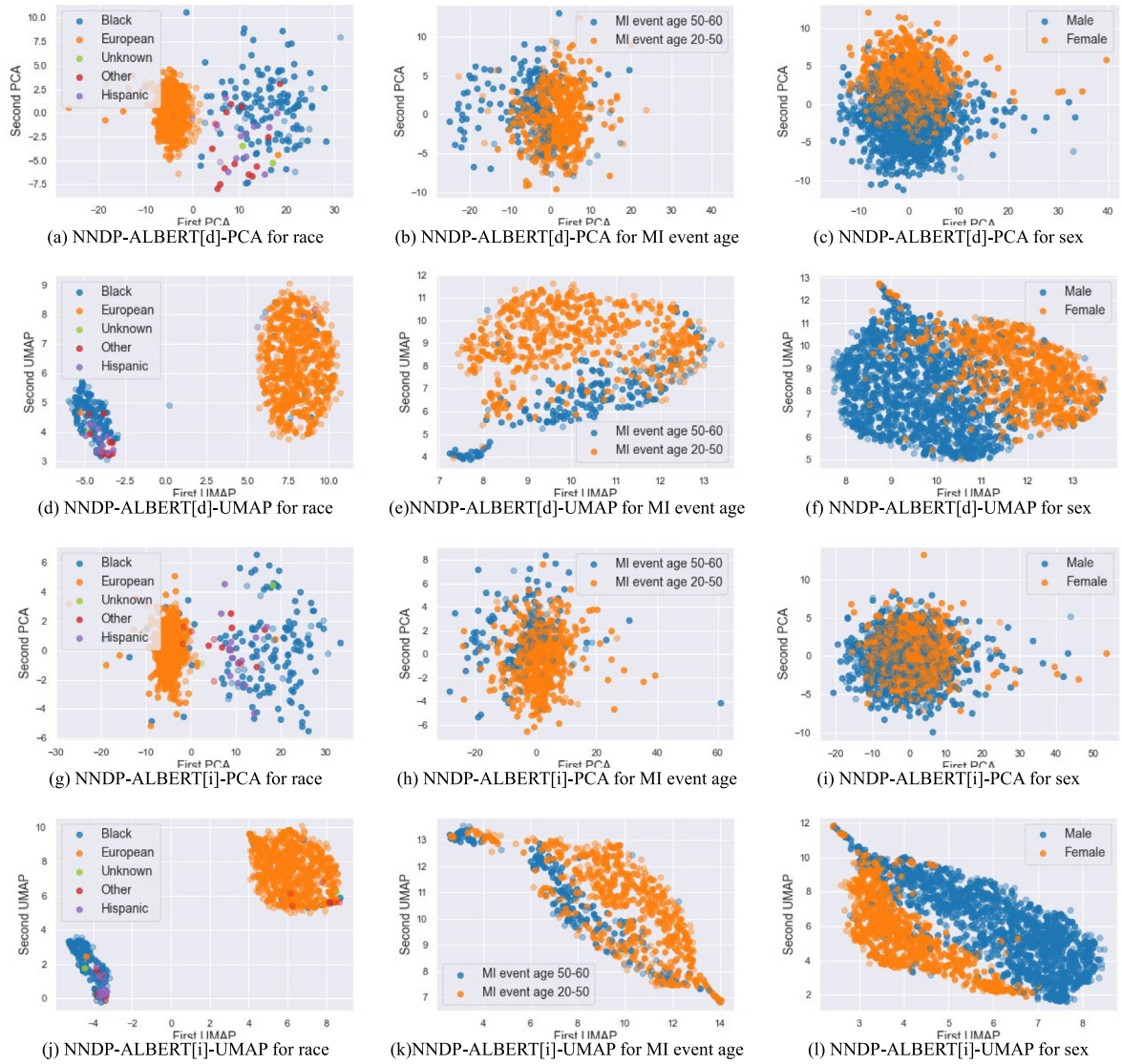
Fig. 7. Visualization results using contextual embedding methods-NNDP-ELMo-PCA, NNDP-ELMo-UMAP, NNDP-GPT-2-PCA, and NNDP-GPT-2-UMAP for race (a, d, g, j), MI event age group (b, e, h, k), and sex (c, f, i, l) categories.

European classes, similar to NNDP-IN and NNDP-OUT. However, for the MI event age and sex categories, both PCA and UMAP on the NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i] showed less clear separations for the young MI event vs. old MI event, and male vs. female when compared to NNDP-IN cluster separations, as shown in Fig. 4. In particular, NNDP-ELMo and ALBERT[i]-PCA performed worst in discriminating male vs. female classes (1st row, third column in Fig. 7, and 2nd row, third column in Fig. 8), respectively. However, the UMAP on NNDP-ALBERT[i] (4th row, third column) nicely separated two sex classes, which suggests that its embedding representations preserved some degree of the correlations between gene names. However, the class separations in the MI event and sex categories using the embedding models were all visually less impressive than those of NNDP-IN.

In order to investigate the feasibility of the effects of literature information on MI risks by our NNDP-PCA and NNDP-UMAP, we additionally examined if MI at different age groups are different or can be separated between Blacks and Europeans. We examined two age groups when MI occurred for the subjects. The age groups are defined as between young (age ranges from 20 to 50) and old (age ranges from above 50 to 60) groups for both races. Fig. 9 shows visualization results of PCA and UMAP on NNDP-IN, NNDP-OUT, and RP, and PCA and UMAP transformations on the data themselves. Fig. 10 shows the visualization results for PCA and UMAP via **conventional** embedding methods—NNDP-Word2Vec, NNDP-GloVe, and NNDP-FastText. Figs. 11–12 show the visualization results transformed by PCA and UMAP via **contextual** embedding methods—NNDP-ELMo, NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i].

The first row and first column of Fig. 9 results are based on NNDP-IN-PCA whereas the same row with the second column is via NNDP-IN-UMAP. The NNDP-IN with either PCA or UMAP appears to separate two MI event age groups by race. There are predominantly four clusters that are well separated between Europeans and Blacks and the two age groups of MI for both races. The other three approaches, NNDP-OUT with either PCA or UMAP (second row), PCA (3rd row, first column), UMAP (3rd row, second column), and RP with either PCA or UMAP (4th row) do not provide as good cluster separation as seen with either NNDP-IN-PCA or NNDP-IN-UMAP in Fig. 9.

Fig. 10 shows PCA and UMAP results via **conventional embedding methods**—NNDP-Word2Vec, NNDP-GloVe, and NNDP-FastText, while Fig. 11 shows PCA and UMAP results via **contextual embedding methods**—NNDP-ELMo, NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i], respectively. We found that all NNDP-based embedding variant methods provided better



(a) NNDP-ALBERT[d]-PCA for race (b) NNDP-ALBERT[d]-PCA for MI event age (c) NNDP-ALBERT[d]-PCA for sex

(d) NNDP-ALBERT[d]-UMAP for race (e)NNDP-ALBERT[d]-UMAP for MI event age (f) NNDP-ALBERT[d]-UMAP for sex

(g) NNDP-ALBERT[i]-PCA for race (h) NNDP-ALBERT[i]-PCA for MI event age (i) NNDP-ALBERT[i]-PCA for sex

(j) NNDP-ALBERT[i]-UMAP for race (k)NNDP-ALBERT[i]-UMAP for MI event age (l) NNDP-ALBERT[i]-UMAP for sex

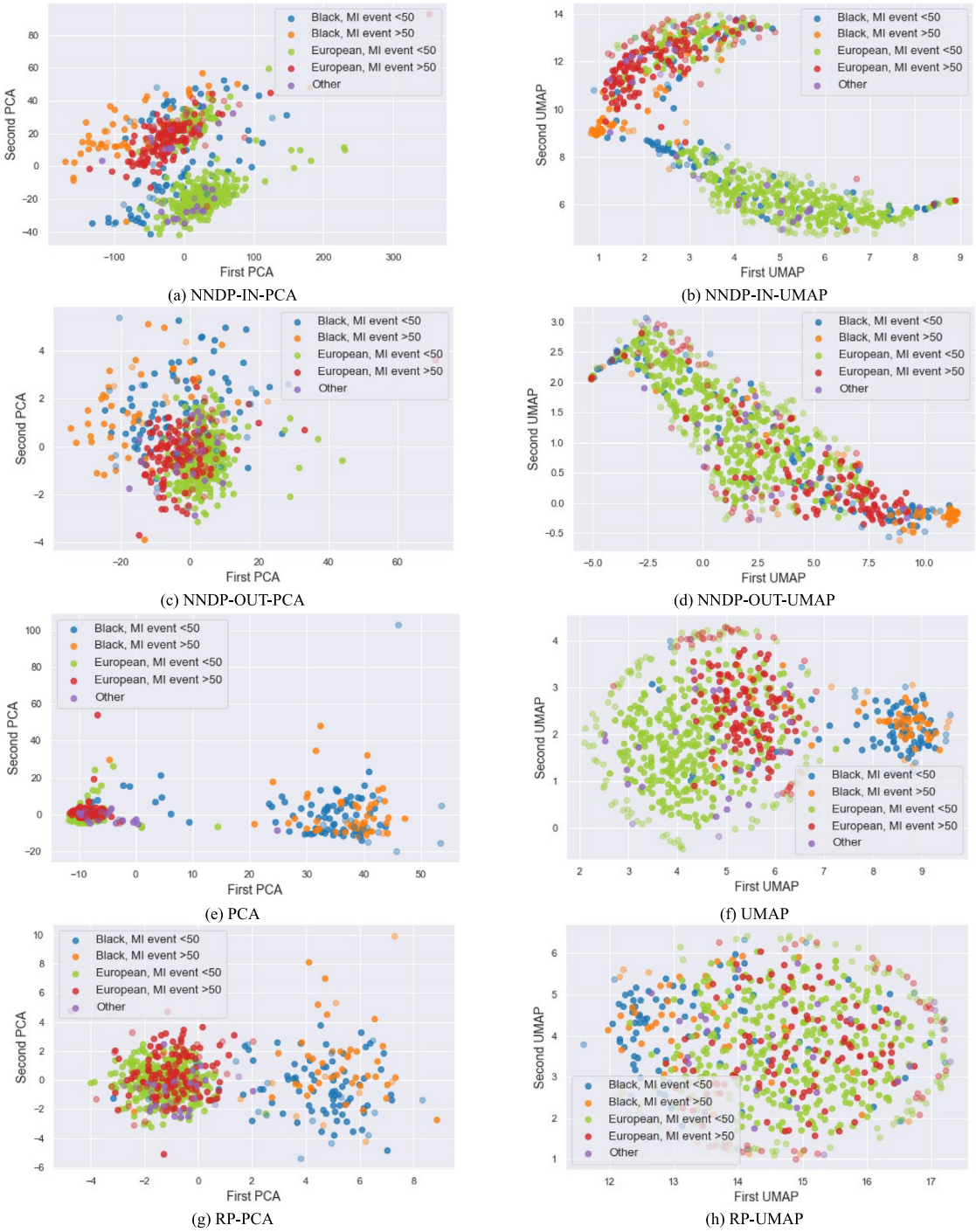
Fig. 8. Visualization results on contextual embedding methods-NNDP-ALBERT[d]-PCA, NNDP-ALBERT[d]-UMAP, NNDP-ALBERT[i]-PCA, and NNDP-ALBERT[i]-UMAP for race (a, d, g, j), MI event age group (b, e, h, k), and sex (c, f, i, l) categories.

separations between Europeans and Blacks and the two age groups of MI for both races than did the sole use of UMAP and PCA on the data visualization. The use of PCA on NNPD-Word2Vec showed similarly good separations between the clusters with NNDP-IN-PCA, but the use of UMAP on NNDP-IN showed even better separations than did NNDP-Word2Vec-UMAP. In summary, Figs. 4–11 indicate that NNDP-IN provides the best visual cluster separation among different classes when compared to all different embedding approaches.

Additionally, an interesting observation is that the distance between young and old Black subjects' MI prevalence is closer to each other than are the Europeans age groups' counterparts, especially seen with NNDP-IN-UMAP (the first row, second column of Fig. 9), NNDP-Word2Vec-UMAP (the first row and second column of Fig. 10), NNDP-GPT-2-UMAP (the second row and second column of Fig. 11), and NNDP-ALBERT[i]-UMAP (the third row and second column of Fig. 11). The implication of this is that a younger Black group may have similar MI risks to the older group. In literature, Black patients presenting with MI tend to be younger than other races (Garcia et al., 2021). Considering that our NNDP method involves better visualization results, as shown in Figs. 4–8, the result that the younger Black MI age group is closely clustered with the older Black MI age group, may suggest that they were affected by unknown risks in the literature word embeddings. Since socioeconomic status has been widely reported in the literature as the most significant factor in the high prevalence of MI in young Black (Garcia et al., 2021), the hidden risks might contain some degree of socioeconomic-related information obtained from literature.

8.2. Evaluation results of semantic correlations in word embeddings

To validate and interpret visualization results, we also examined the textual semantic analysis in literature word embeddings. We calculated the cosine similarities of the following words: 'heart, african + american, european + american, male + female' with all words (241,315 unique words) using each embedding model (NNDP-IN, NNDP-OUT, NNDDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT). For literature embedding models (IN and OUT matrices), Word2Vec, GloVe, and FastText, 241,315 unique word embedding vectors were extracted directly from their pre-trained vocabulary



(a) NNDP-IN-PCA (b) NNDP-IN-UMAP

(c) NNDP-OUT-PCA (d) NNDP-OUT-UMAP

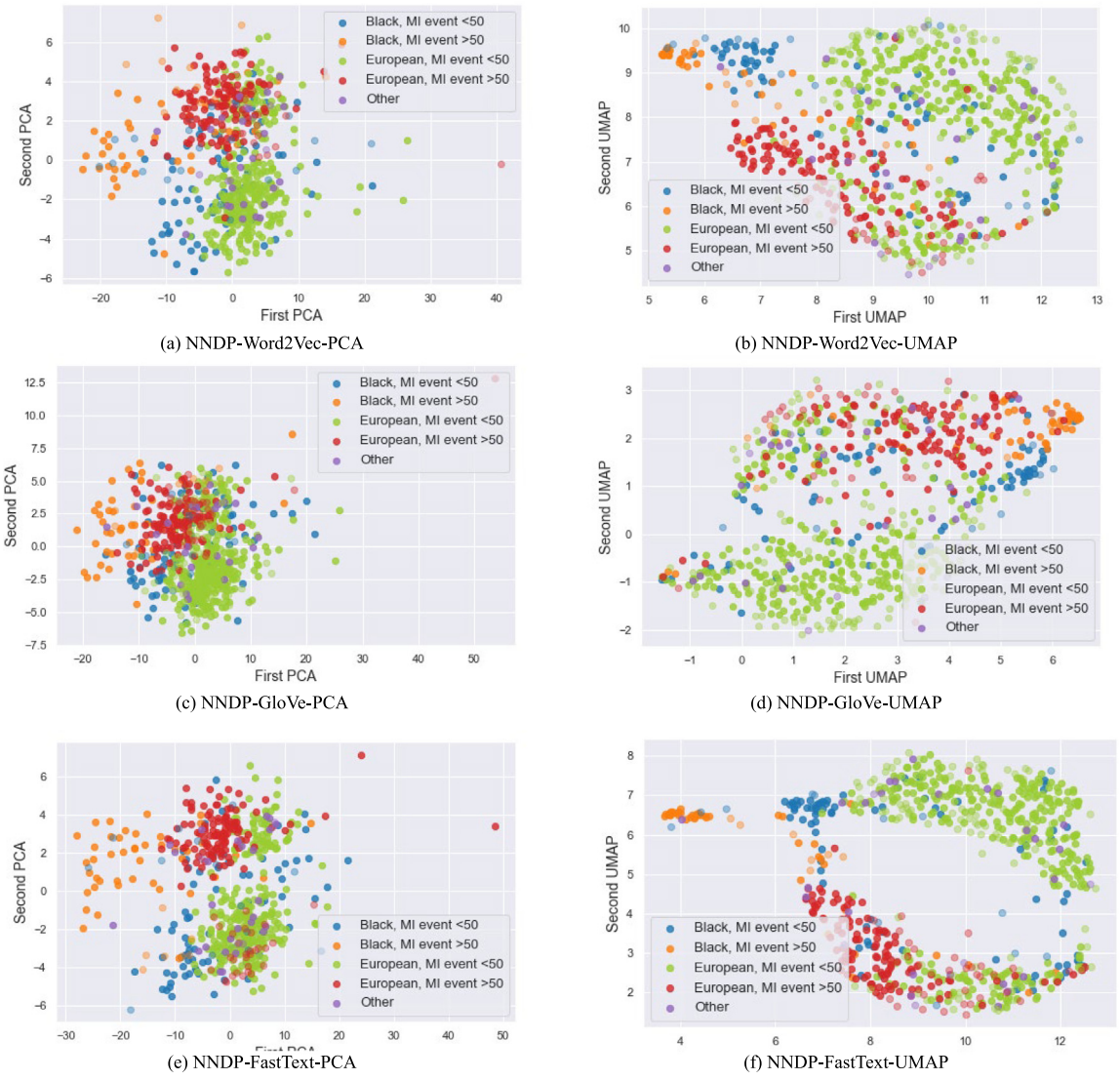
(e) PCA (f) UMAP

(g) RP-PCA (h) RP-UMAP

Fig. 9. NNDP-IN-PCA, NNDP-IN-UMAP, NNDP-OUT-PCA, NNDP-OUT-UMAP, PCA UMAP, RP-PCA, RP-UAMP methods for Young European MI, old European MI, young Black MI, old Black MI, and other class visualizations.

elements. However, ALBERT has only 30,000 word-piece vocabulary elements so we additionally extracted 241,315 unique word embedding vectors using various combinations of 30,000-word piece embedding vectors from the pre-trained ALBERT for the analysis of semantic correlations. For example, a complete unique word 'heartburn' was dissembled into 'heart' and '##burn' by the ALBERT's word piece tokenizer. Then we averaged the 'heart' and '##burn' to obtain a 'heartburn' unique word vector. All models' cosine similarities were computed using Eq. (7). Since GPT-2 has only a 50,257 sub-word vocabulary, GPT-2's 241,315 unique word embedding vectors were computed using various combinations of 50,257 sub-word embedding vectors from the pre-trained GPT-2. ELMo's 241,315 unique word embedding vectors were also extracted using the combinations of the character embedding vectors from the pre-trained ELMo. We sorted top-10 similar words with their similarity score for each query, as shown in Tables 3–4.

As shown in Table 3(a), the U matrix (NNDP-IN) provides semantically related terms for all queries. For 'heart', heart-related abbreviations, for example, vsmcs (vascular smooth muscle cells), and cardiac genetic symbols (mir101a (Pan et al., 2012), barx1 (Gould & Walter, 2000), ddi2 (van der Ende et al., 2018), and nkx2-1 (Yin et al., 2006)) were obtained. The 'hypertrophy' is not



(a) NNDP-Word2Vec-PCA (b) NNDP-Word2Vec-UMAP

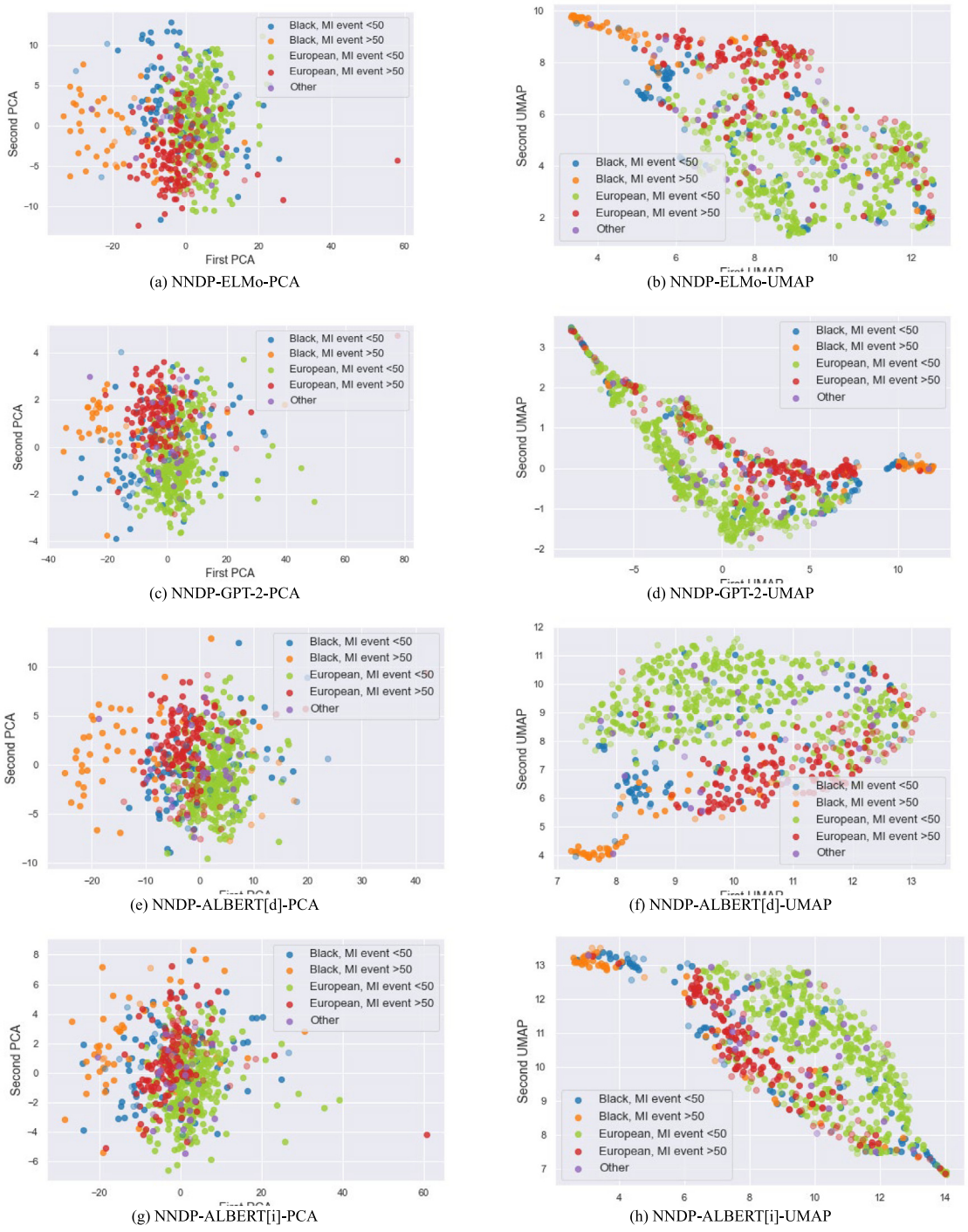
(c) NNDP-GloVe-PCA (d) NNDP-GloVe-UMAP

(e) NNDP-FastText-PCA (f) NNDP-FastText-UMAP

Fig. 10. NNDP-Word2Vec-PCA, NNDP-Word2Vec-UMAP, NNDP-GloVe-PCA, NNDP-GloVe-UMAP, NNDP-FastText-PCA, and NNDP-FastText-UMAP methods for Young European MI, old European MI, young Black MI, old Black MI, and other class visualizations.

Table 3
Query associated top-10 words search using U and V (IN matrix of our model for NNDP-IN and OUT MATRIX of our model For NNDP-OUT).

	"heart"	Score	"african + american"	Score	"european + american"	Score	"male + female"	Score
(a) NNDP-IN (used IN-Matrix U)	**heart**	**1.00**	**african**	**0.62**	**european**	**0.66**	**female**	**0.69**
	cardiac	**0.72**	**american**	**0.62**	**american**	**0.66**	**male**	**0.69**
	ventricular	**0.58**	**association**	**0.39**	**gwas**	**0.47**	**reproductive**	**0.43**
	cardiomyocytes	**0.50**	plce	0.39	**association**	**0.45**	**fertility**	**0.43**
	hypertrophy	**0.45**	#10	0.38	**associations**	**0.43**	**sperm**	**0.43**
	mir101a	**0.44**	**gwas**	**0.38**	#10	0.43	**spermatozoa**	**0.40**
	barx1	**0.44**	**associations**	**0.38**	**population**	**0.43**	**infertile**	**0.40**
	nkx2-1	**0.42**	**population**	**0.37**	**genetic**	**0.42**	#ZBTB32	**0.38**
	ddi2	**0.41**	**genetic**	**0.37**	**loci**	**0.39**	**fertilization**	**0.38**
	vsmcs	**0.41**	**european**	**0.37**	mtefts	0.39	**testis**	**0.38**
(b) NNDP-OUT (used OUT-Matrix V)	**heart**	**1.00**	**american**	**0.99**	**american**	**0.99**	**male**	**0.99**
	cntfets	0.98	**african**	**0.99**	**european**	**0.99**	**female**	**0.99**
	aoah-#	0.98	circ_0043278	0.98	**population-wise**	**0.98**	brahmi	0.98
	gbmscs	0.98	muscle-#	0.98	microrna-34	0.98	pyridin-3-yl	0.98
	sub-saharan	0.98	l104a	0.98	bio-macromolecules	0.98	pneumothorax	0.98
	parker	0.98	flrt1	0.98	ichthyosis-causative	0.98	#ASIP1	0.98
	cdc13	0.98	prrxl1	0.98	14z-eicosatetraenoic	0.98	bc034767	0.98
	hfq	0.98	coaxially	0.98	retgcs	0.98	cgnp	0.98
	single-transmembrane	0.98	igf2r	0.98	pallister-killian	0.98	#AVPR2	0.98
	aries_v4	0.98	#FIP1	0.98	#rs3858145	0.98	#UGT2A3	0.98



(a) NNDP-ELMo-PCA (b) NNDP-ELMo-UMAP

(c) NNDP-GPT-2-PCA (d) NNDP-GPT-2-UMAP

(e) NNDP-ALBERT[d]-PCA (f) NNDP-ALBERT[d]-UMAP

(g) NNDP-ALBERT[i]-PCA (h) NNDP-ALBERT[i]-UMAP

Fig. 11. NNDP-ELMo-PCA, NNDP-ELMo-UMAP, NNDP-GPT-2-PCA, NNDP-GPT-2-UMAP, NNDP-ALBERT[d]-PCA, NNDP-ALBERT[d]-UMAP, NNDP-ALBERT[i]-PCA, and NNDP-ALBERT[i]-UAMP methods for Young European MI, old European MI, young Black MI, old Black MI, and other class visualizations.

directly related to heart, however, the hypertrophy is often referred as 'hypertrophy cardiomyopathy'. For 'african + european', 8 words semantically related to the query were captured: association, gwas, associations, population, genetic, and european. We collected literature data based on gene names, and the genome study is called genome-wide association studies (GWAS). The GWAS is biased towards European populations in a risk prediction study (Peterson et al., 2019). Hence, current studies have aimed to estimate accurate disease risks for other races such as African Americans (Peterson et al., 2019). For the 'european + .american', 8 words that are semantically related to the query were captured. The query shared the same 7 words with 'african + .american' since the 'african' and 'european' have high semantic similarity as a racial category. The 'loci' is a term used in genetics. For 'male + female', the U matrix showed sex-related words–reproductive, fertility, sperm, spermatozoa, infertile, fertilization, testis, and sex-related gene ZBTB32. The ZBTB32 is known as testis zinc finger protein (TZFP) that has a significant role in adult testis

(Van Wyngene et al., 2021). In contrast, as shown in Table 3(b), the V matrix (NNDP-OUT) provides randomized words and similar scores for all captured words. These results suggest that the V matrix does not represent finer correlations between words as we expected. The results can also explain why NNDP-OUT provides inferior performance when compared to NNDP-IN, as shown in Figs. 4 and 9.

Table 4(a) shows that Word2Vec (NNDP-Word2Vec) clusters categorically and semantically similar words closer. For 'heart', categorically similar words such as 'kidney', 'eye', 'liver', and 'organ' were well represented, as were semantically related words such as 'cardiac', 'ventricular', 'myocardium', 'cardiovascular', and 'myocardial'. For 'african + american' and 'european + american', mostly race-related words were captured as they are bolded. For 'male + female', age-related words also appeared: 'adult', 'offspring', '3-month-old', 'young', 'month-old'. These good similarity results could explain why NNDP-Word2Vec has good cluster separation for all categories, as shown in Figs. 6 and 10. However, NNDP-Word2Vec showed smaller cluster separations for MI events and sex categories when compared to NNDP-IN in Figs. 4 and 9. We can find the reasons via Table 3(a) and 4(b): Table 3(a) (for NNDP-IN) shows topically related words for each query while Table 4(a) (for NNDP-Word2Vec) shows categorically related words for each query. Note that a literature embedding model generates embedding vectors by predicting a gene name using associated documents that represent topically related correlations between words, while Word2Vec generates embedding vectors by predicting a center word using context words that represent categorically related correlations between word. The categorically strong representations may reduce the priority for gene-gene correlations for the data visualization, so the cosine similarity result could explain why NNDP-IN is superior to NNDP-Word2Vec for cluster separations.

Table 4(b) shows GloVe (NNDP-GloVe) results for each query. As shown in Table 4(b), similar words appeared to be randomly selected, but the bolded words are co-occurrent with each query keyword. For the first word column: 'dallas (heart dallas)', 'failure-relevant (heart failure-relevant)', 'failure-related (heart failure-related)', 'murmur (heart murmur)', and 'conotruncal (conotrucal heart)' were correctly captured based on the word co-occurrence probability. Moreover, 'jogging' causes changes in heart rate, and 'tsk-1ko' is reported to modulate pressure overload-induced cardiac remodeling (Duan et al., 2020). For the second column words: sweetgum (American sweetgum), diaspora (African diaspora), bulldog (American bulldog), and non-finnish (non-finnish European) were also correctly captured. The 'eurasier' is a dog breed originating from Europe. The 'folklores' is a cultural matter that relates to racial information. For 'male + female', some sex-related genes were also captured: 'hnsc-h9' is a female human neural stem cell (Pottmeier et al., 2020) and cfap69 is an infertility-related gene (Dong et al., 2018). Some topically related word such as 'infertility' and 'neutered' were also captured. Symbol-related words also appeared with 'sex' (i.e., sex symbol). However, the results have no specific patterns (it is not clear whether each word vector was created topically or categorically), hence, these results could explain why NNDP-GloVe-PCA provided worse cluster separations than did NNDP-IN and NNDP-Word2Vec for the MI event age and sex categories, as shown in Figs. 4, 6, 9, and 10.

Table 4(c) shows that FastText (NNDP-FastText) adequately clusters categorically and semantically similar words, but it also captured sub-word-related words. For 'heart', most captured words include 'heart' as well as a semantically similar word 'myocardium'. For 'african + american', 'european + american' and 'male + female', all captured words include sub-words of the query word such as 'afro-american', 'latino-american', and 'f2gfemale'. The sub-word biased results could explain why NNDP-FastText shows more thin shapes of separated clusters for the sex category while unique-word-based embeddings–NNDP-IN, NNDP-Word2Vec, and NNDP-GloVe show wider shapes of separated clusters as shown in Figs. 4, 6, 9, and 10.

Table 4(d) shows that ELMo (NNDP-ELMo) captures semantically and categorically related words as well as sub-word-related words for each query. For 'heart', we show that categorically related words (kidney, eye, brain, and liver as a human body category) and topically related words such as peri-ictal and epicardial (epicardial pacemaker) were correctly captured. However, it contains not-related words such as 'semi-soli' and 'dent'. For 'african + american' and 'european + american', most race-based words were captured. 'wean', 'han', and 'hessian' were also additionally captured but they are not related to the query word. However, considering that both queries contain '-an' at the end of characters, we expected this because of the sub-word information represented by character-based convolutional filters of ELMo. For 'male + female', topically related words such as 'low-fertility' and sex-related risk factors were captured: 'hcc-bearing' as androgen increases the risk of hcc (hepatocellular carcinoma) (Lee et al., 2019) and 'pcd-affected' as pcd (primary ciliary dyskinesia) affects male infertility (Jayasena & Sironen, 2021). Considering ELMo captures contextual information, it provides more topical information when compared with Word2Vec, GloVe, and FastText. However, ELMo shows some collected words that may be related to each other ('march', 'november'), but they are unrelated to input query words (i.e., 'african + european'). This is because bidirectional LSTMs capture information sequentially from left-to-right and right-to-left, and ELMo was trained with fixed-length sentences as its input. However, this work extracted gene name-based representations by feeding each gene name (a word) to ELMo, and the padding was added next to the word to reach the fixed-length input of ELMo. Hence, the ELMo structure may have corrupted information of gene-name-based representation. These results could explain why ELMo provides the worst separation of clusters for MI event and sex categories when compared to other NNDP approaches (except NNDP-OUT), as shown in Figs. 4–11.

Table 4(e) shows that GPT-2 (NNDP-GPT-2) captures mostly topically similar words as well as sub-word-based words. For 'heart', heart-related risk factors and terms were captured: h2afz (H2 A.z), hdsp (heart disease and stroke program), hmgbs (high mobility group boxes), and hffs (heart failure survival score). H2 A.z is related to cardiac myocyte hypertrophy (Chen et al., 2006), and the high mobility group box 1 may aid in treatment of cardiovascular diseases (Raucci et al., 2019). Categorically related hbmec (human brain microvascular endothelial cells) and hleb3 (hydrogen peroxide-treated lens epithelial cells) as a human body category were also captured. On the other hand, 'hot' and 'hatched' are not related to heart. However, 'hot' often appeared with 'heart', and there is a term 'HATCH score' related to heart failure prediction (Shibata et al., 2022). Interestingly, all captured words included a character 'h'. We expect this is because of GPT-2's BPE tokenizer. The BPE replaces the most common pair of consecutive chars with a unique character that does not appear in the training data. We expect pairing these consecutive chars may provide sub-word-biased representations. For 'african + american', 'european + american', and 'male + female', all captured words included sub-words of the query word such as 'american-specific', 'european-derived', and 'female-smokers'. The results could explain why NNDP-GPT-2-UMAP shows more thin shapes of visualized clusters for all categories similar to the visualization results of NNDP-FastText, as shown in Figs. 6, 7, 10, and 11. The results of sub-word-biased representations also may explain why NNDP-GPT-2-UMAP shows less separations for female and male categories, as shown in Figs. 7 and 11.

Table 4
Query associated top-10 words search using Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT.

	"heart"	Score	"african + american"	Score	"european + american"	Score	"male + female"	Score
(a) NNDP-Word2Vec	**heart**	**1.00**	**american**	**0.76**	**american**	**0.80**	**female**	**0.89**
	kidney	**0.68**	**african**	**0.76**	**european**	**0.80**	**male**	**0.89**
	cardiac	**0.66**	**european**	**0.66**	**asian**	**0.63**	**adult**	**0.60**
	ventricular	**0.59**	**asian**	**0.63**	**african**	**0.63**	**males**	**0.59**
	myocardium	**0.59**	**indian**	**0.62**	**caucasian**	**0.61**	**females**	**0.58**
	cardiovascular	**0.55**	**latino**	**0.60**	**latino**	**0.61**	**offspring**	**0.58**
	myocardial	**0.53**	**caucasian**	**0.60**	**japanese**	**0.59**	**3-month-old**	**0.57**
	eye	**0.53**	**african-american**	**0.59**	**african-american**	**0.59**	**young**	**0.57**
	liver	**0.52**	**south**	**0.58**	**afro-caribbean**	**0.58**	**month-old**	**0.56**
	organ	**0.52**	l56rfs	0.57	**australian**	**0.58**	isfahan	0.56
(b) NNDP-GloVe	**heart**	**1.00**	**american**	**0.83**	**american**	**0.83**	**male**	**0.95**
	healthy-type	**0.92**	**african**	**0.83**	**european**	**0.83**	**female**	**0.95**
	dallas	**0.91**	#5681	0.78	#1794	0.82	**hnsc-h9**	**0.95**
	failure-relevant	**0.88**	#1794	0.77	#5681	0.80	**cfap69-knockout**	**0.92**
	jogging	**0.82**	**sweetgum**	**0.77**	**sweetgum**	**0.79**	**neutered**	**0.92**
	failure-related	**0.82**	143e−02	0.75	**folklores**	0.75	**symbolt**	**0.92**
	murmur	**0.81**	**folklores**	0.73	143e−02	0.74	hypo-s-dmrs	0.92
	tsp-1ko	**0.79**	**diaspora**	**0.73**	**non-finnish**	**0.73**	**symbolc**	**0.91**
	conotruncal	**0.78**	**bulldog**	**0.71**	**lines-african**	**0.72**	**infer-tility**	**0.91**
	arvms	**0.77**	l56rfs	0.70	**eurasier**	**0.72**	3244-g	0.91
(c) NNDP-FastText	**heart**	**1.00**	**african**	**0.87**	**european**	**0.85**	**female**	**0.96**
	canheart	**0.80**	**american**	**0.87**	**american**	**0.85**	**male**	**0.96**
	hearth	**0.80**	**west-african**	**0.80**	**euro-american**	**0.80**	**f2gfemale**	**0.87**
	cardiac	**0.76**	**euro-american**	**0.80**	**african**	**0.76**	**pfemale**	**0.81**
	brain-heart	**0.75**	**afro-american**	**0.78**	**latino-american**	**0.76**	**male–female**	**0.79**
	heartworm	**0.75**	**latino-american**	**0.78**	**european-american**	**0.75**	**male-male**	**0.77**
	heart-#	**0.73**	**african-american**	**0.77**	**north-american**	**0.75**	**females**	**0.76**
	life-heart	**0.73**	**non-african**	**0.76**	**afro-american**	**0.74**	**orfemale**	**0.75**
	myocardium	**0.73**	**latin-american**	**0.76**	**indo-european**	**0.74**	**males**	**0.75**
	heartburn	**0.73**	**americans**	**0.75**	**european-australian**	**0.74**	**xx-male**	**0.73**
(d) NNDP-ELMo	**heart**	**1.00**	**african**	**0.90**	**european**	**0.87**	**male**	**0.95**
	kidney	**0.83**	**american**	**0.90**	**american**	**0.87**	**female**	**0.95**
	eye	**0.78**	**mexican**	**0.86**	**african**	**0.83**	**pfemale**	**0.79**
	brain	**0.77**	**asian**	**0.84**	**asian**	**0.83**	**hcc-bearing**	0.78
	peri-ictal	**0.74**	**han**	**0.83**	**mexican**	**0.82**	**pmale**	**0.78**
	hearth	**0.73**	september	0.82	**han**	**0.80**	baf-a1-treated	0.78
	liver	**0.73**	march	0.81	**caucasian**	**0.80**	**low-fertility**	**0.78**
	episcleral	**0.72**	**caucasian**	**0.80**	**hessian**	**0.77**	**pcd-affected**	**0.77**
	semi-solid	0.72	november	0.80	kangaroo	0.77	riboflavin-unresponsive	0.77
	dent	0.72	**european**	**0.79**	**wean**	**0.77**	obese	0.77
(e) NNDP-GPT-2	**heart**	**1.00**	**american**	**0.79**	**american**	**0.76**	**male**	**0.78**
	hot	**0.93**	**african**	**0.79**	**european**	**0.76**	**female**	**0.78**
	hbmec	**0.93**	**american-specific**	**0.77**	**european-derived**	**0.74**	**male-limited**	**0.75**
	hnpgs	**0.93**	**americanum**	**0.75**	**european-frequent**	**0.73**	**female-smokers**	**0.74**
	hatched	**0.93**	**african-ancestry**	**0.74**	**americanus**	**0.73**	**female-limited**	**0.73**
	hle-b3	**0.93**	**african-american**	**0.74**	**americanum**	**0.73**	**female-related**	**0.72**
	h2afz	**0.92**	**african-descent**	**0.74**	**american-specific**	**0.73**	**male-#**	**0.72**
	hdsp	**0.92**	**americanus**	**0.73**	**european-origin**	**0.72**	**female-specific**	**0.71**
	hmgbs	**0.92**	**african-americans**	**0.73**	**european-australians**	**0.71**	**female-biased**	**0.70**
	hffs	**0.92**	**african-origin**	**0.72**	**european-american**	**0.70**	**male-dominant**	**0.70**
(f) NNDP-ALBERT[d]	**heart**	**1.00**	**american**	**0.69**	**european**	**0.73**	**male**	**0.74**
	body	**0.66**	**african**	**0.69**	**american**	**0.73**	**female**	**0.74**
	pancreatic	**0.61**	**french-american-british**	**0.58**	**african-american**	**0.59**	**male–female**	**0.59**
	hearts	**0.61**	**americans**	**0.56**	**americans**	**0.56**	**male-male**	**0.48**
	chd8	**0.57**	**hispanic-americans**	**0.56**	**college**	**0.52**	**male-to-female**	**0.47**
	shr-lx	**0.57**	**african-american**	**0.55**	**french-american-british**	**0.51**	**young**	**0.43**
	epicardial	**0.56**	**east**	**0.52**	**east–west**	**0.51**	**mother-infant**	**0.43**
	dies	**0.56**	**european-american**	**0.51**	**black**	**0.50**	**malep**	**0.42**
	bone-regeneration	**0.55**	**euro**	**0.48**	**african-americans**	**0.49**	**female-smokers**	**0.41**
	whole-muscle	**0.55**	**easter**	**0.48**	**germany**	**0.49**	**maleate**	**0.38**
(g) NNDP-ALBERT[i]	**heart**	**1.00**	**african**	**0.75**	**european**	**0.74**	**male**	**0.86**
	hearth	**0.78**	**american**	**0.75**	**american**	**0.74**	**female**	**0.86**
	heartburn	**0.76**	**african-american**	**0.71**	**european-american**	**0.70**	**male–female**	**0.78**
	hearts	**0.72**	**european-american**	**0.63**	**african-american**	**0.65**	**male-male**	**0.72**
	brain-heart	**0.70**	**hispanic-american**	**0.62**	**hispanic-american**	**0.62**	**maleic**	**0.63**
	heartworm	**0.66**	**european**	**0.60**	**african**	**0.61**	**malep**	**0.62**
	life-heart	**0.62**	**latino-american**	**0.59**	**european-asian**	**0.60**	**maleate**	**0.58**
	cardiac	**0.61**	**americanum**	**0.58**	**european-australian**	**0.60**	**female-predominant**	**0.55**
	heart-directed	**0.55**	**mexican-american**	**0.58**	**latino-american**	**0.59**	**female-limited**	**0.52**
	heart-enriched	**0.55**	**asian**	**0.58**	**americanum**	**0.58**	**male-predominant**	**0.51**

Note: Semantically correlated words are bolded.

Table 5
Quantitative performance of Race, MI event age group, and sex classifications for NNDP-IN, NNDP-OUT, PCA, UMAP, and RP.

		(a) Race				(b) MI Event Age Group				(c) Sex			
Method	Model	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.
Original 64,000-dimensional data	SVM-linear	**0.96**	**0.96**	**0.99**	**0.97**	**0.79**	**0.78**	**0.81**	**0.79**	**0.99**	**1.00**	0.98	**0.99**
	SVM-rbf	0.95	0.95	0.98	0.96	0.71	0.92	0.14	0.35	**0.99**	**1.00**	**0.99**	**0.99**
	SVM-poly	0.95	0.95	0.98	0.96	0.71	0.92	0.14	0.35	**0.99**	**1.00**	**0.99**	**0.99**
	LR	0.94	0.94	0.98	0.96	0.74	**0.93**	0.25	0.46	**0.99**	**1.00**	**0.99**	**0.99**
	MLP	**0.96**	**0.96**	**0.99**	**0.97**	0.74	**0.93**	0.21	0.44	0.96	0.97	0.95	0.96
NNDP (IN)	SVM-linear	**0.95**	**0.95**	**0.98**	0.96	**0.78**	**0.87**	**0.54**	**0.68**	0.98	**0.99**	0.97	0.98
	SVM-rbf	0.94	0.94	0.98	0.96	0.76	0.91	0.36	0.57	**0.99**	**0.99**	**0.98**	**0.99**
	SVM-poly	0.94	0.94	0.98	0.96	0.76	0.91	0.36	0.57	**0.99**	**0.99**	**0.98**	**0.99**
	LR	0.94	0.94	0.98	0.96	0.77	**0.92**	0.38	0.59	**0.99**	**0.99**	**0.98**	**0.99**
	MLP	**0.95**	**0.95**	**0.98**	**0.97**	0.77	0.89	0.44	0.62	0.98	0.98	**0.98**	0.98
	*p-value	**0.43**				**0.69**				**1.12**			
NNDP (OUT)	SVM-linear	**0.95**	**0.95**	**0.98**	0.96	**0.76**	**0.89**	**0.40**	**0.59**	**0.93**	**0.90**	**0.94**	**0.92**
	SVM-rbf	0.94	0.94	**0.98**	0.96	0.72	0.91	0.20	0.42	0.91	0.86	**0.94**	0.90
	SVM-poly	0.94	0.94	**0.98**	0.96	0.72	0.91	0.20	0.42	0.91	0.86	**0.94**	0.90
	LR	**0.95**	**0.95**	**0.98**	0.96	**0.76**	**0.89**	0.39	**0.59**	**0.93**	**0.90**	**0.94**	**0.92**
	MLP	0.94	0.94	**0.98**	0.96	0.73	0.89	0.32	0.53	0.92	0.89	**0.94**	0.91
	*p-value	**0.43**				**0.20**				0.00			
PCA	SVM-linear	**0.93**	**0.93**	**0.98**	0.95	**0.66**	**0.78**	**0.37**	**0.53**	**0.89**	**0.87**	**0.90**	**0.88**
	SVM-rbf	0.22	0.22	0.74	0.40	0.28	0.01	1.00	0.07	0.69	0.11	1.00	0.32
	SVM-poly	0.22	0.22	0.74	0.40	0.28	0.01	1.00	0.07	0.69	0.11	1.00	0.32
	LR	**0.93**	**0.93**	**0.98**	0.95	**0.66**	**0.79**	**0.34**	**0.50**	**0.90**	**0.86**	**0.92**	**0.88**
	MLP	0.28	0.28	0.80	0.47	0.60	0.68	0.38	0.51	0.86	0.75	0.93	0.82
	*p-value	0.02				0.00				0.00			
UMAP	SVM-linear	**0.85**	**0.85**	**0.95**	0.90	**0.77**	**0.84**	**0.57**	**0.68**	0.62	0.34	0.77	0.51
	SVM-rbf	0.83	0.83	0.94	0.89	0.73	0.96	0.14	0.31	**0.65**	0.00	1.00	0.00
	SVM-poly	0.83	0.83	0.94	0.89	0.73	0.96	0.14	0.31	**0.65**	0.00	1.00	0.00
	LR	0.85	0.85	**0.95**	0.90	0.76	0.84	0.54	0.66	0.62	0.26	0.81	0.45
	MLP	**0.86**	**0.86**	**0.95**	0.91	0.76	0.87	0.48	0.64	**0.63**	**0.36**	**0.77**	**0.52**
	*p-value	0.00				**0.40**				0.00			
RP (Gaussian Distribution)	SVM-linear	0.94	0.94	**0.98**	0.96	**0.69**	**0.81**	**0.38**	**0.55**	0.74	0.58	0.83	0.69
	SVM-rbf	**0.95**	**0.95**	**0.98**	**0.96**	0.70	0.91	0.14	0.36	**0.78**	**0.59**	**0.88**	**0.72**
	SVM-poly	**0.95**	**0.95**	**0.98**	**0.96**	0.70	0.91	0.14	0.36	**0.78**	**0.59**	**0.88**	**0.72**
	LR	0.94	0.94	**0.98**	0.96	**0.73**	0.94	0.14	0.35	0.74	0.55	0.85	0.68
	MLP	0.94	0.94	**0.98**	0.96	**0.71**	**0.86**	**0.30**	**0.50**	0.73	0.58	0.81	0.69
	*p-value	**0.43**				0.00				0.00			

*p-value was computed based on the best accuracies between original 6400-dimensional data set and each method.

Table 4(f) shows that the sum matrix of token and positional embedding matrices of ALBERT (NNDP-ALBERT[d]) adequately captures categorically and topically similar words. For 'heart', categorically and topically related words such as pancreatic (pancreatic cardiac), epicardial (epicardial pacemaker), dies, bone-regeneration, and whole-muscles were correctly captured. Cardiac risk factor-related proteins were also captured. For example, chromodomain helicase binding protein 8 (Chd8) contributes to cardiac development (Shanks et al., 2012), and SHR-Lx reduces blood pressure and heart weight (Šeda et al., 2005). For 'african + american' and 'european + american', most of the race-related words were captured. Some of the captured words are not directly related to race but they co-appeared with 'american' including: 'easter', 'east', and 'college'. For 'female + male', as a demographic category, the categorically related word 'age' appeared. However, most query-sub-word-related words (query-word-related pieces) such as 'male-male' were present. These results could explain why ALBERT provides unimpressive cluster separations for the sex category, as shown in Fig. 8(c) and (f).

Table 4(g) shows that the token embedding matrix of ALBERT (NNDP-ALBERT[i]) clusters words mostly based on query word-related sub-words. For 'heart', query-sub-word-related words were correctly captured: 'hearth', 'heartburn', 'hearts', 'brain-heart', 'heartworm', 'life-heart', 'heart-directed', 'heart-enriched', as well as the topically related word 'cardiac'. For 'african + american' and 'european + american', word-pieced-based words were also captured as well as race related words such as 'european', 'asian', and 'african'. as shown in Table 4(c). For 'male + female', query word-piece-based words were all captured: 'male–female', 'male-male', 'maleic', 'malep', 'maleate', 'female-predominant', 'female-limited', and 'male-predominant'. The presence of a large number of sub-word-based similar words can reduce the priority for categorically or topically similar words for each query; the sub-word- based results could explain why NNDP-ALBERT[i]-PCA provided poor class separations for the MI event and the sex categories. The results also could explain how sub-word-based information is related to the visualization results when compared with NNDP-ALBERT[d]-PCA's results.

Table 6
Quantitative performance of Race, MI event age group, and sex classifications for NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT.

		(a) Race				(b) MI Event Age Group				(c) Sex			
Method	Model	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.
NNDP-Word2Vec	SVM-linear	0.95	0.95	0.98	0.96	**0.77**	**0.88**	**0.50**	**0.66**	0.97	0.98	0.96	0.97
	SVM-rbf	0.94	0.94	0.98	0.96	0.76	0.92	0.36	0.57	**0.98**	**0.98**	**0.97**	**0.98**
	SVM-poly	0.94	0.94	0.98	0.96	0.76	0.92	0.36	0.57	**0.98**	**0.98**	**0.97**	**0.98**
	LR	0.94	0.94	0.98	0.96	0.77	0.92	0.38	0.59	**0.98**	**0.99**	**0.97**	**0.98**
	MLP	**0.95**	**0.95**	**0.98**	**0.97**	0.77	0.91	0.40	0.60	0.97	0.96	0.97	0.97
	*p-value	**0.43**				**0.39**				0.00			
NNDP-GLOVE	SVM-linear	**0.95**	**0.95**	**0.98**	**0.97**	**0.77**	0.87	**0.51**	**0.66**	**0.98**	**0.99**	**0.98**	**0.98**
	SVM-rbf	0.94	0.94	0.98	0.96	0.76	0.91	0.37	0.58	**0.98**	**0.98**	**0.98**	**0.98**
	SVM-poly	0.94	0.94	0.98	0.96	0.76	0.91	0.37	0.58	**0.98**	**0.98**	**0.98**	**0.98**
	LR	0.94	0.94	0.98	0.96	**0.78**	0.92	0.41	0.61	**0.98**	**0.98**	**0.98**	**0.98**
	MLP	0.94	0.94	0.98	0.96	0.76	0.90	0.38	0.58	0.98	0.98	0.97	0.98
	*p-value	**0.43**				**0.69**				0.00			
NNDP-FastText	SVM-linear	0.94	0.94	0.98	0.96	**0.78**	**0.87**	**0.52**	**0.67**	**0.98**	**0.99**	**0.97**	**0.98**
	SVM-rbf	0.94	0.94	0.98	0.96	0.78	0.92	0.41	0.61	**0.98**	**0.98**	**0.97**	**0.98**
	SVM-poly	0.94	0.94	0.98	0.96	0.78	0.92	0.41	0.61	**0.98**	**0.98**	**0.97**	**0.98**
	LR	0.94	0.94	0.98	0.96	0.77	0.92	0.39	0.60	**0.98**	**0.98**	**0.98**	**0.98**
	MLP	0.94	0.94	0.98	0.96	0.77	0.92	0.39	0.60	0.97	0.97	0.97	0.97
	*p-value	**0.10**				**0.69**				0.00			
NNDP-ELMo	SVM-linear	0.94	0.94	0.98	0.96	0.73	0.91	0.26	0.48	0.97	0.97	0.97	0.97
	SVM-rbf	0.94	0.94	0.98	0.96	0.73	0.91	0.26	0.48	0.87	0.79	0.91	0.85
	SVM-poly	0.94	0.94	0.98	0.96	0.74	0.90	0.35	0.54	0.87	0.79	0.91	0.85
	LR	0.94	0.94	0.98	0.96	0.72	0.89	0.28	0.49	0.94	0.93	0.95	0.94
	MLP	0.94	0.94	0.98	0.96	0.76	0.89	0.40	0.59	0.91	0.87	0.93	0.90
	*p-value	**0.10**				**0.19**				0.00			
NNDP-GPT-2	SVM-linear	0.95	0.95	0.98	0.96	**0.81**	**0.90**	**0.54**	**0.70**	**0.98**	**0.99**	**0.97**	**0.98**
	SVM-rbf	0.94	0.94	0.98	0.96	0.76	0.94	0.30	0.52	0.97	0.97	0.97	0.97
	SVM-poly	0.94	0.94	0.98	0.96	0.76	0.94	0.30	0.52	0.97	0.97	0.97	0.97
	LR	0.95	0.95	0.98	0.96	0.78	0.93	0.37	0.59	**0.98**	**0.99**	**0.98**	**0.98**
	MLP	0.95	0.95	0.98	0.96	0.79	0.92	0.44	0.63	0.97	0.97	0.97	0.97
	*p-value	**0.43**				**0.37**				0.00			
NNDP-ALBERT[d]	SVM-linear	0.95	0.95	0.98	0.96	**0.78**	**0.88**	**0.53**	**0.68**	**0.98**	**0.99**	**0.98**	**0.98**
	SVM-rbf	0.94	0.94	0.98	0.96	0.78	0.93	0.38	0.59	0.97	0.97	0.98	0.97
	SVM-poly	0.94	0.94	0.98	0.96	0.78	0.93	0.38	0.59	0.97	0.97	0.98	0.97
	LR	0.94	0.94	0.98	0.96	0.79	0.92	0.45	0.64	0.98	0.97	0.98	0.97
	MLP	0.94	0.94	0.98	0.96	0.77	0.91	0.40	0.60	0.97	0.96	0.98	0.97
	*p-value	0.43				**0.69**				0.00			
NNDP-ALBERT[i]	SVM-linear	0.94	0.94	0.98	0.96	**0.79**	0.89	0.54	**0.69**	**0.98**	**0.98**	**0.98**	**0.98**
	SVM-rbf	0.94	0.94	0.98	0.96	0.77	0.92	0.36	0.57	**0.98**	0.97	0.98	0.97
	SVM-poly	0.94	0.94	0.98	0.96	0.77	0.92	0.36	0.57	**0.98**	0.97	0.98	0.97
	LR	0.94	0.94	0.98	0.96	**0.79**	0.92	0.42	0.62	**0.98**	**0.98**	**0.98**	**0.98**
	MLP	0.94	0.94	0.98	0.96	**0.79**	0.92	0.46	0.64	0.97	0.97	0.98	0.97
	*p-value	0.10				**1.05**				0.00			

*p-value was computed based on the best accuracies between original 6400-dimensional data set and each method.

NNDP-ALBERT[i]-PCA was inferior to NNDP-ALBERT[d]-PCA, as NNDP-ALBERT[i] provided more sub-word-related words for sex category. These results of two different embedding matrices of ALBERT (ALBERT[d] and ALBERT[i]) show how sub-word-biased representations could make the shape of visualized clusters thinner; this was also the case for FastText, ELMo, and GPT, as shown in Figs. 7, 10, and 11.

8.3. Classification results with dimensionality reduction

For quantitative evaluation of classification of race, MI age groups, and sex, we show various machine learning (SVM, RF, LR, and MLP) methods applied to both the original and the reduced data dimension using NNDP, PCA, UMAP, and RP. NNDP, PCA, UMAP, and RP reduced the data dimension to 128 from 6400

for both the training and testing datasets. Tables 5–6 show classification results (accuracy, sensitivity, specificity, and G-mean score) with and without dimension-reduced SNP test data. The classification for the race is five classes whereas for MI age group and sex, they are both two-class identification. For example, for the race classification, based on the trained data using NNDP approach, any of the listed machine learning models were used to classify which of the five classes the test data represent. The same question was asked of PCA and RP methods on the classification of race, MI age groups, and sex.

8.3.1. NNDP-IN and -OUT vs. PCA, UMAP, and RP

As shown in Table 5, NNDP-IN provided the best accuracy and G-mean score for race, MI event age group, and sex classifications. NNDP-IN obtained the best accuracy value of 0.95 and 0.97 for G-mean for the race classification, accuracy of 0.78 and G-mean of 0.68 for the MI event age group classification, and accuracy of 0.99 and G-mean score of 0.99 for the sex classification. NNDP-IN also shows a p-value more than 0.05 for all classifications when compared with the original data (without data dimension reduction) suggesting that the dimension-reduced data is essentially the same as the original data despite the fact that SNP data dimension was reduced 50 times. As expected, NNDP-OUT provides inferior performance compared to NNDP-IN, but it outperforms PCA, UMAP, and RP for some of the classifications. NNDP-OUT outperforms PCA and UMAP and has the same performance as RP with accuracy of 0.95, and a G-mean score of 0.96 for the race classification. It outperformed PCA and RP with a 0.76 accuracy and a 0.59 G-mean score for the MI event age group classification, and outperformed all methods other than NNDP-IN with 0.93 accuracy and 0.92 G-mean score for the sex classification.

UMAP provides similar performance (0.77 accuracy and 0.68 G-mean) compared to NNDP-IN for MI event age group classification, however UMAP is inferior to other classifications since UMAP is designed to perform well for local structure estimation (Diaz-Papkovich et al., 2019). PCA provides good performance for the race and sex classifications since race and sex categories represent global structures and PCA is known to work well for delineating global dynamics (Sakaue et al., 2020). Hence, it is expected that PCA provides not as good performance for MI event age group classification (0.66 accuracy and 0.53 G-mean score) since this is considered as categorizing local structures. RP provides better performance than UMAP for the sex classification. However, RP did not fare well for the classification of MI event age group.

8.3.2. NNDP-IN and -OUT vs. NNDP-Word2Vec, -GloVe, -FastText, -ELMo, -GPT-2, and -ALBERT

In Table 6, we show classification results using NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT (NNDP-ALBERT[d] and NNDP-ALBERT[i]) for the race, MI event age, and sex classifications. NNDP-IN showed the most consistent classification performance for all classification categories. NNDP-Word2Vec and NNDP-GloVe showed the same performance (accuracy of 0.95 and G-mean of 0.97) as that of NNDP-IN for the race classification. However, NNDP-Word2Vec and NNDP-GloVe were found to have inferior performance when compared to NNDP-IN for the MI event age and sex classifications. NNDP-ELMo provided the worst performance for MI event and sex categories when compared to other NNDP approaches. We expected this because of ELMo structure's locality bias, as discussed in Section 8.2.

Interestingly, NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i] provided the same or better accuracy and G-mean scores (0.83 and 0.70 for GPT-2, 0.78 and 0.68 for ALBERT[d], 0.79 and 0.69 for ALBERT[i]) when compared to NNDP-IN (0.78 and 0.68) for the MI event age group classification. As described in Section 4.1, processed literature data for NNDP-variants assign gene names at the first location and all associated sentences are located next to each gene name. Transformer models look at all words for each sequence equally via their self-attention mechanisms, so correlations between associated words for each gene name could be captured by transformer structures without the locality bias (i.e., window size of Word2Vec, and sequential representations of ELMo). The transformer's self-attention mechanisms enable accurate capturing of the correlations, gene names and associated words. However, NNDP-IN provided superior performance for the race and sex classifications (0.97 and 0.99 of G-mean scores, respectively) when compared to NNDP-GPT-2, NNDP-ALBERT[d], NNDP-ALBERT[i] (0.96 and 0.98 of G- mean scores for all NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i], respectively).

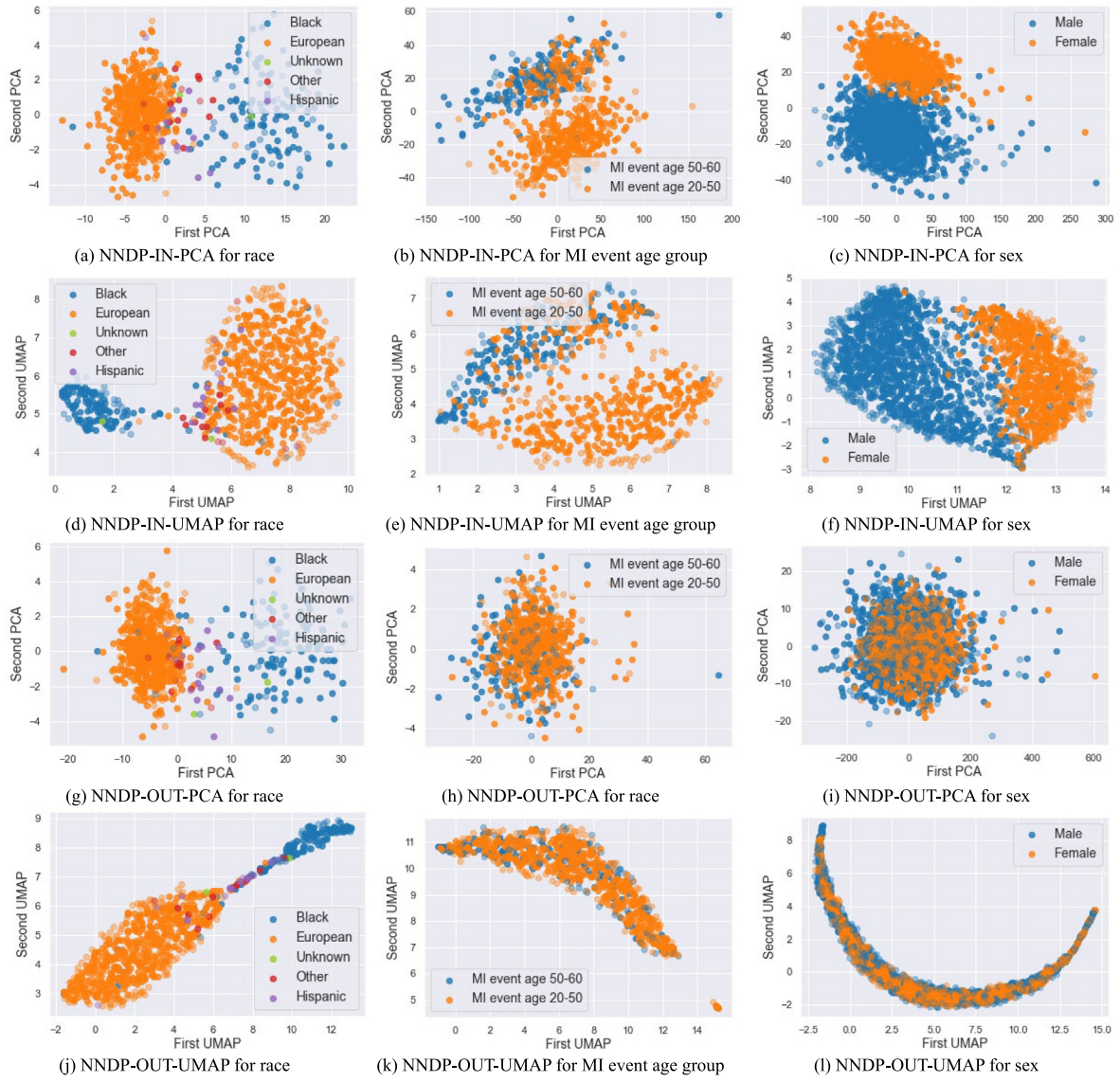
8.4. Missing data simulation results

We validated the generalizability of our NNDP-based methods using missing data. As shown in Figs. 12–16, NNPD-IN provided the best separation capability with 10% of data missing, for all categories: race, MI event age group, and sex. Fig. 13 shows PCA separated the race classes as well as NNPD-IN. However, for the sex category, PCA failed to separate male and female classes as these two classes were clustered together. UMAP with RP performed poorly as it was not able to separate all three categories when missing data were present.

Fig. 14 Shows PCA and UMAP on NNDP-Word2Vec and NNDP-FastText which both performed well in cluster separation for the race, MI event, and sex categories, but the distances between clusters were closer than they were for NNDP-IN-PCA and NNDP-IN-UMAP. PCA on NNDP-GloVe and performed well in cluster separation between European and Black for the race category. PCA and UMAP on NNDP-GloVe resulted in some degree of class separation for the MI event and sex classes, but their cluster separations were not as clear as for the other conventional NNDP approaches. Figs. 15 and 16 show PCA and UMAP with contextual embedding method-based NNDPs—NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT[d], and NNDP-ALBERT[i]. The contextual embedding methods completely failed to separate the MI event age and sex classes. However, NNDP-GPT-2-UMAP and NNDP-ALBERT[i]-UMAP showed new clusters for the sex category when compared to other embedding models' results. The male and female classes were not well separated, although there is some semblance of clusters.

For quantitative comparison of the various methods for the missing data simulation rather than visualizations of classification separation, we show in Tables 7–8 performance metrics similar to Tables 5–6, which used complete data. In Table 7, similar to the visualization results, NNDP-IN showed good performance for all three category classifications. As shown in Tables 7–8, all NNDP's performance metrics are nearly identical to those with the no-dimension-reduced data, shown in the first row, except for the MI event age group classification. This is expected, as 10% of the data were missing, but still good performance overall and especially when compared to the other methods. PCA had the poorest performance for the missing data for all three metrics, as shown in Table 7. Given that as much as 10% missing data did not negatively affect the performance of NNDP, this result further demonstrates the feasibility and generalizability of the NNDP approach.

Interestingly, NNDP-GloVe with 10% missing data showed the best MI event age classification even though NNDP-GloVe visualizations provided poor separation between MI event age groups. We expect this is because GloVe's embedding space is geometrically the most stable when compared to other unique word-based



(a) NNDP-IN-PCA for race (b) NNDP-IN-PCA for MI event age group (c) NNDP-IN-PCA for sex

(d) NNDP-IN-UMAP for race (e) NNDP-IN-UMAP for MI event age group (f) NNDP-IN-UMAP for sex

(g) NNDP-OUT-PCA for race (h) NNDP-OUT-PCA for race (i) NNDP-OUT-PCA for sex

(j) NNDP-OUT-UMAP for race (k) NNDP-OUT-UMAP for MI event age group (l) NNDP-OUT-UMAP for sex

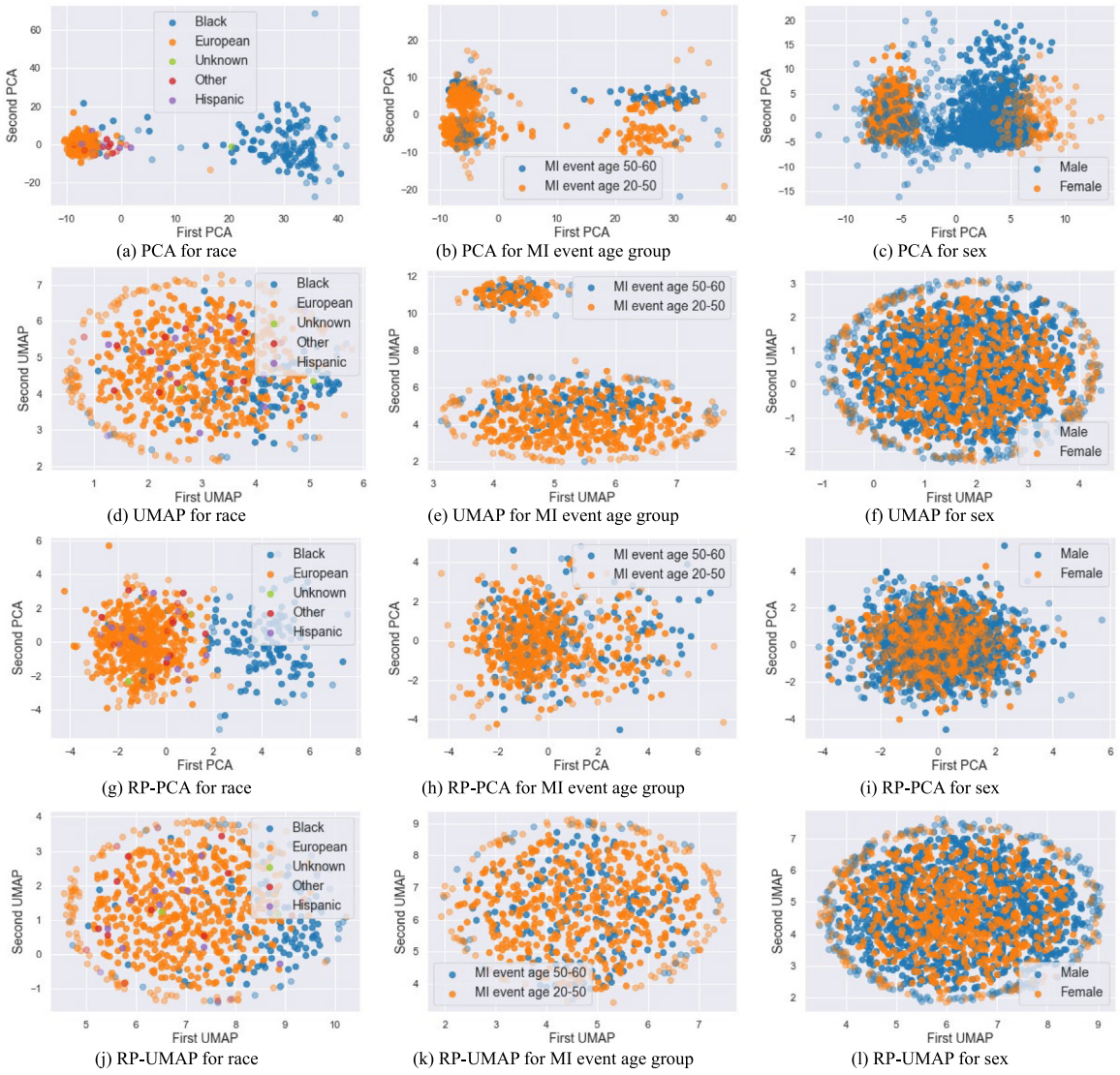
Fig. 12. Visualization result using NNDP methods (NNDP-IN-PCA, NNDP-IN-UMAP, NNDP-OUT-PCA, NNDP-OUT-UMAP) on missing SNP data for race (a, d, g, j), MI event age group (b, e, h, k), and sex (c, f, i, l) categories.

embedding models (Mimno & Thompson, 2017; Wendlandt et al., 2018), as GloVe considers global word-to-word co-occurrence probabilities. However, considering that PCA and UMAP aim to preserve the most significant information from the data (Cheng et al., 2022; Dong et al., 2022; Spencer et al., 2020), GloVe's global word-to-word co-occurrence probabilities may reduce the significant correlations between gene names, which consequently leads to inferior visualization results when compared to NNDP-IN. We show in Table 4(b) the characteristics of the GloVe's global word-to-word co-occurrences via semantic analysis.

We found that transformer-based contextual embedding method-based NNDPs—NNDP-GPT-2, NNDP-ALBERT[d], and NNDL-ALBERT[i] with 10% missing data resulted in good performance for the MI event age and sex classification tasks even though their visualizations failed to separate different clusters, as shown in Fig. 16. Because the transformer's self-attention mechanisms look for all words for each sequence equally, we expected that transformers would enable stable embedding space to represent searched words. However, we found that NNDP-GPT-2, NNDP-ALBERT[d], and NNDP-ALBERT[i] included sub-word-biased representations, as shown in Table 4(e)–(g). We expected the sub-word-biased representation to reduce the correlation between genes and robustness of visualization for those cases with 10% missing data.

9. Discussion and conclusion

In this paper, we proposed an unsupervised literature-based SNP data visualization/dimensionality reduction method – the NNDP method – that can capture both global and local structures. For the method validation, we designed a literature embedding model. We compared the literature-embedding-based NNDP against traditional methods: PCA, UMAP, and RP. We also compared the literature embedding model-based NNDP with NNDP variants using six other popular embedding models: Word2Vec, GloVe, FastText, ELMo, GPT-2, and ALBERT. As far as we are aware,



(a) PCA for race (b) PCA for MI event age group (c) PCA for sex

(d) UMAP for race (e) UMAP for MI event age group (f) UMAP for sex

(g) RP-PCA for race (h) RP-PCA for MI event age group (i) RP-PCA for sex

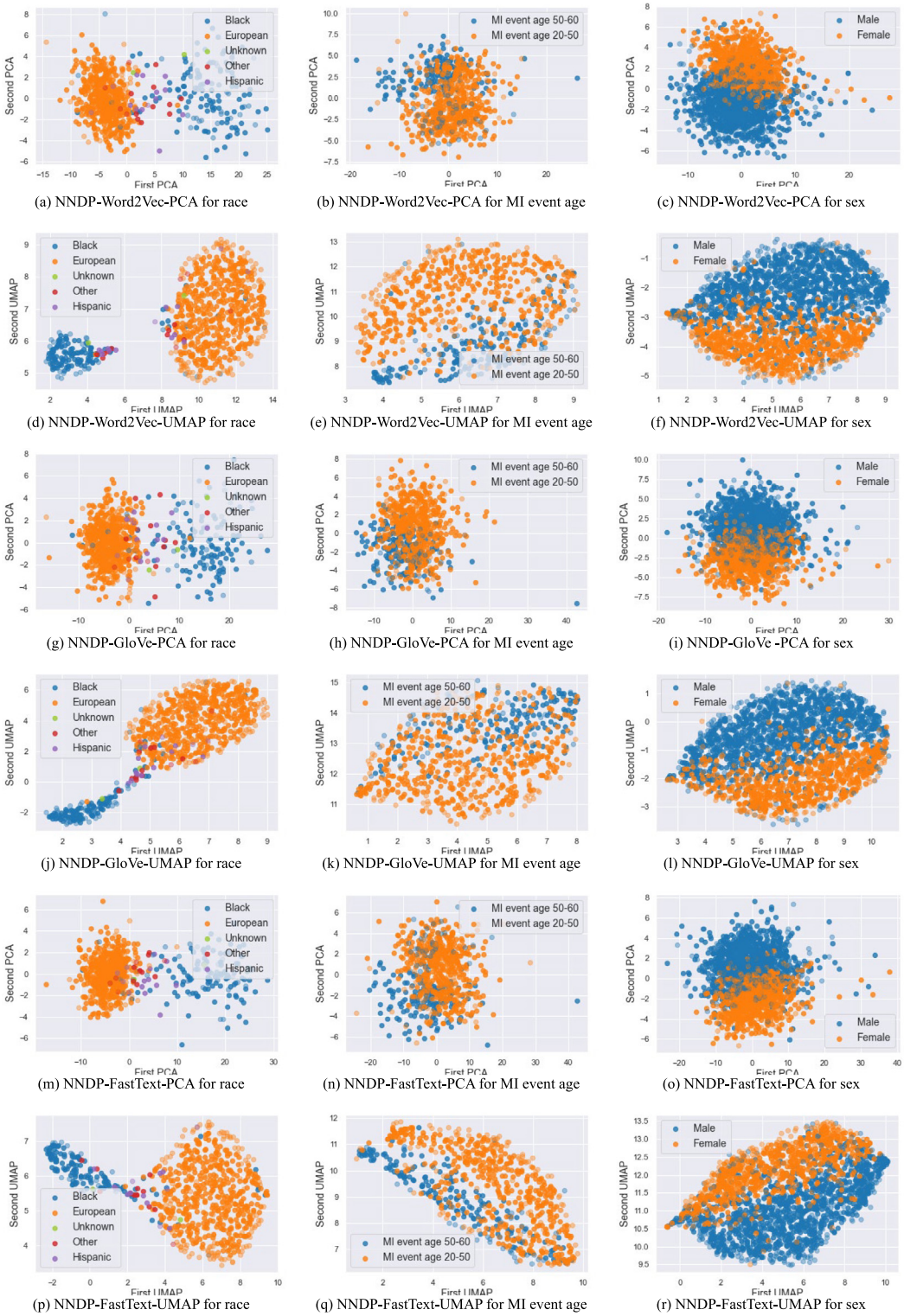
(j) RP-UMAP for race (k) RP-UMAP for MI event age group (l) RP-UMAP for sex

Fig. 13. Visualization results using other methods (basic PCA/UMAP, RP-PCA, RP-UMAP) on missing SNP data for race, MI event age group, and sex categories.

this is one of the first studies to conduct dimensionality reduction and data visualization on different modes of data, especially genetic data, using literature databases along with textual interpretability. Our NNDP-based methods outperformed all other popular dimensionality reduction and visualization models such as PCA and UMAP examined in this work, as determined by both visual and quantitative metrics to separate various categorical classes. Our method was shown to provide the most efficient separation of the genotype data structures in SNP data even when it is confronted with significant missing data. Our NNDP method also provides textual interpretability to understand how each embedding model performs the visualization tasks, via semantic analysis.

NNDP-based methods, especially NNDP-IN, captured both global and local structures whereas PCA and UMAP do not. Even when they are combined to capture both global and local structures, PCA and UMAP visualization results were not as informative or accurate as those of NNDP. Another advantage of our method was that by incorporating hidden risks obtained from literature for MI, we found an interesting observation. NNDP-IN-UMAP, NNDP-Word2Vec-UMAP, NNDP-GPT-2-UMAP, and NNDP-ALBERT-UMAP[i] showed that the younger cohort of Black people had similar high risks for MI as the older Black people, but this was not observed for the younger European subjects. However, our visualization is based on literature word embeddings, and some literature (Dyke, 2018; Garcia et al., 2021; Lee et al., 2016) suggested that young blacks have more MI preva lence than other races, and old age is a risk factor for MI (Hajar, 2017). Considering socioeconomic factors that were reported as the most significant factor to the high prevalence of MI in young Black (Garcia et al., 2021), our visualization results may have correctly captured socioeconomic-related word representations. Understanding hidden risks using statistical and literature analysis is significant for avoiding bias in heritage-based disease risk estimations (Baud et al., 2017). Such informative results were not possible with other visualization techniques.

Our NNDP-based methods can potentially be used in many applications that require the unraveling of hidden structures of population data for accurate risk prediction and genetic data QC



(a) NNDP-Word2Vec-PCA for race (b) NNDP-Word2Vec-PCA for MI event age (c) NNDP-Word2Vec-PCA for sex

(d) NNDP-Word2Vec-UMAP for race (e) NNDP-Word2Vec-UMAP for MI event age (f) NNDP-Word2Vec-UMAP for sex

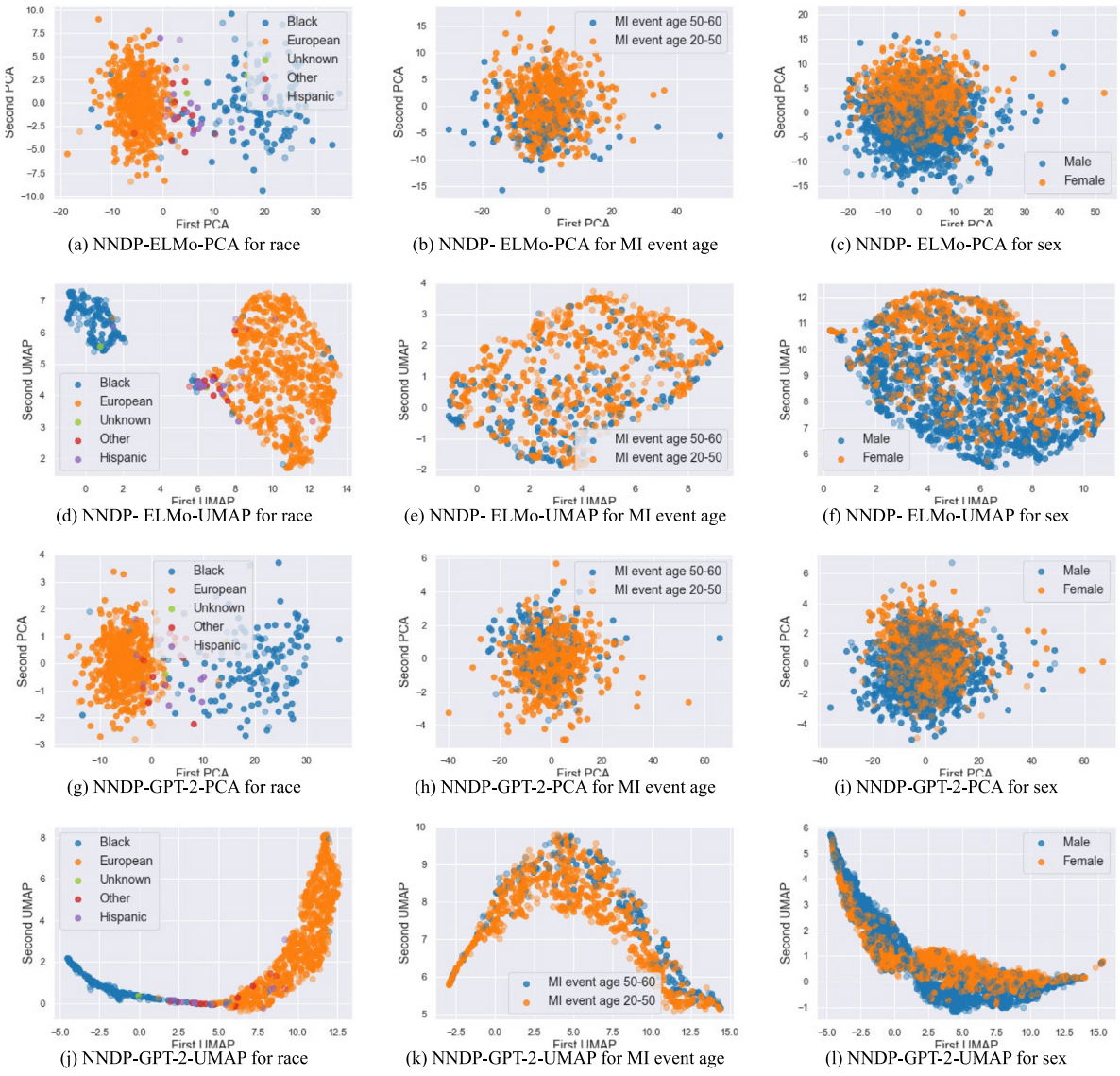
(g) NNDP-GloVe-PCA for race (h) NNDP-GloVe-PCA for MI event age (i) NNDP-GloVe -PCA for sex

(j) NNDP-GloVe-UMAP for race (k) NNDP-GloVe-UMAP for MI event age (l) NNDP-GloVe-UMAP for sex

(m) NNDP-FastText-PCA for race (n) NNDP-FastText-PCA for MI event age (o) NNDP-FastText-PCA for sex

(p) NNDP-FastText-UMAP for race (q) NNDP-FastText-UMAP for MI event age (r) NNDP-FastText-UMAP for sex

Fig. 14. Visualization results using NNDP-Word2Vec-PCA, NNDP-Word2Vec-UMAP, NNDP-GloVe-PCA, NNDP-GloVe-UMAP, NNDP-FastText-PCA, and NNDP-FastText-UMAP on missing SNP for race (a, d, g, j, m, p), MI event age group (b, e, h, k, n, q), and sex (c, f, i, l, o, r) categories.



(a) NNDP-ELMo-PCA for race (b) NNDP- ELMo-PCA for MI event age (c) NNDP- ELMo-PCA for sex

(d) NNDP- ELMo-UMAP for race (e) NNDP- ELMo-UMAP for MI event age (f) NNDP- ELMo-UMAP for sex

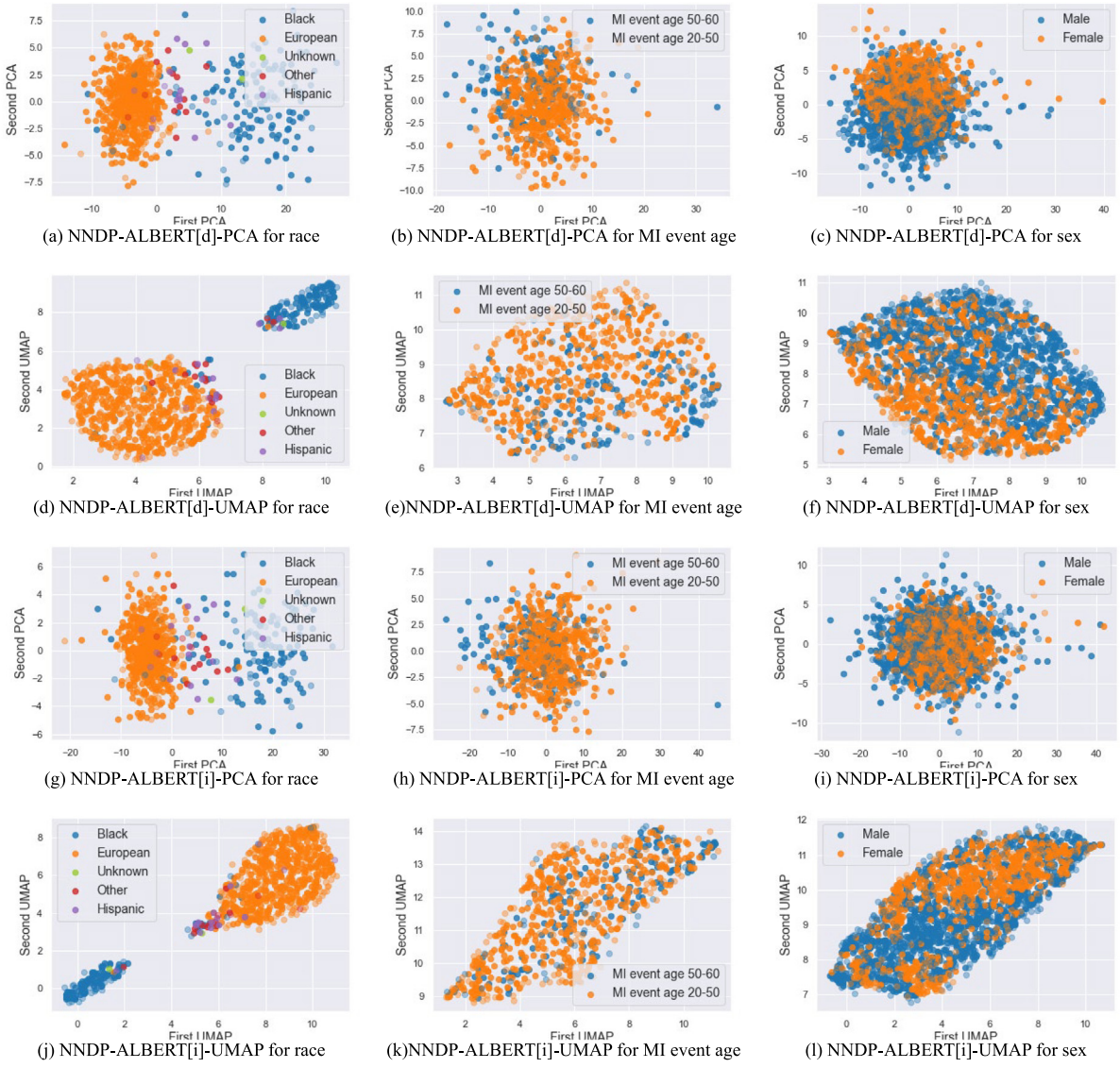
(g) NNDP-GPT-2-PCA for race (h) NNDP-GPT-2-PCA for MI event age (i) NNDP-GPT-2-PCA for sex

(j) NNDP-GPT-2-UMAP for race (k) NNDP-GPT-2-UMAP for MI event age (l) NNDP-GPT-2-UMAP for sex

Fig. 15. Visualization results using NNDP-ELMo-PCA, NNDP- ELMo-UMAP, NNDP-GPT-2-PCA, and NNDP-GPT-2-UMAP on missing SNP for race (a, d, g, j), MI event age group (b, e, h, k), and sex (c, f, i, l) categories.

procedures. QC is typically used to remove false positives arising from poor quality DNA, hidden confounders, and genotyping artifacts (Morris et al., 2010). We have simulated this type of scenario by deliberately removing 10% of SNP data. Both PCA and UMAP were unable to handle missing data whereas NNDP was still able to provide good separations of different categories for both visualization and classification tasks. Furthermore, PCA and UMAP are not robust for large SNP data, as shown in Appendix. Our NNDP method provides not only better visualization performance but also textual interpretability of visualization results, which facilitates explainable artificial intelligence. As genomic sequencing technologies advance, there will be a rapid increase in the volume and complexity of genetic data, hence, more advanced visualization approaches that can handle high-dimensional genetic data with better interpretability will become more critical (O'Donoghue et al., 2018).

Our proposed NNDP method has shown robust visualization with textual interpretability. This capability will enable better classification and categorization, as complex biological datasets can be better visualized and interpreted. Our approach can be combined with ChatGPT (Brown et al., 2020) since all modern ChatBot models are trained using large embedding models such as GPT-3 (Brown et al., 2020; Suhaili et al., 2021). A ChatBot system, when combined with a better visualization approach, has the potential to not only obtain relevant information quickly and efficiently, but also to unravel those complex structures into visually interpretable results. Moreover, since textual representations are also used to generate text scripts for a given image or image scenes for a text set (Yu et al., 2022), our NNDP approach can be used with medical image and bio-signals such as electrocardiogram for more interpretable disease risk analysis and decision-making on various datasets with ChatBot systems. Additionally, since automatic speech recognition systems use textual representation to display machine-recognized utterances for given speech sets (Bang et al., 2022; Moon et al., 2019; Nassif



(a) NNDP-ALBERT[d]-PCA for race (b) NNDP-ALBERT[d]-PCA for MI event age (c) NNDP-ALBERT[d]-PCA for sex

(d) NNDP-ALBERT[d]-UMAP for race (e)NNDP-ALBERT[d]-UMAP for MI event age (f) NNDP-ALBERT[d]-UMAP for sex

(g) NNDP-ALBERT[i]-PCA for race (h) NNDP-ALBERT[i]-PCA for MI event age (i) NNDP-ALBERT[i]-PCA for sex

(j) NNDP-ALBERT[i]-UMAP for race (k)NNDP-ALBERT[i]-UMAP for MI event age (l) NNDP-ALBERT[i]-UMAP for sex

Fig. 16. Visualization results using NNDP-ALBERT[d]-PCA, NNDP-ALBERT[d]-UMAP, NNDP-ALBERT[i]-PCA, and NNDP-ALBERT[i]-UMAP on missing SNP for race (a, d, g, j, m, p), MI event age group (b, e, h, k, n, q), and sex (c, f, i, l, o, r) categories.

et al., 2019), our NNDP approach can be combined with conversational artificial intelligence assistant systems to aid human tasks more efficiently.

Our semantic analysis showed that a well-trained embedding model successfully captures the dynamics of SNP data, and that the visualization results can be interpreted via textual analysis. Our method's limitation is that the model requires the SNPs to be identified and referenced in the literature. Since some SNPs do not have any references in the literature, the SNPs are not identifiable. However, the limitation can be addressed with advanced algorithms such as SNP interaction algorithms (Elgart et al., 2022; Silva et al., 2022). Moreover, in our prior work (Moon et al., 2023, 2021), we trained a skip-gram-based literature-based embedding model to reduce the dimension of phenotype data four times without any deterioration of the classification results. It was shown that our NNDP approach can be efficiently used for not only genetic data but also other various data without any prior knowledge of the data.

In summary, there are several novelties with the proposed method. First, we developed an efficient method to comb through literature data such as the published abstracts to extract relevant information such as genes associated with specific diseases, which in our case was MI. Our approach also allowed interpretability of the visualized results using semantic correlations. While a simple back-propagation neural network was used for NNDP-IN, we were able to efficiently reduce the dimension of high-dimensional data without any degradation in the performance of the proposed method. NNDP-IN showed even more impressive visualization results when compared to the state-of-the-art transformer neural network-based NNDP (GPT-2 and ALBERT). Considering NNDP-IN's embedding vectors captured topically related words for each query in the semantic analysis, the visualization performance of NNDP-IN may improve by making its window size bigger since conventional word embedding models' bigger window size can capture more topical information between words (Levy & Goldberg, 2014).

Table 7
Quantitative performance of Race, MI event age group, and sex classifications With 10% of Data Missing Simulation for NNDP-IN, NNDP-OUT, PCA, UMAP, and RP.

		(a) Race				(b) MI Event Age Group				(c) Sex			
Method	Model	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.
Original 64000-dimensional data	SVM-linear	**0.96**	**0.96**	**0.99**	**0.97**	**0.75**	**0.76**	**0.73**	**0.74**	0.97	1.00	0.95	0.97
	SVM-rbf	0.94	0.94	0.98	0.96	0.73	1.00	0.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**
	SVM-poly	0.94	0.94	0.98	0.96	0.73	1.00	0.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**
	LR	0.94	0.94	0.98	0.96	**0.77**	0.96	0.26	0.49	0.99	1.00	0.99	0.99
	MLP	**0.96**	**0.96**	**0.99**	**0.97**	0.75	0.91	0.33	0.53	0.91	0.90	0.92	0.91
NNDP (IN)	SVM-linear	**0.94**	**0.94**	**0.98**	**0.96**	0.74	0.83	0.51	0.64	0.94	0.93	0.94	0.94
	SVM-rbf	**0.94**	**0.94**	**0.98**	**0.96**	0.75	0.88	0.41	0.58	**0.96**	**0.95**	**0.96**	**0.96**
	SVM-poly	**0.94**	**0.94**	**0.98**	**0.96**	0.75	0.88	0.41	0.58	**0.96**	**0.95**	**0.96**	**0.96**
	LR	**0.94**	**0.94**	**0.98**	**0.96**	0.75	0.85	0.50	0.65	0.93	0.90	0.95	0.93
	MLP	**0.94**	**0.94**	**0.98**	**0.96**	**0.77**	**0.86**	**0.51**	**0.66**	0.95	0.94	0.95	0.94
	*p-value	**0.10**				**1.05**				0.00			
NNDP (OUT)	SVM-linear	**0.94**	**0.94**	**0.98**	**0.96**	**0.72**	**0.88**	**0.30**	**0.51**	**0.74**	**0.57**	**0.83**	**0.69**
	SVM-rbf	**0.94**	**0.94**	**0.98**	**0.96**	0.73	0.97	0.06	0.23	0.73	0.34	0.93	0.56
	SVM-poly	**0.94**	**0.94**	**0.98**	**0.96**	0.73	0.97	0.06	0.23	0.73	0.34	0.93	0.56
	LR	**0.94**	**0.94**	**0.98**	**0.96**	0.71	0.91	0.18	0.39	0.74	0.56	0.83	0.68
	MLP	**0.94**	**0.94**	**0.98**	**0.96**	0.72	0.90	0.25	0.47	**0.72**	**0.60**	**0.78**	**0.69**
	*p-value	**0.10**				0.09				0.00			
PCA	SVM-linear	**0.91**	**0.91**	**0.97**	**0.94**	0.58	0.67	0.35	0.47	**0.42**	**0.24**	**0.51**	**0.29**
	SVM-rbf	0.16	0.16	0.72	0.33	0.28	0.01	0.99	0.06	0.65	0.00	1.00	0.00
	SVM-poly	0.16	0.16	0.72	0.33	0.28	0.01	0.99	0.06	0.65	0.00	1.00	0.00
	LR	0.85	0.85	0.95	0.90	**0.61**	**0.71**	**0.34**	**0.48**	0.42	0.22	0.53	0.27
	MLP	0.66	0.66	0.90	0.77	0.65	0.80	0.26	0.43	0.47	0.19	0.62	0.26
	*p-value	0.00				0.00				0.00			
UMAP	SVM-linear	**0.83**	**0.83**	**0.95**	**0.88**	**0.72**	**0.83**	**0.41**	**0.58**	0.63	0.09	0.91	0.27
	SVM-rbf	0.80	0.80	0.93	0.87	0.73	0.97	0.06	0.15	0.65	0.00	1.00	0.00
	SVM-poly	0.80	0.80	0.93	0.87	0.73	0.97	0.06	0.15	0.65	0.00	1.00	0.00
	LR	0.81	0.81	0.94	0.87	0.71	0.84	0.37	0.55	0.63	0.06	0.93	0.21
	MLP	0.82	0.82	0.94	0.88	0.72	0.85	0.36	0.54	**0.60**	**0.19**	**0.81**	**0.39**
	*p-value	0.00				0.01				0.00			
RP (Gaussian Distribution)	SVM-linear	0.92	0.92	0.97	0.94	**0.66**	**0.80**	**0.28**	**0.47**	**0.69**	**0.47**	**0.81**	**0.62**
	SVM-rbf	**0.94**	**0.94**	**0.98**	**0.96**	0.73	1.00	0.01	0.03	0.71	0.40	0.88	0.59
	SVM-poly	**0.94**	**0.94**	**0.98**	**0.96**	0.73	1.00	0.01	0.03	0.71	0.40	0.88	0.59
	LR	0.93	0.93	0.98	0.95	0.71	0.93	0.11	0.32	0.70	0.42	0.84	0.60
	MLP	0.91	0.91	0.97	0.94	0.66	0.83	0.22	0.41	0.67	0.47	0.78	0.61
	*p-value	**0.10**				0.01				0.00			

*p-value was computed based on the best accuracies between original 6400-dimensional data set and each method.

Since transformer structures such as GPT-2 and ALBERT are appropriate for a large dataset, developing transformer-based literature embedding models would be significant for more advanced genetic data visualization. It should be noted, however, that the original GPT-2 and ALBERT structures provided inferior performance for the sex category when compared to NNDP-IN. The reasons why transformer-based models did not provide good distinction between male and female classes can be explained by semantic analysis. Table 4 shows that GPT-2 and ALBERT provided sub-word-biased related words for the query 'male + female' instead of topically and categorically related words. Hence, our textual interpretability using semantic analysis could provide significant insights into designing appropriate transformer-structure-based literature embedding models for better genetic data visualization capabilities.

In addition, the NNDP-based methods remained robust even when we simulated having as much as 10% of the data missing, while PCA and UMAP fared poorly in this simulation. Given that literature text data and gene data all lead to high-dimensional data, interpretable and explainable visualization of data to unravel hidden structures and dynamics are of high importance and is an unmet need. The proposed NNDP is a solution for these purposes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Table 8
Quantitative performance of Race, MI event age group, and sex classifications With 10% of Data Missing Simulation for NNDP-Word2Vec, NNDP-GloVe, NNDP-FastText, NNDP-ELMo, NNDP-GPT-2, and NNDP-ALBERT.

		(a) Race				(b) MI Event Age Group				(c) Sex			
Method	Model	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.	Acc.	Sens.	Spec.	G.
NNDP-Word2Vec	SVM-linear	**0.95**	**0.95**	**0.98**	**0.96**	0.76	0.86	0.52	**0.66**	**0.94**	0.94	0.94	**0.94**
	SVM-rbf	0.94	0.94	0.98	0.96	0.77	0.91	0.40	0.60	**0.94**	**0.95**	0.94	**0.94**
	SVM-poly	0.94	0.94	0.98	0.96	0.77	0.91	0.40	0.60	**0.94**	**0.95**	0.94	**0.94**
	LR	**0.95**	**0.95**	**0.98**	**0.96**	**0.78**	0.92	0.42	0.62	**0.94**	0.93	**0.95**	**0.94**
	MLP	0.94	0.94	0.98	0.96	0.77	0.89	0.45	0.63	0.93	0.92	0.93	0.93
	*p-value	**0.43**				**0.69**				0.00			
NNDP-GLOVE	SVM-linear	**0.95**	**0.95**	**0.98**	**0.96**	0.76	0.84	0.53	**0.67**	**0.96**	**0.96**	**0.96**	**0.96**
	SVM-rbf	0.94	0.94	0.98	0.96	0.76	0.90	0.38	0.58	0.95	0.94	0.95	0.95
	SVM-poly	0.94	0.94	0.98	0.96	0.76	0.90	0.38	0.58	0.95	0.94	0.95	0.95
	LR	0.94	0.94	0.98	0.96	0.76	0.89	0.42	0.61	0.95	0.94	0.96	0.95
	MLP	0.94	0.94	0.98	0.96	0.76	0.88	0.46	0.63	0.95	0.95	0.95	0.95
	*p-value	**0.43**				**0.70**				0.00			
NNDP-FastText	SVM-linear	**0.95**	**0.95**	**0.98**	**0.96**	**0.76**	**0.85**	**0.52**	**0.67**	0.94	**0.94**	**0.95**	**0.94**
	SVM-rbf	0.94	0.94	0.98	0.96	0.75	0.92	0.31	0.53	**0.95**	0.93	0.95	**0.94**
	SVM-poly	0.94	0.94	0.98	0.96	0.75	0.92	0.31	0.53	**0.95**	0.93	0.95	**0.94**
	LR	0.94	0.94	0.98	0.96	0.77	0.90	0.45	0.63	0.94	0.92	0.95	**0.94**
	MLP	0.93	0.93	0.98	0.96	0.76	0.88	0.46	0.63	0.93	0.91	0.95	0.93
	*p-value	**0.43**				**0.70**				0.00			
NNDP-ELMo	SVM-linear	0.94	0.94	0.98	0.96	**0.76**	**0.88**	**0.47**	**0.63**	**0.93**	**0.92**	**0.94**	**0.93**
	SVM-rbf	0.94	0.94	0.98	0.96	0.72	0.92	0.17	0.38	0.78	0.59	0.88	0.72
	SVM-poly	0.94	0.94	0.98	0.96	0.72	0.92	0.17	0.38	0.78	0.59	0.88	0.72
	LR	0.94	0.94	0.98	0.96	0.74	0.90	0.33	0.52	0.90	0.85	0.92	0.89
	MLP	0.92	0.92	0.98	0.95	0.70	0.86	0.30	0.50	0.83	0.75	0.88	0.81
	*p-value	**0.10**				**0.70**				0.00			
NNDP-GPT2	SVM-linear	0.94	0.94	0.98	0.96	**0.76**	**0.86**	**0.49**	**0.65**	**0.95**	**0.93**	**0.95**	**0.94**
	SVM-rbf	0.94	0.94	0.98	0.96	0.73	0.93	0.22	0.45	0.93	0.90	**0.95**	0.92
	SVM-poly	0.94	0.94	0.98	0.96	0.73	0.93	0.22	0.45	0.93	0.90	**0.95**	0.92
	LR	0.94	0.94	0.98	0.96	**0.78**	**0.89**	**0.49**	**0.65**	0.94	0.92	**0.95**	0.94
	MLP	0.92	0.92	0.98	0.95	0.75	0.88	0.42	0.60	0.93	0.92	0.94	0.93
	*p-value	**0.10**				**0.69**				0.00			
NNDP-ALBERT[d]	SVM-linear	0.94	0.94	0.98	0.96	0.76	0.85	0.51	**0.66**	**0.95**	**0.94**	**0.96**	**0.95**
	SVM-rbf	0.94	0.94	0.98	0.96	**0.77**	**0.92**	0.39	0.58	0.94	0.90	0.96	0.93
	SVM-poly	0.94	0.94	0.98	0.96	**0.77**	**0.92**	0.39	0.58	0.94	0.90	0.96	0.93
	LR	0.94	0.94	0.98	0.96	0.76	0.90	0.40	0.60	0.95	0.92	0.96	0.94
	MLP	0.94	0.94	0.98	0.96	0.76	0.87	0.47	0.63	0.93	0.89	0.95	0.92
	*p-value	**0.10**				0.70				**0.00**			
NNDP-ALBERT[i]	SVM-linear	**0.95**	**0.95**	**0.98**	**0.96**	0.76	0.85	0.51	**0.66**	**0.95**	**0.95**	0.95	**0.95**
	SVM-rbf	0.94	0.94	0.98	0.96	**0.77**	**0.92**	0.39	0.58	0.95	0.93	**0.96**	0.94
	SVM-poly	0.94	0.94	0.98	0.96	**0.77**	**0.92**	0.39	0.58	0.95	0.93	**0.96**	0.94
	LR	0.94	0.94	0.98	0.96	0.76	0.90	0.40	0.60	0.95	0.93	**0.96**	0.94
	MLP	0.94	0.94	0.98	0.96	0.76	0.87	0.47	0.63	0.95	0.92	**0.96**	0.94
	*p-value	**0.43**				**1.05**				0.00			

*p-value was computed based on the best accuracies between original 6400-dimensional data set and each method.

Acknowledgments

Datasets in this study were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000883.v1.p1 and dbGaP accession phs000279.v2.p1 by participating in "Big Data Analysis Challenge: Creating New Paradigms for Heart Failure Research" created by NHLBI in 2020".

Appendix

As a further investigation, we classified SNP data based on race, MI event age group, and sex in order to examine the classification performance and computation time as the dimension size varied (128*Q-dimension where Q ranged from 50 to 250 and each step size $Q = 50$), as shown in Tables A.1 and A.2.

Table A.1
Classification performance (averaged G-mean score ± its standard deviation [MIN and MAX G-Scores for all G-mean scores of each model] for all ML models) on each 128*Q-dimensional SNP data set (Q ranges from 50 to 250).

	Model	N (original data's dimension = 128*Q)				
		6400 (Q=50)	12,800 (Q=100)	19,200 (Q=150)	25,600 (Q=200)	32,000 (Q=250)
Race (*)	**NNDP (IN)**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**
	NNDP (OUT)	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]
	NNDP (Word2Vec)	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.96]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.96 ± 0.00 [0.96, 0.97]**
	NNDP (GloVe)	**0.96 ± 0.00 [0.96, 0.97]**	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]
	NNDP (FastText)	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	**0.96 ± 0.00 [0.96, 0.97]**
	NNDP (ELMo)	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.95, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]
	NNDP (GPT-2)	**0.96 ± 0.00 [0.96, 0.97]**	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.95, 0.96]	0.96 ± 0.00 [0.96, 0.96]	**0.96 ± 0.00 [0.96, 0.97]**
	NNDP (ALBERT[d])	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]
	NNDP (ALBERT[i])	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]
	PCA	0.63 ± 0.26 [0.40, 0.95]	0.66 ± 0.23 [0.46, 0.95]	0.65 ± 0.23 [0.46, 0.96]	0.64 ± 0.21 [0.45, 0.95]	0.65 ± 0.20 [0.45, 0.92]
	UMAP	0.90 ± 0.01 [0.89, 0.91]	0.88 ± 0.01 [0.87, 0.89]	0.87 ± 0.00 [0.86, 0.87]	0.85 ± 0.02 [0.82, 0.87]	0.87 ± 0.00 [0.87, 0.87]
	RP	0.96 ± 0.00 [0.96, 0.96]	0.96 ± 0.00 [0.96, 0.96]	0.95 ± 0.00 [0.95, 0.96]	0.95 ± 0.00 [0.95, 0.96]	0.95 ± 0.00 [0.95, 0.96]
MI event age group	**NNDP (IN)**	**0.61 ±0.04 [0.57, 0.68]**	**0.57 ± 0.06 [0.52, 0.66]**	**0.52 ± 0.10 [0.40, 0.64]**	**0.48 ± 0.11 [0.35, 0.60]**	**0.41 ± 0.14 [0.24, 0.55]**
	NNDP (OUT)	0.51 ± 0.08 [0.42, 0.59]	0.45 ± 0.11 [0.32, 0.55]	0.33 ± 0.21 [0.07, 0.55]	0.26 ± 0.16 [0.07, 0.48]	0.18 ± 0.12 [0.06, 0.34]
	NNDP (Word2Vec)	0.60 ± 0.03 [0.57, 0.66]	0.56 ± 0.06 [0.49, 0.66]	0.51 ± 0.10 [0.39, 0.62]	0.44 ± 0.15 [0.26, 0.58]	0.39 ± 0.19 [0.16, **0.57**]
	NNDP (GloVe)	0.60 ± 0.03 [0.58, 0.66]	**0.57 ± 0.05 [0.52, 0.66]**	**0.52 ± 0.10 [0.40, 0.63]**	**0.48 ± 0.15 [0.3, 0.63]**	**0.40 ± 0.23 [0.12, 0.61]**
	NNDP (FastText)	0.61 ± 0.03 [0.60, 0.67]	**0.57** ±0.07 [0.49, **0.69**]	0.49 ±0.15 **[0.31, 0.65]**	0.42 ±0.17 **[0.21, 0.61]**	**0.36** ±0.23 [0.08, **0.57**]
	NNDP (ELMo)	0.52 ± 0.05 [0.48, 0.59]	**0.39** ±0.07 [0.30, 0.47]	**0.25** ±0.13 [0.10, 0.46]	**0.17** ±0.10 [0.06, 0.32]	**0.16** ±0.09 [0.07, 0.30]
	NNDP (GPT-2)	0.59 ± 0.07 [0.52, **0.70**]	0.50 ±0.13 [0.35, 0.65]	0.41 ±0.25 **[0.11, 0.67]**	**0.35** ±0.22 [0.09, **0.62**]	**0.29** ±0.20 [0.06, **0.59**]
	NNDP (ALBERT[d])	**0.62** ±0.03 [0.59, **0.68**]	0.56 ±0.09 [0.45, **0.67**]	0.51 ±0.13 [0.36, **0.65**]	0.42 ±0.16 [0.23, 0.57]	0.40 ±0.21 [0.14, **0.58**]
	NNDP (ALBERT[i])	**0.62 ± 0.05 [0.57, 0.69]**	0.53 ± 0.10 [0.42, 0.66]	0.48 ± 0.17 [0.27, 0.66]	0.42 ± 0.22 [0.15, 0.61]	0.39 ± 0.26 [0.07, 0.62]
	PCA	0.34 ± 0.22 [0.07, 0.53]	0.40 ± 0.09 [0.29, 0.49]	0.34 ± 0.18 [0.12, 0.51]	0.33 ± 0.17 [0.12, 0.51]	0.26 ± 0.21 [0.00, 0.46]
	UMAP	0.52 ± 0.17 [0.31, 0.68]	0.35 ± 0.21 [0.10, 0.55]	0.15 ± 0.12 [0.0, 0.29]	0.16 ± 0.13 [0.0, 0.28]	0.24 ± 0.19 [0.00, 0.41]
	RP	0.42 ± 0.08 [0.35, 0.55]	0.37 ± 0.10 [0.26, 0.49]	0.26 ± 0.17 [0.06, 0.45]	0.25 ± 0.17 [0.05, 0.43]	0.22 ± 0.15 [0.04, 0.39]
Sex	**NNDP (IN)**	**0.99 ± 0.00 [0.98, 0.55]**	**0.96 ± 0.00 [0.96, 0.97]**	**0.94 ± 0.00 [0.93, 0.94]**	**0.90 ± 0.00 [0.89, 0.90]**	**0.87 ± 0.00 [0.87, 0.88]**
	NNDP (OUT)	0.91 ± 0.01 [0.90, 0.92]	0.81 ± 0.03 [0.78, 0.84]	0.72 ± 0.06 [0.65, 0.78]	0.65 ± 0.08 [0.55, 0.73]	0.59 ± 0.10 [0.47, 0.69]
	NNDP (Word2Vec)	0.98 ± 0.00 [0.97, 0.98]	0.94 ± 0.00 [0.94, 0.95]	0.90 ± 0.01 [0.89, 0.91]	0.86 ± 0.00 [0.86, 0.87]	0.83 ± 0.01 [0.82, 0.85]
	NNDP (GloVe)	0.98 ± 0.00 [0.98, 0.98]	0.95 ± 0.01 [0.94, 0.96]	0.92 ± 0.01 [0.91, 0.94]	**0.89 ± 0.02 [0.87, 0.92]**	**0.86 ± 0.02 [0.84, 0.89]**
	NNDP (FastText)	0.98 ± 0.00 [0.97, 0.98]	0.94 ± 0.00 [0.94, 0.95]	0.90 ± 0.01 [0.89, 0.92]	0.86 ± 0.01 [0.85, 0.87]	0.82 ± 0.02 [0.80, 0.84]
	NNDP (ELMo)	0.90 ± 0.05 [0.85, 0.97]	0.81 ± 0.08 [0.72, 0.92]	0.72 ± 0.12 [0.59, 0.88]	0.65 ± 0.15 [0.48, 0.85]	0.62 ± 0.14 [0.45, 0.80]
	NNDP (GPT-2)	0.97 ± 0.01 [0.97, 0.98]	0.94 ± 0.02 [0.92, 0.94]	0.90 ± 0.04 [0.85, **0.94**]	0.86 ± 0.05 [0.79, **0.91**]	0.81 ± 0.08 [0.71, **0.89**]
	NNDP (ALBERT[d])	0.97 ± 0.00 [0.97, 0.98]	0.94 ± 0.01 [0.93, 0.96]	0.90 ± 0.02 [0.88, 0.93]	0.86 ± 0.02 [0.84, 0.89]	0.84 ± 0.03 [0.81, 0.87]
	NNDP (ALBERT[i])	0.97 ± 0.00 [0.97, 0.98]	0.95 ± 0.01 [0.94, 0.96]	0.92 ± 0.01 [0.91, 0.94]	0.89 ± 0.02 [0.87, 0.91]	0.86 ± 0.03 [0.82, 0.90]
	PCA	0.64 ± 0.27 [0.32, 0.88]	0.23 ± 0.09 [0.12, 0.32]	0.45 ± 0.11 [0.31, 0.55]	0.43 ± 0.05 [0.37, 0.48]	0.49 ± 0.01 [0.48, 0.50]
	UMAP	0.30 ± 0.24 [0.00, 0.52]	0.73 ± 0.05 [0.67, 0.78]	0.75 ± 0.02 [0.72, 0.77]	0.65 ± 0.03 [0.61, 0.69]	0.52 ± 0.08 [0.43, 0.60]
	RP	0.70 ± 0.02 [0.68, 0.72]	0.60 ± 0.02 [0.58, 0.63]	0.55 ± 0.03 [0.51, 0.58]	0.46 ± 0.07 [0.37, 0.53]	0.42 ± 0.09 [0.32, 0.51]

Table A.2
Dimensionality reduction operation time for each category's 128*Q-dimensional SNP data set (Q ranges from 50 to 250).

Category	N (dimension = 128*Q)	NNDP (IN)	NNDP (OUT)	NNDP (Word2Vec)	NNDP (Glo-Ve)	NNDP (Fast-Text)	NNDP (ELMo)	NNDP (GPT-2)	NNDP (ALBERT[d])	NNDP (ALBERT[i])	PCA	UMAP	RP
Race (*)	50 (6,400)	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	0.07	0.85	0.02
	100 (12,800)	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	0.12	0.94	0.02
	150 (19,200)	**0.02**	**0.02**	**0.03**	**0.02**	**0.02**	**0.03**	**0.03**	**0.02**	**0.02**	0.18	1.07	0.03
	200 (25,600)	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	0.24	1.22	0.03
	250 (32,000)	**0.03**	**0.03**	**0.04**	**0.03**	**0.04**	**0.03**	**0.04**	**0.04**	**0.03**	0.30	1.40	0.04
MI event age group (*)	50 (6,400)	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	0.07	1.02	0.02
	100 (12,800)	**0.03**	**0.03**	**0.02**	**0.02**	**0.02**	**0.02**	**0.03**	**0.02**	**0.03**	0.14	0.97	0.03
	150 (19,200)	**0.03**	**0.04**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	0.17	1.14	0.03
	200 (25,600)	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.04**	**0.04**	**0.04**	**0.03**	0.20	1.21	0.03
	250 (32,000)	**0.03**	**0.03**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**	**0.04**	0.26	1.27	0.04
Sex	50 (6,400)	**0.07**	**0.07**	**0.06**	**0.06**	**0.06**	**0.08**	**0.07**	**0.06**	**0.06**	0.20	4.94	0.07
	100 (12,800)	**0.12**	**0.12**	**0.11**	**0.13**	**0.13**	**0.12**	**0.13**	**0.13**	**0.13**	0.38	8.48	0.17
	150 (19,200)	**0.19**	**0.18**	**0.18**	**0.17**	**0.18**	**0.18**	**0.25**	**0.18**	**0.18**	0.58	12.27	0.25
	200 (25,600)	**0.27**	**0.24**	**0.24**	**0.27**	**0.24**	**0.24**	**0.26**	**0.25**	**0.25**	0.74	15.93	0.30
	250 (32,000)	**0.29**	**0.31**	**0.29**	**0.29**	**0.31**	**0.32**	**0.34**	**0.31**	**0.29**	0.91	18.99	0.37

All performance evaluations were computed based on the 5-fold test data sets. Table A.1 shows averaged G-mean scores from the ML models examined—SVC-linear, SVC-kernel, SVC-poly, LR, and MLP. Table A.2 shows the computation time as the dimension was reduced from 128*Q-dimensional SNP data to 128-dimensional SNP data.

As shown in Table A.1, NNDPs, excepting NNDP-OUT, provided the best performance of averaged G1 score for all Q cases for the 128*Q-dimensional data (Q ranges from 50 to 250) for all race, MI event age group, and sex classifications. UMAP provided similar good performance on MI event age group classification when compared to NNDP at Q=50. However, UMAP's performance significantly degraded from Q>=100 while NNDP maintained its good performance even for all higher-dimensional data. As shown in Table A.2, NNDP provided the fastest computation time (∼65.50 times) on the test data set. RP provided similarly efficient computation time compared to NNDP-IN. UMAP fared the worst, as it resulted in significantly higher computational times compared to either RP or NNDPs.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation* (pp. 265–283). https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Al-Husain, L., & Hafez, A. M. (2015). Dimensionality reduction approach for genotypic data. In *2015 IEEE conference on computational intelligence in bioinformatics and computational biology* (pp. 1–5). http://dx.doi.org/10.1109/CIBCB.2015.7300305.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In A. El Moataz, D. Mammass, A. Mansouri, & F. Nouboud (Eds.), *Image and signal processing* (pp. 317–325). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-51935-3_34.

Allen, L., Atkinson, J., Jayasundara, D., Cordiner, J., & Moghadam, P. Z. (2021). Data visualization for Industry 4.0: A stepping-stone toward a digital future, bridging the gap between academia and industry. *Patterns*, *2*(5), Article 100266. http://dx.doi.org/10.1016/j.patter.2021.100266.

Amalia, A., Sitompul, O. S., Nababan, E. B., & Mantoro, T. (2020). An efficient text classification using fasttext for bahasa Indonesia documents classification. In *2020 international conference on data science, artificial intelligence, and business analytics* (pp. 69–75). http://dx.doi.org/10.1109/DATABIA50434.2020.9190447.

Ambroziak, M., Niewczas-Wieprzowska, K., Maicka, A., & Budaj, A. (2020). Younger age of patients with myocardial infarction is associated with a higher number of relatives with a history of premature atherosclerosis. *BMC Cardiovascular Disorders*, *20*(1), 410. http://dx.doi.org/10.1186/s12872-020-01677-w.

de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, *17*(R2), R122–R128. http://dx.doi.org/10.1093/hmg/ddn288.

Bang, J.-U., Lee, M.-K., Yun, S., & Kim, S.-H. (2022). Improving end-to-end speech translation model with bert-based contextual information. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 6227–6231). http://dx.doi.org/10.1109/ICASSP43922.2022.9746117.

Baud, A., Mulligan, M. K., Casale, F. P., Ingels, J. F., Bohl, C. J., Callebert, J., Launay, J.-M., Krohn, J., Legarra, A., Williams, R. W., & Stegle, O. (2017). Genetic variation in the social environment contributes to health and disease. *PLOS Genetics*, *13*(1), Article e1006498. http://dx.doi.org/10.1371/journal.pgen.1006498.

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 245–250). http://dx.doi.org/10.1145/502512.502546.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. http://dx.doi.org/10.1162/tacl_a_00051.

Borah, A., Barman, M. P., & Awekar, A. (2021). Are word embedding methods stable and should we care about it? In *Proceedings of the 32st ACM conference on hypertext and social media* (pp. 45–55). http://dx.doi.org/10.1145/3465336.3475098.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 1877–1901).

Chen, I.-Y., Lypowy, J., Pain, J., Sayed, D., Grinberg, S., Alcendor, R. R., Sadoshima, J., & Abdellatif, M. (2006). Histone H2A.z is essential for cardiac myocyte hypertrophy but opposed by silent information regulator $2\alpha^*$. *Journal of Biological Chemistry*, *281*(28), 19369–19377. http://dx.doi.org/10.1074/jbc.M601443200.

Cheng, L., Qiu, Y., Schmidt, B. J., & Wei, G.-W. (2022). Review of applications and challenges of quantitative systems pharmacology modeling and machine learning for heart failure. *Journal of Pharmacokinetics and Pharmacodynamics*, *49*(1), 39–50. http://dx.doi.org/10.1007/s10928-021-09785-6.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 670–680). http://dx.doi.org/10.18653/v1/D17-1070.

De, R., Verma, S. S., Drenos, F., Holzinger, E. R., Holmes, M. V., Hall, M. A., Crosslin, D. R., Carrell, D. S., Hakonarson, H., Jarvik, G., Larson, E., Pacheco, J. A., Rasmussen-Torvik, L. J., Moore, C. B., Asselbergs, F. W., Moore, J. H.,

Ritchie, M. D., Keating, B. J., & Gilbert-Diamond, D. (2015). Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (QMDR). *BioData Mining*, *8*(1), 41. http://dx.doi.org/10.1186/s13040-015-0074-0.

Deagen, M. E., McCusker, J. P., Fateye, T., Stouffer, S., Brinson, L. C., McGuinness, D. L., & Schadler, L. S. (2022). FAIR and interactive data graphics from a scientific knowledge graph. *Scientific Data*, *9*(1), Article 1. http://dx.doi.org/10.1038/s41597-022-01352-z.

Devore, J. L. (2000). *Probability and statistics for engineering and the sciences* (5th ed.). Duxbury, https://www.gettextbooks.com/isbn/9780538733526/.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, *15*(11), Article e1008432. http://dx.doi.org/10.1371/journal.pgen.1008432.

Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. *Journal of Human Genetics*, *66*(1), Article 1. http://dx.doi.org/10.1038/s10038-020-00851-4.

Dong, F. N., Amiri-Yekta, A., Martinez, G., Saut, A., Tek, J., Stouvenel, L., Lorès, P., Karaouzène, T., Thierry-Mieg, N., Satre, V., Brouillet, S., Daneshipour, A., Hosseini, S. H., Bonhivers, M., Gourabi, H., Dulioust, E., Arnoult, C., Touré, A., Ray, P. F., ... Coutton, C. (2018). Absence of CFAP69 causes male infertility due to multiple morphological abnormalities of the flagella in human and mouse. *American Journal of Human Genetics*, *102*(4), 636–648. http://dx.doi.org/10.1016/j.ajhg.2018.03.007.

Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 975–983). http://dx.doi.org/10.1109/CVPR.2017.110.

Dong, W., Wozniak, M., Wu, J., Li, W., & Bai, Z. (2022). De-noising aggregation of graph neural networks by using principal component analysis. *IEEE Transactions on Industrial Informatics*, 1. http://dx.doi.org/10.1109/TII.2022.3156658.

Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, *11*(1), 1537. http://dx.doi.org/10.1038/s41467-020-15351-4.

Duan, W., Hicks, J., Makara, M. A., Ilkayeva, O., & Abraham, D. M. (2020). TASK-1 and TASK-3 channels modulate pressure overload-induced cardiac remodeling and dysfunction. *American Journal of Physiology. Heart and Circulatory Physiology*, *318*(3), H566–H580. http://dx.doi.org/10.1152/ajpheart.00739.2018.

Dürrschnabel, D., Hanika, T., & Stubbemann, M. (2022). FCA2vec: Embedding techniques for formal concept analysis. In R. Missaoui, L. Kwuida, & T. Abdessalem (Eds.), *Complex data analytics with formal concept analysis* (pp. 47–74). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-93278-7_3.

Dyke, M. V. (2018). Heart disease death rates among blacks and whites aged ≥35 years—United States, 1968–2015. *vol. 67*, In *MMWR. Surveillance summaries*. http://dx.doi.org/10.15585/mmwr.ss6705a1.

Egecioglu, O., Ferhatosmanoglu, H., & Ogras, U. (2004). Dimensionality reduction and similarity computation by inner-product approximations. *IEEE Transactions on Knowledge and Data Engineering*, *16*(6), 714–726. http://dx.doi.org/10.1109/TKDE.2004.9.

Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., Lin, H. J., Raffield, L., Gao, Y., Chen, H., de Vries, P., Lloyd-Jones, D. M., Lange, L. A., Peloso, G. M., Fornage, M., Rotter, J. I., Rich, S. S., Morrison, A. C., Psaty, B. M., ... Sofer, T. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology*, *5*(1), Article 1. http://dx.doi.org/10.1038/s42003-022-03812-z.

van der Ende, M. Y., Said, M. A., van Veldhuisen, D. J., Verweij, N., & van der Harst, P. (2018). Genome-wide studies of heart failure and endophenotypes: Lessons learned and future directions. *Cardiovascular Research*, *114*(9), 1209–1225. http://dx.doi.org/10.1093/cvr/cvy083.

Foody, G. M. (2009). Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, *113*(8), 1658–1663. http://dx.doi.org/10.1016/j.rse.2009.03.014.

Fuhrman, J. D., Gorre, N., Hu, Q., Li, H., Naqa, I. El., & Giger, M. L. (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics*, *49*(1), 1–14. http://dx.doi.org/10.1002/mp.15359.

Garcia, M., Almuwaqqat, Z., Moazzami, K., Young, A., Lima, B. B., Sullivan, S., Kaseer, B., Lewis, T. T., Hammadah, M., Levantsevych, O., Elon, L., Bremner, J. D., Raggi, P., Shah, A. J., Quyyumi, A. A., & Vaccarino, V. (2021). Racial disparities in adverse cardiovascular outcomes after a myocardial infarction in Young or middle-aged patients. *Journal of the American Heart Association*, *10*(17), Article e020828. http://dx.doi.org/10.1161/JAHA.121.020828.

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., & König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*, *44*(2), 125–138. http://dx.doi.org/10.1002/gepi.22279.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (4th ed.). The Johns Hopkins University Press.

Gould, D. B., & Walter, M. A. (2000). Cloning, characterization, localization, and mutational screening of the human BARX1 gene. *Genomics*, *68*(3), 336–342. http://dx.doi.org/10.1006/geno.2000.6307.

Hajar, R. (2017). Risk factors for coronary artery disease: Historical perspectives. *Heart Views : The Official Journal of the Gulf Heart Association*, *18*(3), 109–114. http://dx.doi.org/10.4103/HEARTVIEWS.HEARTVIEWS_106_17.

Huang, H., Wang, Y., Rudin, C., & Browne, E. P. (2022). Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications Biology*, *5*(1), Article 1. http://dx.doi.org/10.1038/s42003-022-03628-x.

Jayasena, C. N., & Sironen, A. (2021). Diagnostics and management of male infertility in primary ciliary dyskinesia. *Diagnostics*, *11*(9), 1550. http://dx.doi.org/10.3390/diagnostics11091550.

Johnson, William B. (1984). Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, *26*, 189–206.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), http://dx.doi.org/10.1098/rsta.2015.0202.

Jombart, T. S., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, *101*, 92–103. http://dx.doi.org/10.1038/hdy.2008.34.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers* (pp. 427–431). https://aclanthology.org/E17-2068.

Kowsher, M., Sobuj, M. S. I., Shahriar, M. F., Prottasha, N. J., Arefin, M. S., Dhar, P. K., & Koshiba, T. (2022). An enhanced neural word embedding model for transfer learning. *Applied Sciences*, *12*(6), Article 6. http://dx.doi.org/10.3390/app12062848.

Krämer, A., Green, J., Billaud, J.-N., Pasare, N. A., Jones, M., & Tugendreich, S. (2022). Mining hidden knowledge: Embedding models of cause–effect relationships curated from the biomedical literature. *Bioinformatics Advances*, *2*(1), http://dx.doi.org/10.1093/bioadv/vbac022, vbac022.

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019). An automated new approach in fast text classification (fastText): A case study for Turkish text classification without pre-processing. In *Proceedings of the 2019 3rd international conference on natural language processing and information retrieval* (pp. 1–4). http://dx.doi.org/10.1145/3342827.3342828.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*. arXiv:1909.11942.

Lee, A. K., Corneille, M. A., Hall, N. M., Yancu, C. N., & Myers, M. (2016). The stressors of being young and black: Cardiovascular health and black young adults. *Psychology & Health*, *31*(5), 578–591. http://dx.doi.org/10.1080/08870446.2015.1127373.

Lee, S. R., Lee, Y. H., Yang, H., Lee, H. W., Lee, G.-S., An, B.-S., Jeung, E.-B., Park, B.-K., & Hong, E.-J. (2019). Sex hormone-binding globulin suppresses NAFLD-triggered hepatocarcinogenesis after menopause. *Carcinogenesis*, *40*(8), 1031–1041. http://dx.doi.org/10.1093/carcin/bgz107.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 302–308). http://dx.doi.org/10.3115/v1/P14-2050.

Li, W., Cerise, J. E., Yang, Y., & Han, H. (2017). Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology*, *15*(4), Article 1750017. http://dx.doi.org/10.1142/S0219720017500172.

Li, J., Dan, J., Li, C., & Wu, R. (2014). A model-free approach for detecting interactions in genetic association studies. *Briefings in Bioinformatics*, *15*(6), 1057–1068. http://dx.doi.org/10.1093/bib/bbt082.

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, *38*, 4–9. http://dx.doi.org/10.1016/j.tacc.2021.02.007.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2019). Can deep learning techniques improve classification performance of vandalism detection in wikipedia? *Engineering Applications of Artificial Intelligence*, *78*, 248–259. http://dx.doi.org/10.1016/j.engappai.2018.11.012.

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction*. arXiv:1802.03426.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. ICLR.

Mimno, D., & Thompson, L. (2017). The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2873–2878). http://dx.doi.org/10.18653/v1/D17-1308.

Monk, B., Rajkovic, A., Petrus, S., Rajkovic, A., Gaasterland, T., & Malinow, R. (2021). A machine learning method to identify genetic variants potentially associated with alzheimer's disease. *Frontiers in Genetics*, *12*(642), http://dx.doi.org/10.3389/fgene.2021.647436.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, *37*(12), 1482–1492. http://dx.doi.org/10.1038/s41587-019-0336-3.

Moon, J., Posada-Quintero, H. F., & Chon, K. H. (2023). A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information. *Expert Systems with Applications*, *213*, Article 118930. http://dx.doi.org/10.1016/j.eswa.2022.118930.

Moon, J., Posada-Quintero, H. F., Kim, I., & Chon, K. H. (2021). Preliminary analysis of the risk factor identification embedding model for cardiovascular disease. *vol. 2021*, In *Annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology society. Annual international conference* (pp. 1946–1949). http://dx.doi.org/10.1109/EMBC46164.2021.9630039.

Moon, J., Yun, S., Lee, D., & Kim, S. (2019). A preliminary study on topical model for multi-domain speech recognition via word embedding vector. In *2019 34th international technical conference on circuits/systems, computers and communications* (pp. 1–4). http://dx.doi.org/10.1109/ITC-CSCC.2019.8793299.

Morris, J. A., Randall, J. C., Maller, J. B., & Barrett, J. C. (2010). Evoker: A visualization tool for genotype intensity data. *Bioinformatics (Oxford, England)*, *26*(14), 1786–1787. http://dx.doi.org/10.1093/bioinformatics/btq280.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, *7*, 19143–19165. http://dx.doi.org/10.1109/ACCESS.2019.2896880.

O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., Swedlow, J. R., Vuong, J., & Procter, J. B. (2018). Visualization of biomedical data. *Annual Review of Biomedical Data Science*, *1*(1), 275–304. http://dx.doi.org/10.1146/annurev-biodatasci-080917-013424.

Pan, Z., Sun, X., Shan, H., Wang, N., Wang, J., Ren, J., Feng, S., Xie, L., Lu, C., Yuan, Y., Zhang, Y., Wang, Y., Lu, Y., & Yang, B. (2012). MicroRNA-101 inhibited postinfarct cardiac fibrosis and improved left ventricular compliance via the FBJ osteosarcoma oncogene/transforming growth factor-β1 pathway. *Circulation*, *126*(7), 840–850. http://dx.doi.org/10.1161/CIRCULATIONAHA.112.094524.

Patel, N., Jhadav, B., Aljouie, A., & Roshan, U. (2015). Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning. In *2015 IEEE international conference on bioinformatics and biomedicine* (pp. 1367–1374). http://dx.doi.org/10.1109/BIBM.2015.7359878.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). http://dx.doi.org/10.3115/v1/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 2227–2237). http://dx.doi.org/10.18653/v1/N18-1202.

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., ... Duncan, L. E. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunitie, methods, pitfalls, and recommendations. *Cell*, *179*(3), 589–603. http://dx.doi.org/10.1016/j.cell.2019.08.051.

Pottmeier, P., Doszyn, O., Peuckert, C., & Jazin, E. (2020). Increased expression of Y-encoded demethylases during differentiation of human male neural stem cells. *Stem Cells and Development*, *29*(23), 1497–1509. http://dx.doi.org/10.1089/scd.2020.0138.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Rauber, P. E., Fadel, S. G., Falcão, A. X., & Telea, A. C. (2017). Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 101–110. http://dx.doi.org/10.1109/TVCG.2016.2598838.

Raucci, A., Maggio, S. Di., Scavello, F., D'Ambrosio, A., Bianchi, M. E., & Capogrossi, M. C. (2019). The janus face of HMGB1 in heart disease: A necessary update. *Cellular and Molecular Life Sciences*, *76*(2), 211–229. http://dx.doi.org/10.1007/s00018-018-2930-9.

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., & Vilo, J. (2017). Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS ONE*, *12*(7), Article e0179238. http://dx.doi.org/10.1371/journal.pone.0179238.

Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., & Mitra, M. (2018). Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1835–1838). http://dx.doi.org/10.1145/3269206.3269277.

Rožanec, J. M., Fortuna, B., & Mladenić, D. (2022). Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI). *Information Fusion*, *81*, 91–102. http://dx.doi.org/10.1016/j.inffus.2021.11.015.

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Too, C. Lai., Arayssi, T., Hammoudeh, M., Al Emadi, S., Masri, B. K., Halabi, H., Badsha, H., Uthman, I. W., Saxena, R., Padyukov, L., Hirata, M., Matsuda, K., Murakami, Y., Kamatani, Y., & Okada, Y. (2020). Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nature Communications*, *11*(1), 1569. http://dx.doi.org/10.1038/s41467-020-15194-z.

Sang, S., Liu, X., Chen, X., & Zhao, D. (2020). A scalable embedding based neural network method for discovering knowledge from biomedical literature. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1. http://dx.doi.org/10.1109/TCBB.2020.3003947.

Šeda, O., Liska, F., Sedova, L., Kazdová, L., Krenová, D., & Kren, V. (2005). A 14-gene region of rat chromosome 8 in SHR-derived polydactylous congenic substrain affects muscle-specific insulin resistance, dyslipidaemia and visceral adiposity. *Folia Biologica*, *51*, 53–61.

Shanks, M. O., Lund, L. M., Manni, S., Russell, M., Mauban, J. R. H., & Bond, M. (2012). Chromodomain helicase binding protein 8 (Chd8) is a novel A-kinase anchoring protein expressed during rat cardiac development. *PLoS One*, *7*(10), Article e46316. http://dx.doi.org/10.1371/journal.pone.0046316.

Shibata, N., Kondo, T., Morimoto, R., Kazama, S., Sawamura, A., Nishiyama, I., Kato, T., Kuwayama, T., Hiraiwa, H., Umemoto, N., Asai, T., Okumura, T., & Murohara, T. (2022). Clinical value of the HATCH score for predicting adverse outcomes in patients with heart failure. *Heart and Vessels*, *37*(8), 1363–1372. http://dx.doi.org/10.1007/s00380-022-02035-w.

Shimizu, R., Matsutani, M., & Goto, M. (2022). An explainable recommendation framework based on an improved knowledge graph attention network with massive volumes of side information. *Knowledge-Based Systems*, *239*, Article 107970. http://dx.doi.org/10.1016/j.knosys.2021.107970.

Silva, P. P., Gaudillo, J. D., Vilela, J. A., Roxas-Villanueva, R. M. L., Tiangco, B. J., Domingo, M. R., & Albia, J. R. (2022). A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. *Scientific Reports*, *12*(1), Article 1. http://dx.doi.org/10.1038/s41598-022-19708-1.

Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2021). Fast WordPiece tokenization. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2089–2103). http://dx.doi.org/10.18653/v1/2021.emnlp-main.160.

Soumare, H., Rezgui, S., Gmati, N., & Benkahla, A. (2021). New neural network classification method for individuals ancestry prediction from SNPs data. *BioData Mining*, *14*(1), 30. http://dx.doi.org/10.1186/s13040-021-00258-7.

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, *6*, Article 2055207620914777. http://dx.doi.org/10.1177/2055207620914777.

Suhaili, S. Mohamad., Salim, N., & Jambli, M. N. (2021). Service chatbots: A systematic review. *Expert Systems with Applications*, *184*, Article 115461. http://dx.doi.org/10.1016/j.eswa.2021.115461.

Thioulouse, J., Chessel, D., & Champely, S. (1995). Multivariate analysis of spatial patterns: A unified approach to local and global structures. *Environmental and Ecological Statistics*, *2*(1), 1–14. http://dx.doi.org/10.1007/BF00452928.

Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, *302*, Article 103627. http://dx.doi.org/10.1016/j.artint.2021.103627.

Van Wyngene, L., Vanderhaeghen, T., Petta, I., Timmermans, S., Corbeels, K., Van der Schueren, B., Vandewalle, J., Van Looveren, K., Wallaeys, C., Eggermont, M., Dewaele, S., Catrysse, L., van Loo, G., Beyaert, R., Vangoitsenhoven, R., Nakayama, T., Tavernier, J., De Bosscher, K., & Libert, C. (2021). ZBTB32 performs crosstalk with the glucocorticoid receptor and is crucial in glucocorticoid responses to starvation. *IScience*, *24*(7), Article 102790. http://dx.doi.org/10.1016/j.isci.2021.102790.

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 1–73.

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, http://dx.doi.org/10.1017/ATSIP.2019.12.

Wang, S., Zhou, W., & Jiang, C. (2020). A survey of word embeddings based on deep learning. *Computing*, *102*(3), 717–740. http://dx.doi.org/10.1007/s00607-019-00768-7.

Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 2092–2102). http://dx.doi.org/10.18653/v1/N18-1190.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). HuggingFace's transformers: State-of-the-art natural language. In *Processing*. arXiv:1910.03771.

Yang, C.-H., Chuang, L.-Y., & Lin, Y.-D. (2018). Multiobjective multifactor dimensionality reduction to detect SNP–SNP interactions. *Bioinformatics*, *34*(13), 2228–2236. http://dx.doi.org/10.1093/bioinformatics/bty076.

Yao, L., Zhang, Y., Chen, Q., Qian, H., Wei, B., & Hu, Z. (2017). Mining coherent topics in documents using word embeddings and large-scale text data. *Engineering Applications of Artificial Intelligence*, *64*, 432–439. http://dx.doi.org/10.1016/j.engappai.2017.06.024.

Yin, Z., Gonzales, L., Kolla, V., Rath, N., Zhang, Y., Lu, M. M., Kimura, S., Ballard, P. L., Beers, M. F., Epstein, J. A., & Morrisey, E. E. (2006). Hop functions downstream of nkx2.1 and GATA6 to mediate HDAC-dependent negative regulation of pulmonary gene expression. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, *291*(2), L191–L199. http://dx.doi.org/10.1152/ajplung.00385.2005.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Zhang, Y., & Yao, Q. (2022). Knowledge graph reasoning with relational digraph. In *Proceedings of the ACM web conference 2022* (pp. 912–924). http://dx.doi.org/10.1145/3485447.3512008.

Zhou, B., Yang, G., Shi, Z., & Ma, S. (2022). Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*, 1–17. http://dx.doi.org/10.1109/RBME.2022.3210270.