5-2022

Non-Negative Discriminative Data Analytics

Md Imrul Kaish
The University of Texas Rio Grande Valley

NON-NEGATIVE DISCRIMINATIVE DATA ANALYTICS

A Thesis

by

MD IMRUL KAISH

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Electrical Engineering

The University of Texas Rio Grande Valley

May 2022

NON-NEGATIVE DISCRIMINATIVE DATA ANALYTICS

A Thesis
by
MD IMRUL KAISH

COMMITTEE MEMBERS

Dr. Mostafizur Rahman
Chair of Committee

Dr. Mark Chu
Committee Member

Dr. Nantakan Wongkasem
Committee Member

Dr. Jia Chen
Committee Member

May 2022

ABSTRACT

Kaish, Md Imrul, Non-negative Discriminative Data Analytics. Master of Science in Engineering (MSE), May, 2022, 52 pp., 4 tables, 23 figures, references, 87 titles.

Due to advancements in data acquisition techniques, collecting datasets representing samples from multi-views has become more common recently (Jia et al. 2019). For instance, in genomics, a lymphoma patient's dataset may include data on gene expression, single nucleotide polymorphism (SNP), and array Comparative genomic hybridization (aCGH) measurements. Learning from multiple views about the same objective, in general, obtains a better understanding of the hidden patterns of the data compared to learning from a single view data. Most of the existing multi-view learning techniques such as canonical correlation analysis (Hotelling et al. 1936) and multi-view support vector machine (Farquhar et al. 2006), multiple kernel learning (Zhang et al. 2016) are focused on extracting the shared information among multiple datasets.

However, in some real-world applications, it's appealing to extract the discriminative knowledge of multiple datasets, namely discriminative data analytics. For example, consider the one dataset as gene-expression measurements of cancer patients, and the other dataset as the gene-expression levels of healthy volunteers and the goal is to cluster cancer patients according to the molecular sub-types. Performing a single view analysis such as principal component analysis (PCA) on any of the dataset yields information related to the common knowledge between the two datasets (Garte et al. 1996). Addressing such challenge, contrastive PCA (Abid et al. 2017) and discriminative (d) PCA in (Jia et al. 2019) are proposed in to extract one dataset-specific information often missed by PCA.

Inspired by dPCA, we propose a novel discriminative multi-view learning algorithm, namely Non-negative Discriminative Analysis (DNA), to extract the unique information of one dataset (a.k.a. view) with respect to the other dataset. This boils down to solving a non-negative matrix factorization problem. Furthermore, we apply the proposed DNA framework in various real-world down-stream machine learning applications such as feature selections, dimensionality reduction, classification, and clustering.

DEDICATION

The completion of my master’s studies would not have been possible without the love and support of my family, continuous mentoring from Dr. Jia Chen and Dr. Mostafizur Rahman. Thank you for your continuous support.

ACKNOWLEDGMENTS

I will always be grateful to Dr. Jia Chen, former chair of my thesis committee, and Dr. Mostafizur Rahman, current chair of my thesis committee, for all their mentoring and advice. They encouraged me to complete this process through their patience and guidance. I learned to be more humble, positive, perform continuous learning and never give up, from Dr. Chen. From Dr. Rahman, I learned to be punctual, got more ideas about real-world scenario in professional life, to be more reliable in professional settings and skip procrastination. I am very grateful because they added some important values in my life. Also, my thanks go to my other thesis committee members: Dr. Nantakan Wongkasem, and Dr. Mark Chu for their guidance. Their advice, input, and comments on my dissertation helped to ensure the quality of my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

After reading about discriminative analysis and observing its remarkable performance in information detection in noisy environments, we became interested in exploring its application potential. COVID-19 has been active since March 2020 but raising awareness among the general public has been slow. Having information about the pandemic early would allow us to be more prepared and prevent its severity. Motivated by this scenario and after a detailed understanding of what discriminative analysis can do, we developed a novel discriminative analysis method that can detect unique information about the COVID-19 era with respect to previous times. Through regular discriminative analysis between two consecutive time periods, it would be possible to detect any outbreaks early and take necessary precautions to prevent them as much as possible. Due to advancement in data acquisition techniques, it takes a lot of time if we want to train a model with all data available. Proposed discriminative analysis method assigns a weight to each feature in a dataset which can be very useful in applications like feature selection. Analyzing only the most important features reduced the learning time and resulted in better prediction performance.

Multi-view learning is a new approach to machine learning that incorporates learning from several perspectives in order to increase generalization performance. Data fusion or data integration from several feature sets is another name for multi-view learning. When it comes to web page classification, there are usually two ways to describe a page: the text content of the page itself and the anchor text of any web page that links to it. A well-designed multi-view learning technique can provide better performance in classification when both information is available. From literature, we find three types of multi-view learning algorithms: 1) co-training (Yu et al. 2007), 2) multiple kernel learning (Gonen et al. 2011), and 3) subspace learning (Hotelling et al. 1936). Canonical Correlation Analysis (CCA) (Hotelling et al. 1936) explore basis vectors for two sets of variables by mutually maximizing the correlations between the projections onto these basis vectors in order to find the shared latent subspace. CCA has been used to find risk factor for recurrence of breast cancer in (Sadoughi et al. 2016), audio visual synchronization (Sargin et al. 2007).

However, CCA only reveals the correlation between pairwise samples, which cannot adequately describe the similarity between samples in the same class or evaluate the dissimilarity between samples in different classes. Discriminative multi-view learning methods (e.g., discriminative CCA) can tackle this problem. It can learn the latent subspace where within-class correlation is maximized, and the inter-class correlation is minimized. Discriminative CCA has been used for feature extraction for recognition of course categories in Web-KB course dataset in (Sun et al. 2007). Also, discriminative Principal Component Analysis (dPCA) has been proposed in (Chen et al. 2019) which can provide least-squares optimal in recovering the latent subspace specific to target data against background data. In that paper, dPCA has been applied in health data to distinguish between two types of body movement: lying down and frontal elevation of arms.

Discriminative analysis is a dimensionality reduction problem in a broad manner. The transfer of data from a high-dimensional space to a low-dimensional space so that the low-dimensional representation retains most significant aspects of the original data, is known as dimensionality reduction. Feature selection (Kumar et al. 2014), matrix factorization (Lee et al. 1999) and manifold learning (Cayton et al. 2005) are some areas of application for dimensionality reduction. Principal Component Analysis (PCA) (Hotelling et al. 1936) and Non-negative Matrix Factorization (NMF) (Lee et al. 1999) are among some popular classical dimensionality reduction methods. Discriminative analysis is getting more attention recently as a dimensionality reduction method. In a 2019 paper (Zhao et al. 2019), leveraging low dimensional representation, using discriminative analysis, classification of face data has been performed.

Discriminative analysis can often be very useful to obtain unique discriminative information in between two datasets. For example, in 2019, the word "COVID-19" wasn't even familiar but in 2020 it had shown great domination all over the world. In this case, discriminative information can provide knowledge about possible outbreaks in the whole world. On the other hand, discriminative analysis is also capable of finding how important each feature is, in order to represent the target concept in a particular dataset (Luo et al. 2016). From different experimental results it has been shown that only $10 - 20\%$ of the original features actually are responsible for the quality of performance (T. Huang et al. 2012). So, if we can rank features in a dataset, it is possible to separate important features and reduce the size of the dataset and convergence time of the model, in a great extent.

In this paper, we developed DNA (Non-negative Discriminative Analysis), a novel non-negative discriminative principal component analysis, which takes two datasets as inputs: background and target dataset and give us discriminative information of target dataset in contrast to background dataset. The objective of the thesis is followings:

1. To propose a novel discriminative analysis model named DNA.
2. To investigate Google Trends COVID-19 symptoms dataset for year of 2018 to 2020 and to extract unique information related to COVID-19 using DNA.
3. To generate feature importance or feature ranking set in some standard datasets (e.g., CIFAR10).
4. To demonstrate the efficacy of feature selection using DNA in supervised and unsupervised learning settings.

This manuscript is organized in five chapters. In chapter one, we discussed background of the thesis, motivations, and contributions of the thesis. Chapter two discusses detailed literature review of the topics involved in the thesis. Extensive discussion on methodology can be found in chapter three. In chapter four we have demonstrated numerical results using different datasets. Finally, chapter five concludes our thesis and opens directions for future work.

CHAPTER II

LITERATURE REVIEW

Multi-view learning is a new approach to machine learning that incorporates learning from several perspectives to increase generalization performance. Data fusion or data integration from numerous feature sets are other terms for it. As single-view data cannot adequately convey the information of all samples, several data are frequently collected through various measuring methods. For example, any video with audio can be considered as a multi-view data: one view is the video frames, and another view is the audio signal. The title, keywords, and citations in a journal's dataset can be thought of as three independent perspectives on a single paper (R. Bro et al. 1997). A well-designed multi-view learning technique can help to increase performance of the learning model. Multi-view learning has a wide range of applications, including dimensionality reduction (Chen X et al. 2012; Hardoon D et al. 2011; White M et al. 2012), semi-supervised learning, supervised learning, and clustering. Because the background dataset and the target dataset can be regarded as two viewpoints, discriminative learning is a special sort of multi-view learning. This chapter gives a quick summary of previous works in dimensionality reduction and discriminative analysis from several angles.

2.1 Dimensionality Reduction

The dimensionality of real-world data, such as speech signals, digital pictures, or functional Magnetic Resonance Imaging (fMRI) scans, is typically high. The dimensionality of such real-world data must be decreased to handle it properly. The transformation of high-dimensional data into a comprehensible representation with decreased dimensionality is known as dimensionality reduction. It is the translation of data from a high-dimensional space to a low-dimensional space in such a way that the low-dimensional representation retains maximum amount of the original data's relevant qualities as possible. Working with high-dimensional environments can be inconvenient for a variety of reasons. The curse of dimensionality (Xu et al. 2013) and other undesirable properties of high-dimensional spaces are mitigated via dimensionality reduction, which is popular in many disciplines (L.O. Jimenez et al. 1997). Dimensionality reduction makes it easier to classify, visualize, and compress high-dimensional data. Traditionally, linear techniques such as Principal Components Analysis (PCA) (K. Pearson et al. 1901), factor analysis (C. Spearman et al. 1904), and classical scaling (W.S. Torgerson et al. 1952) were used to reduce dimensionality. Signal processing, speech recognition, neuro-informatics, and bioinformatics are among domains that use dimensionality reduction to deal with huge numbers of observations and/or variables.

2.1.1 Principal Component Analysis

Among variety of approaches have been developed for dimensionality reduction, principal component analysis (PCA) is one of the oldest and most utilized. The goal is to minimize a dataset's dimensionality while keeping as much variability (i.e., statistical information) as possible. This means that, keeping as much variety as feasible, involves identifying new variables that are linear functions of the original dataset's variables, maximize variance sequentially, and are uncorrelated with one another. Finding correlation manually in thousands of features is nearly impossible, frustrating, and time-consuming. PCA does this efficiently. Also, PCA helps in overcoming the overfitting issue by reducing the number of features as overfitting occurs when there is a lot of features. After implementing the PCA on the dataset, all the Principal Components (PCs) are independent of one another. In (S. Zhang et al. 2013), PCA has been used for Speech Emotion Recognition (SER) to transform the high dimensional feature space to a lower dimension. The principal components are a type of new variable that can be discovered by solving an eigenvalue/eigenvector issue. (Pearson K. et al. 1901) and (Hotelling H. et al. 1933) are the first papers on PCA, but it wasn't until electronic computers became widely available that it was computationally practical to utilize it on datasets that weren't trivially small. It performs a linear mapping of the data to a lower-dimensional space to maximize the data's variance in the low-dimensional representation. In practice, the covariance (and, on rare occasions, correlation) matrix of the data is constructed, and the eigenvectors are computed on this matrix. The principal components (eigenvectors that correspond to the biggest eigenvalues) can now be utilized to recover a significant percentage of the original data's variance. Furthermore, because the first few eigenvectors commonly contribute most of the system's energy, particularly in low-dimensional systems, they can frequently be interpreted in terms of the system's large-scale physical behavior.

2.1.2 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a matrix decomposition technique that splits a non-negative matrix into two low-rank non-negative matrices (Lee et al. 1999; Suvrit et al. 2006). It has been a promising tool in fields where only non-negative signals exist, such as astronomy. The nonnegativity of NMF, as opposed to other factorization methods such as Singular Value Decomposition (SVD), enables only additive combinations of intrinsic components (i.e., hidden features). This is shown in (Lee DD et al. 1999) where NMF learns face pieces and a face is naturally represented as an additive linear combination of distinct parts (Paatero et al. 1997; Tapper et al. 1994). Negative combinations might not feel as natural or intuitive as positive combinations. The resulting decomposed matrices have less entries than the original matrix, hence NMF is also a dimension reduction approach. This suggests that a decomposition does not require all the entries in the original matrix, hence NMF should be able to accommodate missing items in the target matrix. If the desired loss function is a sum of per-entry losses, such as mean square error (MSE) or Kullback-Leibler (KL) divergence, factorization can be achieved by deleting the loss items corresponding to the missing entries. NMF has become an essential tool in multivariate data analysis and has been widely used in the fields of machine learning (A. Cichocki et al. 2009), data mining (M. Berry et al. 2007), signal processing (I. Buciu et al. 2008), image engineering and computer vision (I. Buciu et al. 2008) due to the enhanced semantic interpretability provided by nonnegativity and the resulting sparsity.

2.2 Feature Selection

(Girish Chandrashekar et al. 2014) gives a complete overview of feature selection with filter, wrapper, and embedded approach. They also talked about applications of feature selection in the real world in text categorization, remote sensing, intrusion detection, genomic analysis, and image retrieval. In addition to that, the benefits, and drawbacks of feature selection methods for dealing with the various aspects of real-world applications, have been discussed. More information is not always better in machine learning applications, as feature selection techniques demonstrate. To compare feature selection approaches, the classifier accuracy and the number of decreased features has been used in (Kumar et al. 2014). Four categories of feature selection methods have been discussed in (Jundong et al. 2017): similarity-based, information theoretical-based, sparse-learning-based, and statistical-based methods. Recent feature selection research on conventional data, structured data, heterogenous data and streaming data, have been discussed in that paper. Thirty-two existing feature selection methods have been discussed in (Dash et al. 1997) which are categorized based on the combinations of generation procedure and evaluation function. Feature selection techniques used in bioinformatics are discussed in (Yvan et al. 2007). Bioinformatics researchers face two major problems dealing with related datasets: large input dimensionality and small sample size. Researchers in bioinformatics, machine learning, and data mining have devised a plethora of feature selection strategies to address these issues. Relief (Kira et al. 1992), a new algorithm was proposed which can pick relevant features using a statistical manner. Relief is noise-tolerant, does not rely on heuristics, and is accurate even when features are interrelated.

2.3 Shared Information Extraction using Multi-View Learning

Subspace learning-based approaches try to obtain a latent subspace shared by numerous views by assuming that the input views are derived from that latent subspace. It is especially effective in lowering the curse of dimensionality since the dimensionality of the latent subspace is lower than any input view (Xu et al. 2013). Given this subspace, it is very beneficial to carry out classification and clustering objectives. Canonical Correlation Analysis (CCA) and some tensor methods on finding latent subspace have been discussed in this section.

2.3.1 Canonical Correlation Analysis

The challenge of identifying basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized was proposed in (Hotelling et al. 1936; Borga et al. 1999). They represented the eigenproblem as two eigenvalue equations, because this reduces the calculation time and size of the eigenvectors. Kernel canonical correlation analysis (KCCA) (Akaho et al. 2001; Melzer et al. 2001) is an extension of Canonical Correlation Analysis (CCA) in which maximally correlated nonlinear projections are discovered, which CCA is unable to perform. Deep canonical correlation analysis (DCCA) has been proposed in (Andrew et al. 2013) which serves similar purpose like KCCA, but it does not require an inner product and has the benefits of a parametric method: training time scales well with data size, and the training data does not need to be referred when computing representations of unseen instances. CCA, KCCA, DCCA can only find correlation between two datasets. Deep Generalized Canonical Correlation Analysis (DGCCA) has been proposed in (Benton et al. 2017) which is the first multi-view representation learning technique based on CCA that combines the flexibility of nonlinear (deep) representation learning with the statistical power of combining data from several independent sources.

2.3.2 Tensor Methods

More recently, tensor decompositions for learning latent variable models, particularly topic models, have received a lot of attention. A general introduction to tensor decomposition is presented in (Bro et al. 1997). The application of Canonical Polyadic (CP) to sensor array processing was studied by Sidiropoulos, Bro, and Giannakis (N. Sidiropoulos et al. 2000). We explored Parallel Factor Analysis (PARAFAC) (Bro et al. 1997) algorithm for tensor decomposition which is a generalization of PCA to higher order arrays. The PARAFAC2 tensor decomposition is a generalization of the PARAFAC/CP tensor decomposition which is proposed in (Harshman et al. 1972). This tensor decomposition represents a collection of related matrix decompositions with one mode in common, i.e., one of the components varies along the set of matrices (tensor slices), while the other remains constant. However, the lack of efficient algorithms that can handle large-scale datasets has been its principal drawback to date. SPARTan: Scalable PARAFAC2 for Large & Sparse Data was proposed in (Perros et al. 2017) that bridges this gap by creating a scalable approach for computing the PARAFAC2 decomposition of large, sparse datasets. In the 1990s, signal processing became prominent, but it wasn't until about a decade ago that the computer science community (particularly those working in machine learning, data mining, and computing) realized the value of tensor decompositions (T. G. Kolda et al. 2005; E. Acar et al. 2005). There has been a lot of work on tensors decompositions recently, for learning latent variable models specially on Topic models (A. Anandkumar et al. 2014; Blei et al. 2003) discusses about the connections between orthogonal tensor decomposition and the method of moments for computing the Latent Dirichlet Allocation (LDA – a widely used topic model).

2.4 Discriminative Information Extraction using Multi-View Learning

In this section, we have discussed two discriminative multi-view analysis methods: contrastive Principal Component Analysis (cPCA) and discriminative Principal Component Analysis (dPCA). Purpose of these algorithms to extract latent subspace where discriminative information from multiple datasets is available.

2.4.1 Contrastive Principal Component Analysis

In order to find hidden patterns in one dataset relative to another dataset, contrastive PCA with one free parameter, alpha, has been proposed in (Abid et al. 2017). A wide variety of experiments was conducted in that paper using cPCA to find directions in which the target data varies significantly, but the background data does not. While PCA aims for identifying dominating trends in just one dataset, cPCA can find it in two datasets. The principal applications of cPCA are the same as those for which PCA is commonly used: efficiently reducing dimensions to enable visualization and exploratory data analysis. It takes about the same amount of time to compute a specific cPCA as it does to compute a standard PCA. cPCA does not attempt to classify individual data points; instead, it aims to display patterns unique to the target. Geographic ancestry clusters within Mexico have been visualized using cPCA using genotyping data (Abid et al. 2017). Subgroup discovery in protein expression data, single cell RNA-Seq data has been performed in another paper (Abid et al. 2018). An automatic scheme based on clustering of subspaces for selecting the most informative values of contrast parameter, alpha, has been proposed in that paper.

2.4.2 Discriminative Principal Component Analysis

Discriminative Principal Component Analysis (dPCA), proposed in (Jia et al. 2019), is least-squares optimal in recovering the latent subspace vector unique to the target data compared to background data under specific conditions. dPCA models for one or more background datasets are generalized using kernel-based learning to account for nonlinear data correlations. It can extract low-dimensional discriminative structure that is unique to the target data but not to numerous sets of background data. This is accomplished by increasing the variation of anticipated target data while lowering the sum of all projected background data variances. Applications of dPCA in classification of applied health data, sensor data, and face picture datasets, presented in (Jia et al. 2019). dPCA has also been used in near-infrared (NIR) spectroscopy datasets and performed multi-class classification (Liu et al. 2020). Discriminative properties have been used for wrist-hand movement detection for standalone, battery-powered EMG wearables in (Raurale et al. 2020).

CHAPTER III

PROPOSED WORK

In this chapter we have discussed about the problem statement of the thesis and our proposed model. Algorithm for the model and explanation on how we get unique information or feature importance set from a dataset, has been provided.

3.1 Problem Statement

Principal Component Analysis (PCA) (Hotelling H. et al. 1933) projects the data onto low dimensions and is especially powerful as an approach to visualize patterns, such as clusters and clines, in a dataset (Jolliffe et al. 2002). For this task, it is important to find out the correlation among the features (correlated variables). In contrast to PCA, the projected coefficients that are obtained using the Non-negative Matrix Factorization (NMF) method are only positive. The parts-based representation may be holistic, rather than local, depending on the type and nature of the data being studied. It is also possible that a parts-based, local representation may require fully hierarchical models with multiple levels of hidden variables rather than the single level used in this approach. When an analyst has multiple datasets (or multiple conditions in one dataset to compare), then the current state-of practice is to perform PCA (or t-distributed Stochastic Neighbor Embedding, Multi-dimensional Scaling, etc.) on each dataset separately, and then manually compare the various projections to explore if there are interesting similarities and differences across datasets (Chen et al. 2016; Zhou et al. 2007). Unlike PCA or NMF, discriminative analysis is not only capable of finding better representation of data but also it can extract unique information of one dataset relative to another. For example, in a genomics context, the foreground data could be gene expression measurements from patients with a disease, and the background data could be measurements from healthy patients (Twine et al., 2011; Zheng et al., 2017; Young et al., 2018). Clearly, PCA is not suitable in this contrastive setting because PCA only identifies structure that exists across the union of the two groups or structure in each group in isolation. Then the goal can be informally stated as finding directions in which the target data varies significantly, but the background data does not, and it efficiently identifies lower-dimensional subspaces that capture structure specific to the target data. Contrastive PCA is a good approach which can offer discriminative information in low dimensional structure. If the involved hyper-parameter is properly selected, which is carried out via singular value decomposition (SVD), cPCA can disclose dataset-specific information that is often ignored by normal PCA. Though it is possible to determine the optimum hyper-parameter from a list of potential values automatically, executing SVD numerous times in large-scale situations can be computationally intensive. Moreover, it fails to obtain unique information from real data in some cases (e.g., COVID-19 google trends data) (Md et al. 2021). It is worthy of investigation on a novel discriminative analysis method which can provide solutions to the existing problems and good performance in real data. We proposed non-negative discriminative analysis (DNA) which is well capable of finding unique information from one dataset relative to another dataset in real world setting and can perform feature selection efficiently.

3.2 Non-negative Discriminative Analysis

In this section, we have discussed our proposed algorithm for discriminative analysis. Consider two datasets: background dataset (denoted as $\{ y_i \in \mathbb{R}^{\mathrm{D}} \}_{i=1}^{n}$) which contains the information of Flu, e.g., Google Trends data in 2019 or 2018 when there was no COVID-19 but Flu, and target dataset (denoted as $\{ x_i \in \mathbb{R}^{\mathrm{D}} \}_{i=1}^{m}$) having the information of both Flu and COVID-19, e.g., Google Trends data in 2020. Here, D denotes the number of searched symptoms and i is time index. In literature, discriminative (d) principal component analysis (PCA) (Jia et al. 2018) and contrastive (c) PCA (Abid et al. 2017) performing such discriminative analysis on both the target and background datasets. Discriminative PCA seeks a projection matrix so that the ratio of the projected target data variance over that of the background data is maximized, while cPCA maximizes the difference between the target data variance and the background data variance.

Specifically, dPCA approach searches for subspace vectors, namely the columns of $U \in \mathbb{R}^{\mathrm{D}\times\mathrm{d}}$ with number of dimensions, $d \leq D$ by solving equation 8 (Jia et al. 2018).

$${}^{Max\ Tr}_{\qquad U}[\left(U^T C_y U\right)^{-1} U^T C_x U] \qquad (8)$$

where $C_x := \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_x)\,(x_i - \mu_x)^T \,\epsilon\, \mathbb{R}^{D\times D}$ representing the sample covariance of the target data with μ_x denoting the corresponding sample mean; C_y is the sample covariance of the background data. This is ratio trace maximization problem, and the columns of the optimal U are the right eigenvectors of $C_y^{-1} C_x$ associated with the top-d eigenvalues. The projections $\{ U^T x_i \in \mathbb{R}^D \}$ are the sought lower-dimensional representations of $\{ x_i \}$, where the $(r, l)th$ entry of U reveals the importance of the r-th feature/symptom to the l-th projected dimension. The challenge of using dPCA directly to uncover such symptom importance (in other words uncover discriminative symptoms of COVID-19 w.r.t. Flu) is the sign ambiguity. To bypass this challenge, we propose a novel nonnegative discriminative analysis, namely DNA, by performing nonnegative matrix factorization (NMF) on $C_y^{-1}C_x$. Specifically, we learn two nonnegative factorization matrices $W \in \mathbb{R}^{D \times d}$and $H \in \mathbb{R}^{d \times D}$ so that

$$C_y^{-1}C_x \approx WH \qquad (9)$$

Table 1: Algorithm for DNA

Algorithm 1: DNA.

```
1: Input: Nonzero-mean target and background data $\{x_i\}_{i=1}^{m}$ and $\{y_i\}_{i=1}^{n}$; number of dimensions $d$.
2: Construct covariance matrices of $\{x_i\}$ and $\{y_i\}$ to obtain $C_x$ and $C_y$.
3: Perform non-negative matrix decomposition on $C_y^{-1}C_x$ to obtain the two factorization components W and H.
4: Output: W and H.
```

One popular approach to solve (9) is to use the Kullback–Leibler (KL) divergence metric. The sought W will be used to estimate the importance of each symptom. Our DNA for nonnegative discriminative analytics of two datasets is summarized in Algorithm 4.

CHAPTER IV

NUMERICAL TESTS

In this section, we have demonstrated our results and findings we got from our study on unique information extraction and feature selection using DNA. Feature selection procedure is verified in both supervised and unsupervised learning environment. At first, we have discussed about the datasets used and then results obtained using those datasets.

4.1 Datasets

In this thesis, we have used three datasets to satisfy our works. Brief description of each dataset is provided here. For each case, there is a target dataset and a background dataset. Three datasets are Covid-19 google trends symptoms dataset, MNIST handwritten digits dataset and Cifar10 object images dataset.

4.1.1 Google Trends Covid-19 Symptom's Dataset

Google Trends is a free and easy-to-use dataset provided by Google (Google LLC. 2021). This anonymized, aggregated dataset depicts trends in symptom search patterns and is meant to aid academics in better understanding COVID-19's impact. Trends in search behaviors, according to public health specialists, may be useful in gaining a better understanding of how COVID-19 affects communities and perhaps in spotting epidemics sooner. It is not possible to presume that the data is a record of real-world clinical events or to use it for medical diagnosis, prognosis, or therapy. This information shows the number of Google searches for a wide range of symptoms, indications, and illnesses. We'll refer to all of them as symptoms in this paper to make things easy. The data covers hundreds of symptoms, including fever, difficulty breathing, and stress, and is based on the following factors: the prevalence of a symptom in Google searches, data quality, and privacy concerns. Number of searches linked to each of these symptoms are counted on a daily basis and categorize the data by geographic location. The generated dataset is a daily or weekly time series for each location that shows the relative frequency of symptom queries. A single search query can be associated with multiple symptoms. A search for "acid reflux with coughing up mucus," for example, yields three symptoms: cough, gastroesophageal reflux disease, and heartburn. Even though the dataset is being released in English, searches are being counted in other languages.

4.1.2 MNIST Handwritten Digits Dataset

A training set of 60,000 samples and a test set of 10,000 examples are available in the MNIST database of handwritten digits. It's a subset of the NIST's larger collection. In a fixed-size image, the digits have been size-normalized and centered. The original NIST black and white (bilevel) photos were size adjusted to fit in a 20x20 pixel box while maintaining the aspect ratio. By determining the center of mass of the pixels and translating the image to position this point at the center of the 28x28 field, the images were centered in a 28x28 image. When the digits are centered by bounding box rather than center of mass, the error rate improves with various classification methods (especially template-based approaches like Support Vector Machines (SVM) and K-nearest Neighbors (KNN)). The MNIST database was built using binary pictures of handwritten numbers from NIST's Special Database 3 and Special Database 1. SD-3 was originally designated as the training set, while SD-1 was designated as the test set by NIST.

SD-3 is far cleaner and easier to distinguish than SD-1 because SD-3 was obtained from Census Bureau personnel, whereas SD-1 was acquired from high school students. The MNIST training set consists of 30,000 SD-3 patterns and 30,000 SD-1 patterns. 5,000 patterns from SD-3 and 5,000 patterns from SD-1 made up the test set. Around 250 writers were represented in the 60,000-pattern training set. It was made sure that the training and test sets of writers were not the same. The training set was supplemented with purposely distorted replicas of the original training data. Shifts, scaling, skewing, and compression are used to create the distortions.

4.1.3 Cifar10 Object Image Dataset

The CIFAR-10 dataset is a well-known computer vision dataset for object recognition. It contains 60,000 32x32 color images divided into ten classes, each with 6,000 images. There are 50,000 images for training and 10,000 images for testing. Complete dataset is divided into five training batches and one test batch, each with 10,000 images. The test batch contains exactly 1000 photographs from each class, chosen at random. The remaining photographs are distributed in training batches in a random order; however, some batches may contain more images from one class than others. The training batches contain exactly 5,000 photos from each class between them. The classes are mutually exclusive in every way. Automobiles and trucks are not interchangeable. Sedans, SUVs, and other vehicles fall under the umbrella of "automobile." Only large trucks are classified as "trucks." Pickup trucks aren't included in either of these categories.

4.2 Unique Information Extraction from Google Trends Dataset

In this section, we used a subset of the COVID-19 Search Trends symptoms dataset to test the effectiveness of our proposed method. We selected the number of searches of three symptoms unique for COVID-19 including ageusia, shortness of breath, and anosmia and six symptoms which are shared by COVID-19 and Flu including vomiting, diarrhea, cough, fever, fatigue, and headache from all the 51 US states, on a daily basis in years 2018, 2019, and 2020. Our goal is to identify these three distinct symptoms in order to give evidence to substantiate the COVID-19 epidemic. Note that there was known spread of the COVID-19 virus in 2020 but not in 2018 or early 2019. In figure 1, we have shown time series data of some relevant symptoms (two unique COVID-19 symptoms and two symptoms shared by COVID-19 and Flu) from Google Trends that represents relative frequency of searches. Clearly, we can see that there is a rapid increase of those symptoms' searches when COVID-19 started to hit in March 2020.

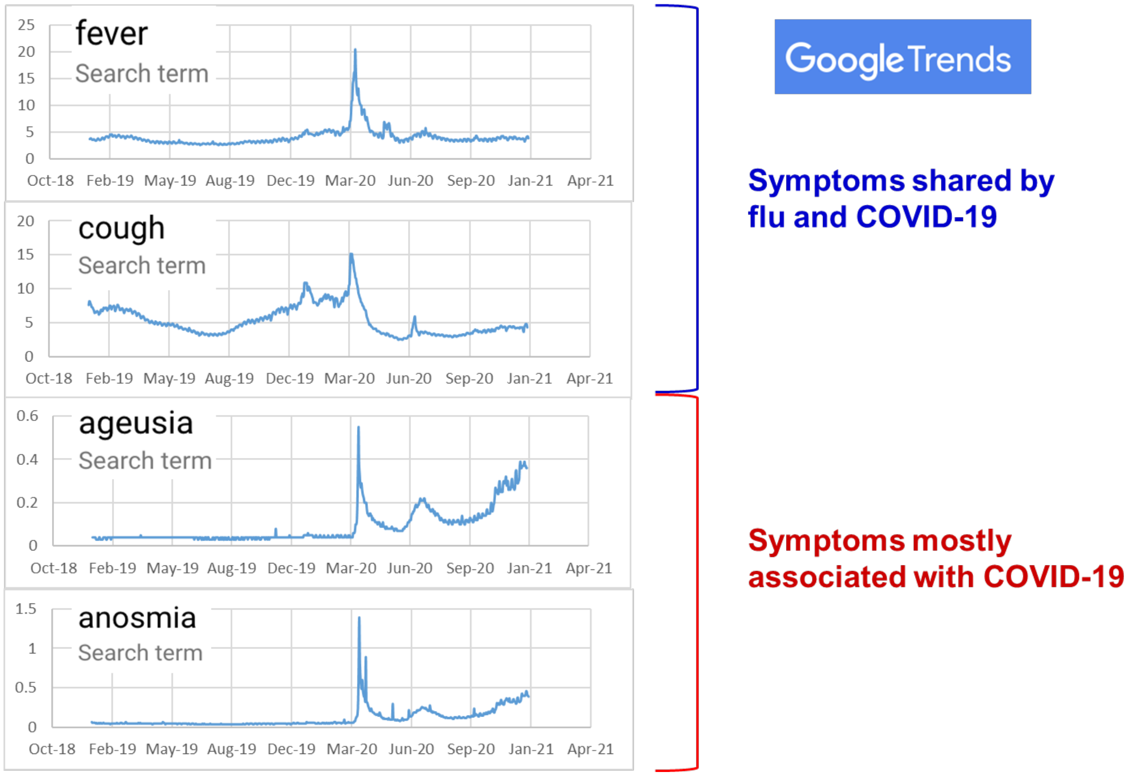


Figure 1: Covid-19 symptoms in Google trends data

Figure 2 shows Google trends data for four symptoms from year 2019 and 2020 side by side to demonstrate the growing search trend in 2020 with respect to 2019. 2020 data was chosen as the target data due to an increase in the number searches for COVID-19 symptoms within that period, whereas 2019 data was used as the background data because COVID-19 symptoms were not prevalent at that time.

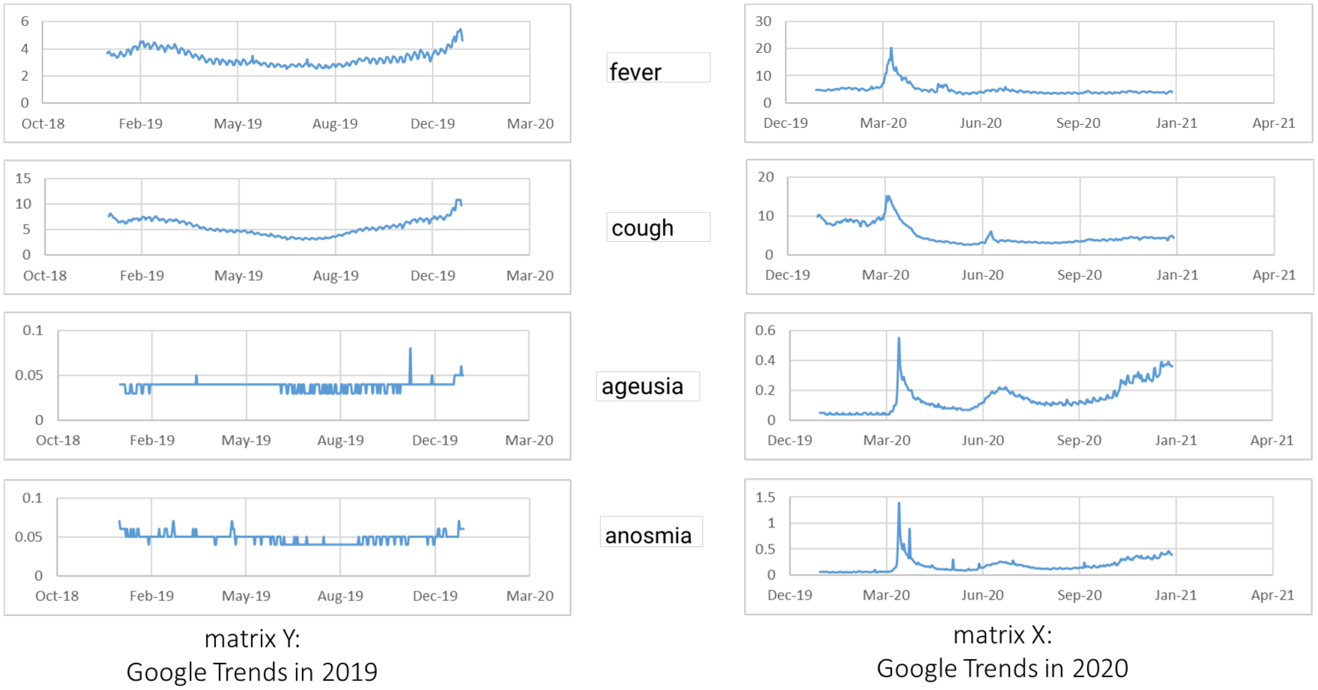


Figure 2: Generating Target and Background Data

4.2.1 Symptoms coefficients using DNA

Background data and target data have been denoted as $\{y_i\}_{i=1}^{n}$ and $\{x_i\}_{i=1}^{m}$ respectively with number of symptoms considered: $D = 9$, number of examples for target data: $m = 19, 032$ and number of examples for background data: $n = 18, 980$. The resulting mean and standard derivation of the symptom coefficients (a.k.a., the column values of W) after running the proposed DNA for 200 Monte Carlos tests which are shown in figure 3. We discovered that the mean of symptom coefficients for unique COVID-19 symptoms were 0.45, 0.36, and 0.68 respectively, but symptom coefficients for shared ones between COVID-19 and Flu symptoms were 0.020, 0.006, 0.001, 0.120, 0.005, and 0.018, respectively. COVID-19 symptoms have 89.76% contribution to the total symptom coefficients. Similarly, we set the 2018 search data as background data and 2020 searches as target data and plotted the results in figure 3 (right). Similar to the case when background was 2019 data, DNA can discover the unique symptoms of COVID-19 relative to those associated with Flu as unique symptoms have 90.18% contribution. It's worth to mention that the standard derivations of the symptoms with high coefficients in figure 3 are high because DNA doesn't admit unique solution and the order of the top symptoms vary a lot during different experiments.

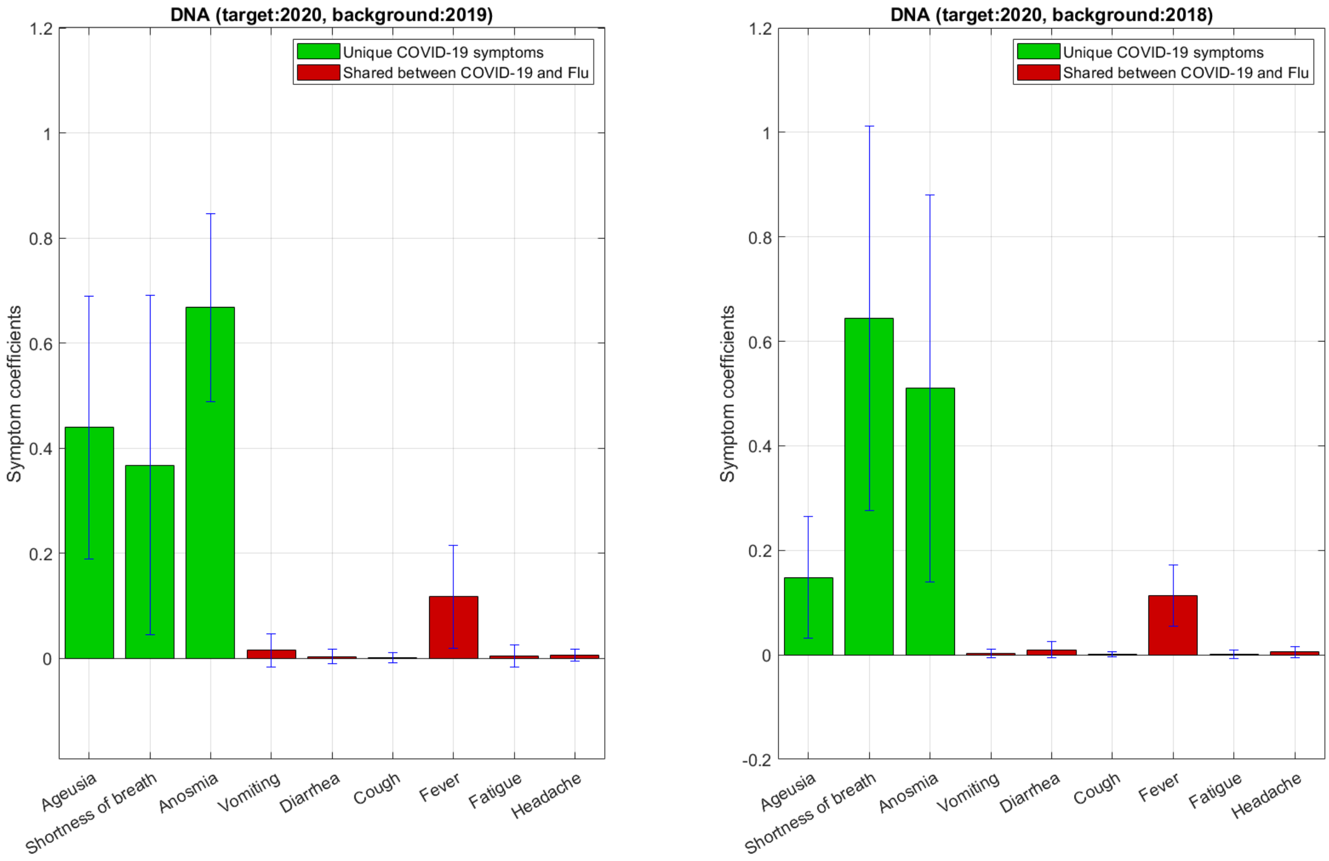


Figure 3: Symptom coefficients using DNA on 2020 data as target

Furthermore, when we set 2019 searches as the target data and 2020 and 2018 searches as the background data; see the results in the figure 4, as expected, DNA doesn't fully return the unique symptoms. This is because the unique COVID-19 symptoms are not the discriminative information of 2019 data relative to 2020 data. When background is 2020 data, contribution of unique symptoms is only 29.47%.

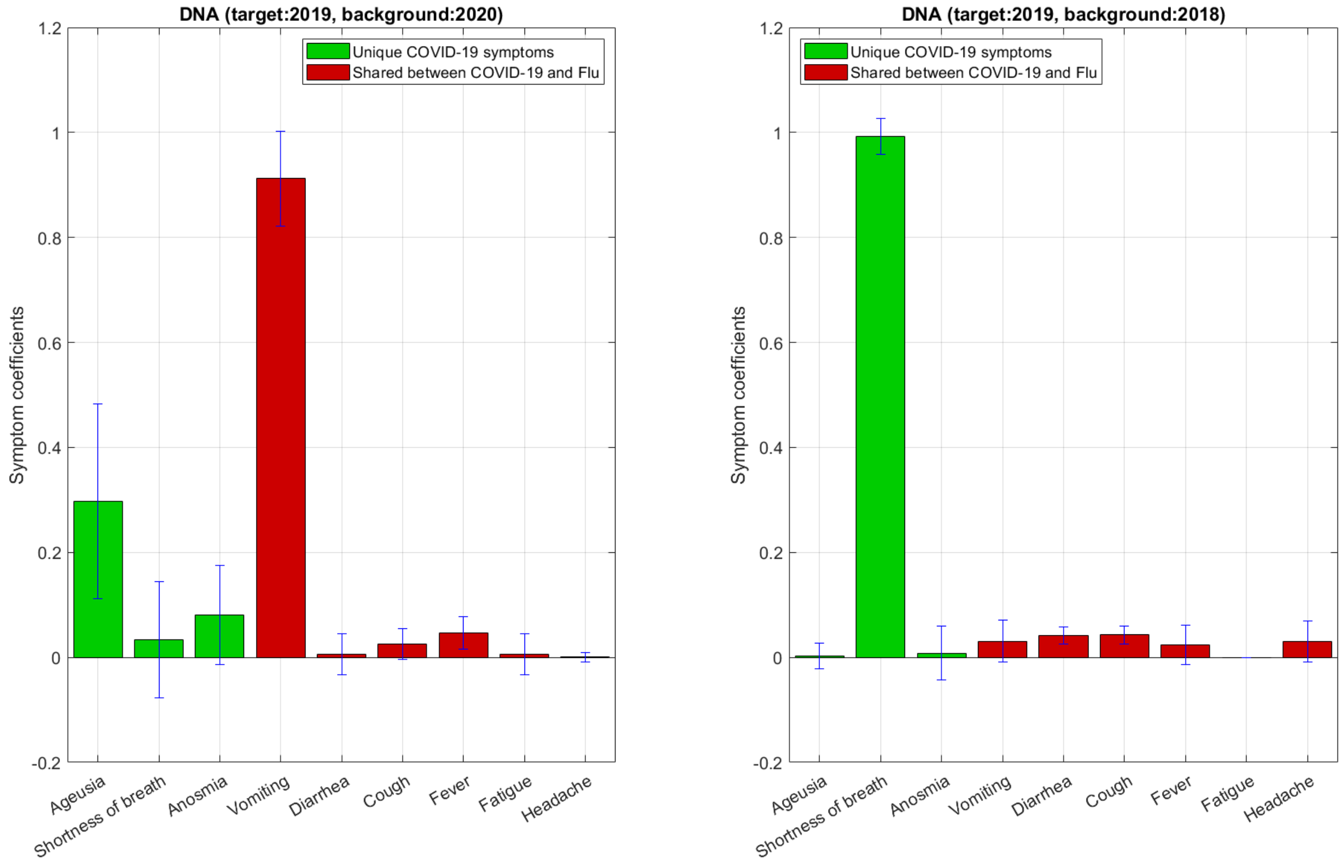


Figure 4: Symptom coefficients using DNA on 2019 data as target

In figure 5, we have considered 2018 searches as target data and 2019/2020 searches as background data. Like the previous case, it was expected to see that DNA doesn't return unique COVID-19 symptoms as 2018 is the year when we didn't even hear about COVID-19 and number of searches on unique symptoms was as usual during that period. Contribution of unique symptom coefficients is low here.

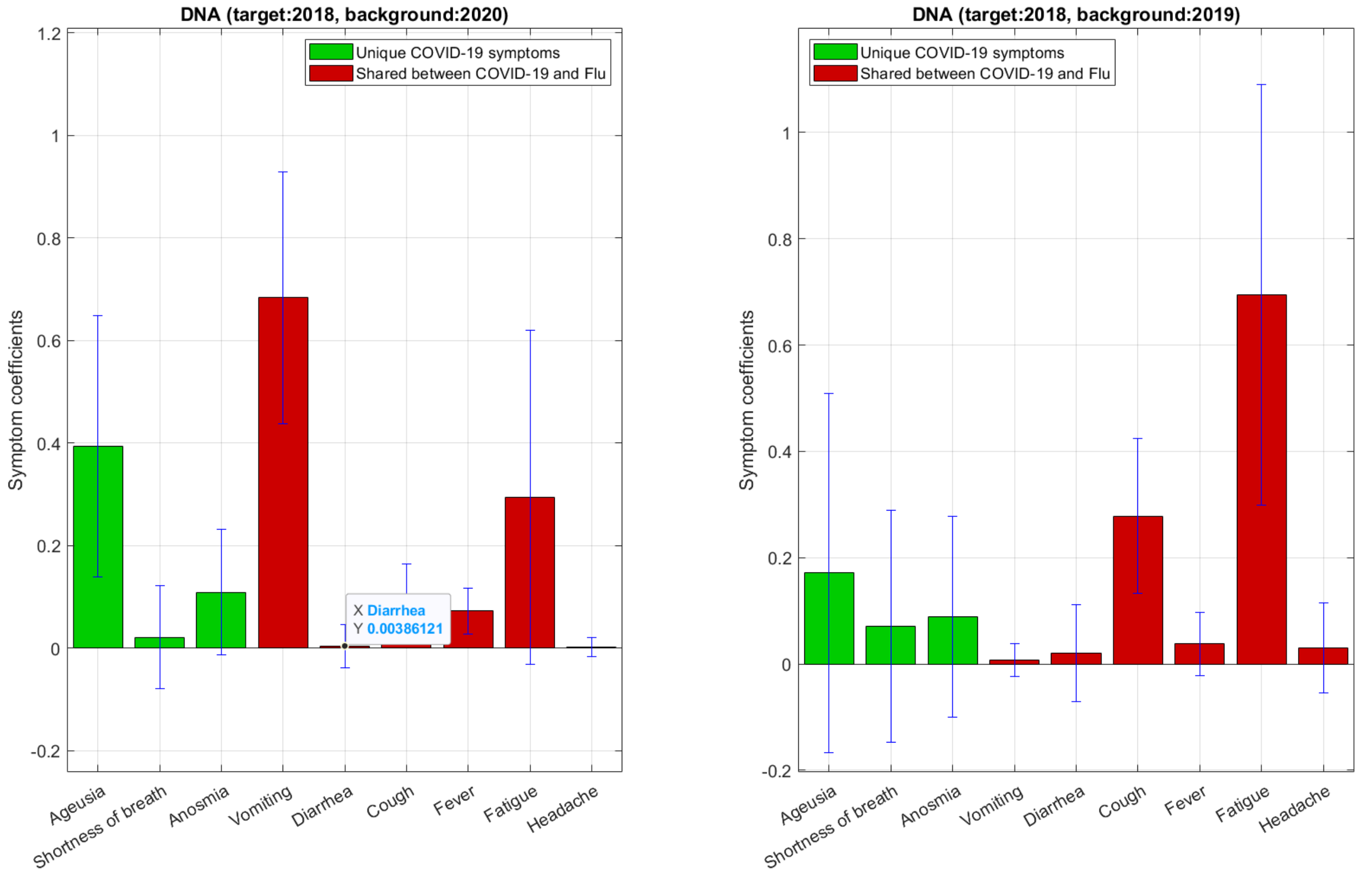


Figure 5: Symptom coefficients using DNA on 2018 data as target

4.2.2 Symptoms coefficients using NMF and PCA

As comparison, the alternative methods such as Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) are also tested using 2018, 2019, or 2020 dataset. In figure 6, we have shown symptoms coefficients using PCA and NMF on 2020 searches data. Unlike DNA, PCA and NMF cannot find uniqueness of COVID-19 symptoms. Both algorithms put more emphasis on symptoms those are shared between COVID-19 and Flu, and they are unable to extract discriminative information effectively. Using NMF and PCA, unique symptom coefficients have only 4.67% and 3.49%.

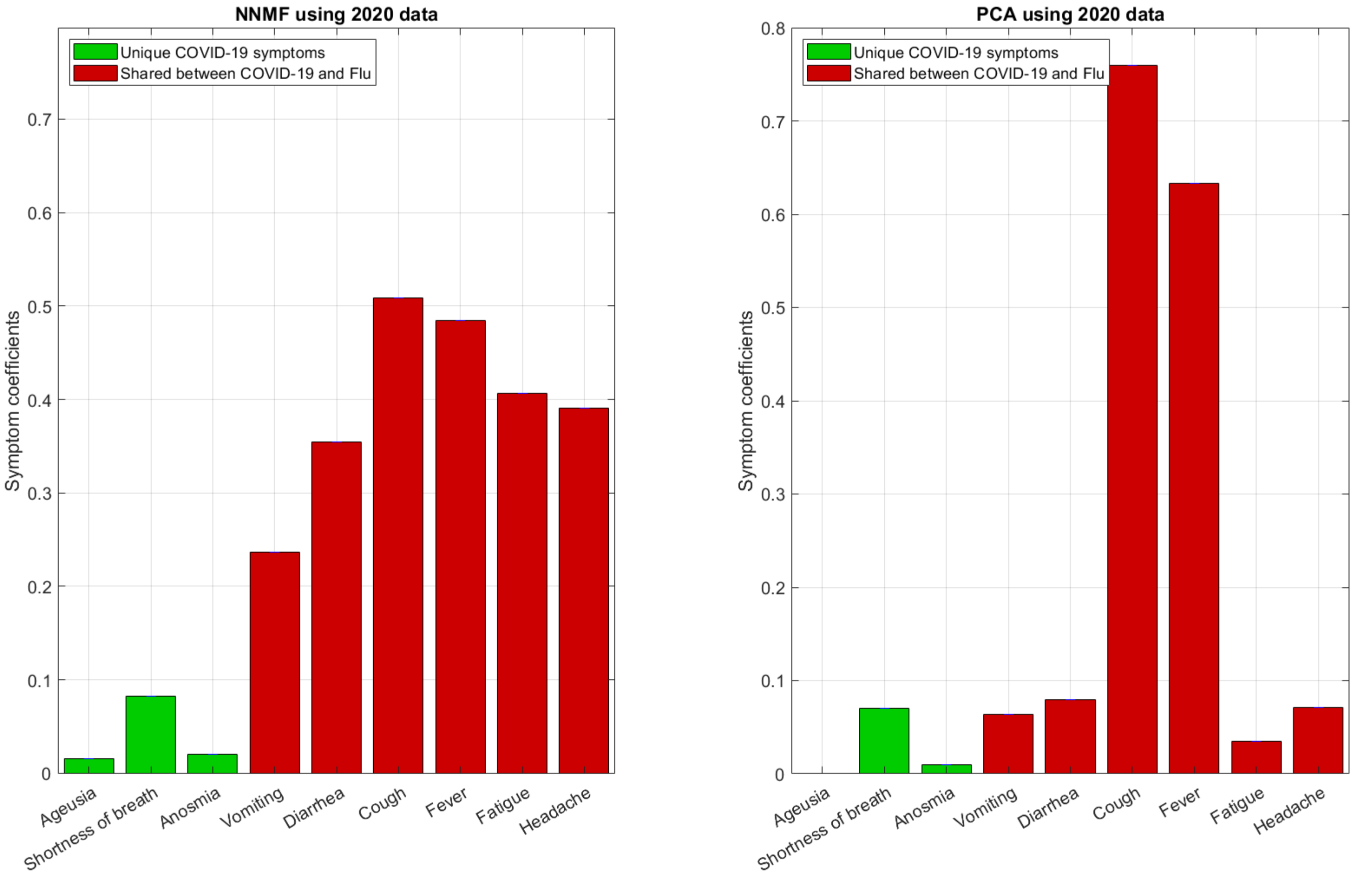


Figure 6: Symptom coefficients using NMF and PCA on 2020 data

In figure 7, we have shown symptom coefficients obtained using NMF and PCA on 2019 data. Similar to the analysis on 2020 data, NMF and PCA cannot find unique COVID-19 symptoms rather focus on shared symptoms on 2019 search data.

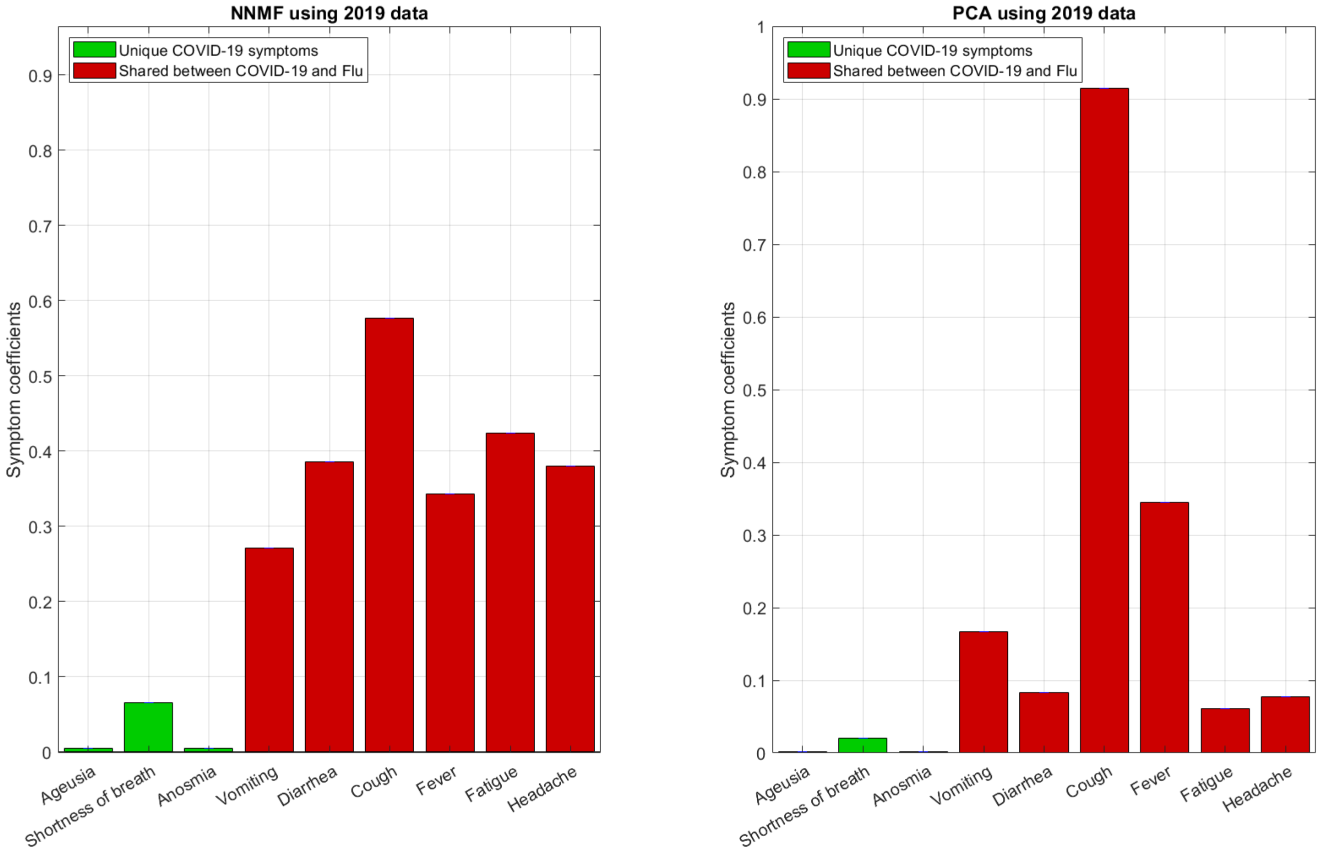


Figure 7: Symptom coefficients using NMF and PCA on 2019 data

In figure 8, we have shown results obtained from 2018 searches data using NMF and PCA. From the analysis on 2018, 2019 and 2020 data, we can see that NMF or PCA cannot detect COVID-19 breakout in 2020. For example, from the analysis using NMF, coefficients for Ageusia were 0.0160, 0.0044 and 0.0045 from 2018 to 2020 data respectively and coefficients for Cough were 0.484, 0.343 and 0.348 respectively. So, shared symptoms are dominant over COVID-19 symptoms for each year. Similar results obtained using PCA from 2018 to 2020 data. For both algorithms, we found that symptoms coefficients are larger in case of 2020 data, but they failed to detect unique COVID-19 symptoms against shared ones.

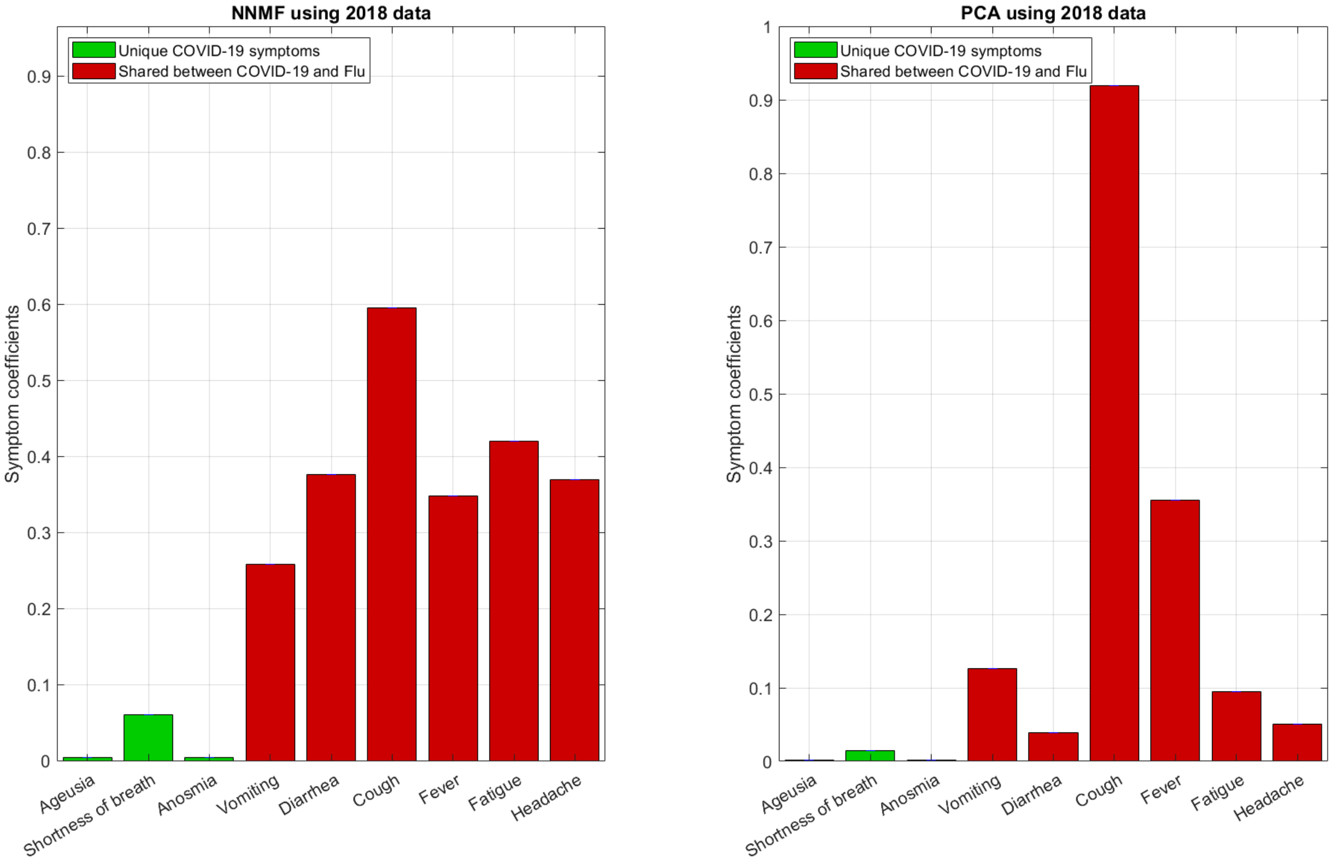


Figure 8: Symptom coefficients using NMF and PCA on 2018 data

4.2.3 Symptoms coefficients using cPCA

Figure 9 shows the results of contrastive PCA using multiple background-target data configurations with contrast parameter (alpha) equal to 0.5. cPCA, like NMF and PCA, was unable to place a strong emphasis on specific COVID-19 symptoms. When compared to Ageusia and Anosmia, the symptoms coefficients for Cough and Fever are relatively large in each case. 5.05% contribution in total symptom coefficient values has been observed for unique COVID-19 symptoms.

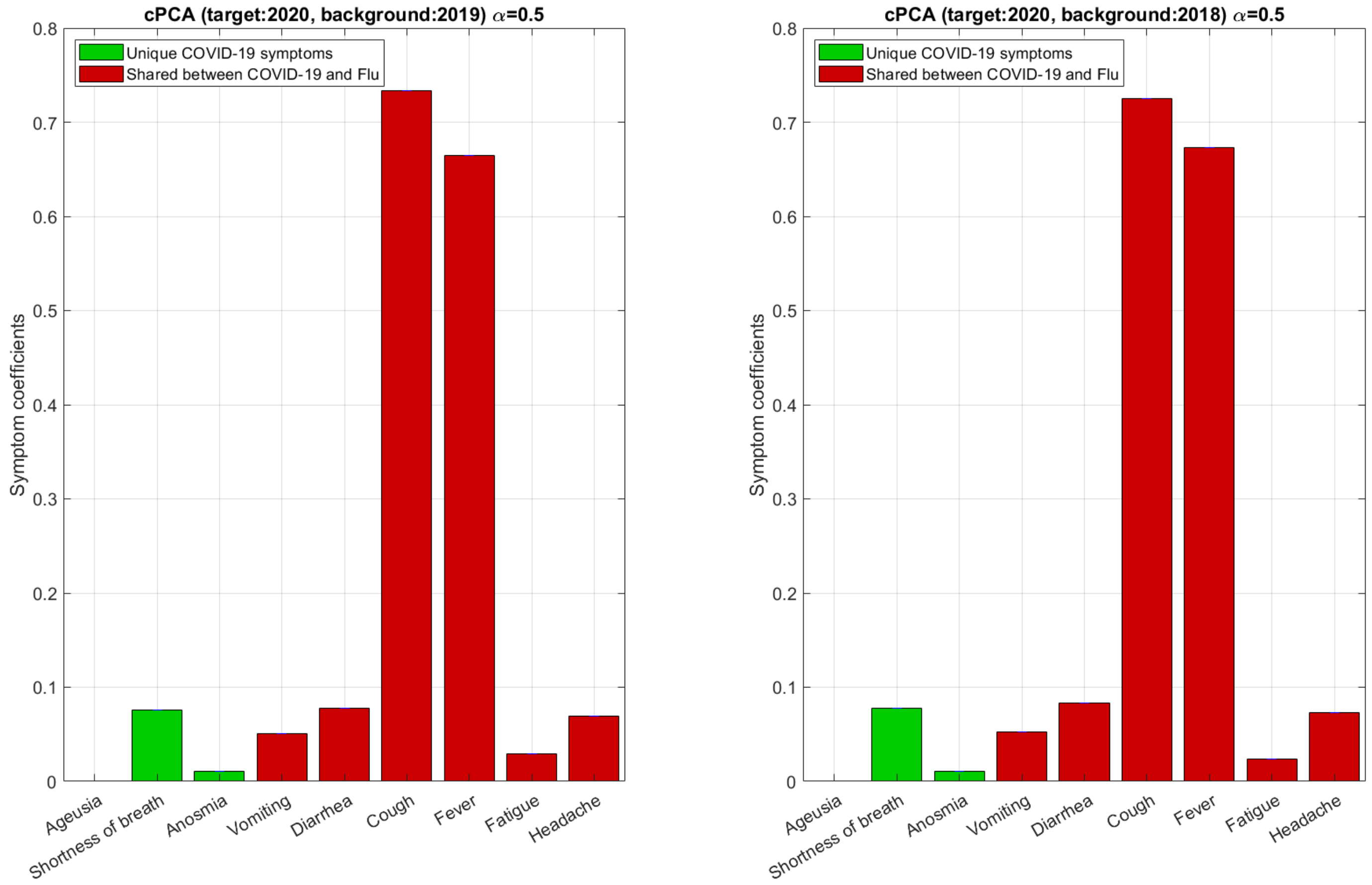
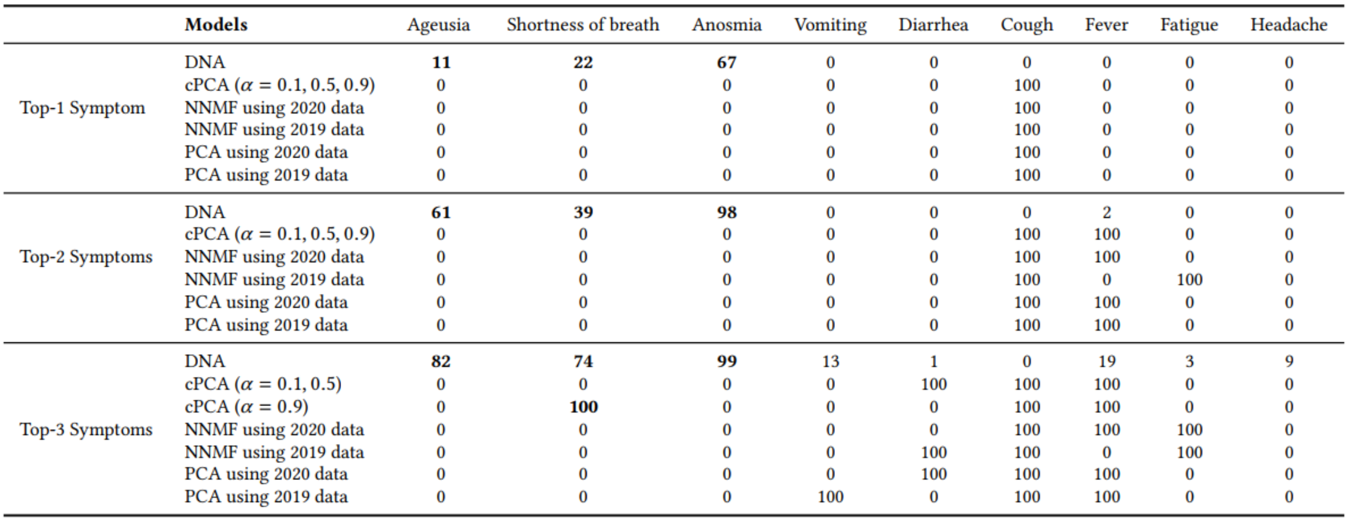


Figure 9: Symptom coefficients using cPCA on 2020 data as target with alpha = 0.5

Next, after running each method for 100 independent times while setting the background and target as the 2019 and 2020 Google Trends data, respectively, we investigate the frequencies of the symptoms showing up as the top-1, top-2, and top-3 by sorting the corresponding coefficients in a descending order. From the experiment results in Table 1, it can be further concluded that the proposed DNA outperforms the existing alternatives in terms of higher frequencies of successfully searching for the discriminative symptoms.

Table 2: Top-k symptom frequencies after 100 Monte Carlo experiments for different models



	Models	Ageusia	Shortness of breath	Anosmia	Vomiting	Diarrhea	Cough	Fever	Fatigue	Headache
Top-1 Symptom	DNA	**11**	**22**	**67**	0	0	0	0	0	0
	cPCA ($\alpha = 0.1, 0.5, 0.9$)	0	0	0	0	0	100	0	0	0
	NNMF using 2020 data	0	0	0	0	0	100	0	0	0
	NNMF using 2019 data	0	0	0	0	0	100	0	0	0
	PCA using 2020 data	0	0	0	0	0	100	0	0	0
	PCA using 2019 data	0	0	0	0	0	100	0	0	0
Top-2 Symptoms	DNA	**61**	**39**	**98**	0	0	0	2	0	0
	cPCA ($\alpha = 0.1, 0.5, 0.9$)	0	0	0	0	0	100	100	0	0
	NNMF using 2020 data	0	0	0	0	0	100	100	0	0
	NNMF using 2019 data	0	0	0	0	0	100	0	100	0
	PCA using 2020 data	0	0	0	0	0	100	100	0	0
	PCA using 2019 data	0	0	0	0	0	100	100	0	0
Top-3 Symptoms	DNA	**82**	**74**	**99**	13	1	0	19	3	9
	cPCA ($\alpha = 0.1, 0.5$)	0	0	0	0	100	100	100	0	0
	cPCA ($\alpha = 0.9$)	0	**100**	0	0	0	100	100	0	0
	NNMF using 2020 data	0	0	0	0	0	100	100	100	0
	NNMF using 2019 data	0	0	0	0	100	100	0	100	0
	PCA using 2020 data	0	0	0	0	100	100	100	0	0
	PCA using 2019 data	0	0	0	100	0	100	100	0	0

4.3 Feature Selection in Supervised Learning

Using discriminative properties of DNA, feature importance set can be created, which can lead us to use smaller set of features for classification. We performed feature selection procedure on MNIST and CIFAR10 datasets. DNA was implemented on training data only to get feature importance set and classification accuracy was calculated using testing data. We have shown the convergence time and classification accuracy with different number of features for both datasets in table 6.

Table 3: Classification Performance on MNIST and CIFAR10 using LR

Number of Features		Convergence time (s)		Classification accuracy (%)	
CIFAR10	MNIST	CIFAR10	MNIST	CIFAR10	MNIST
3072	784	29.65	18.99	85.10	99.91
2500	600	25.91	18.19	86.77	99.87
2000	500	23.10	17.45	85.35	99.65
1500	400	18.66	16.28	85.34	99.38
1000	300	14.32	15.09	86.75	97.33
500	200	11.45	14.63	87.90	99.54
100	100	7.86	12.91	87.00	97.72
25	25	5.98	7.26	83.64	95.37

4.3.1 Feature Selection in classification on CIFAR10 data

Figure 10 depicts the CIFAR10 dataset, which contains only two classes of data: airplane and automobile. For classification, we used logistic regression, and all features and examples in test data. Left image in figure 10 shows CIFAR10 dataset with two labels in two dimensions and right image shows predicted labels after classification using logistic regression.

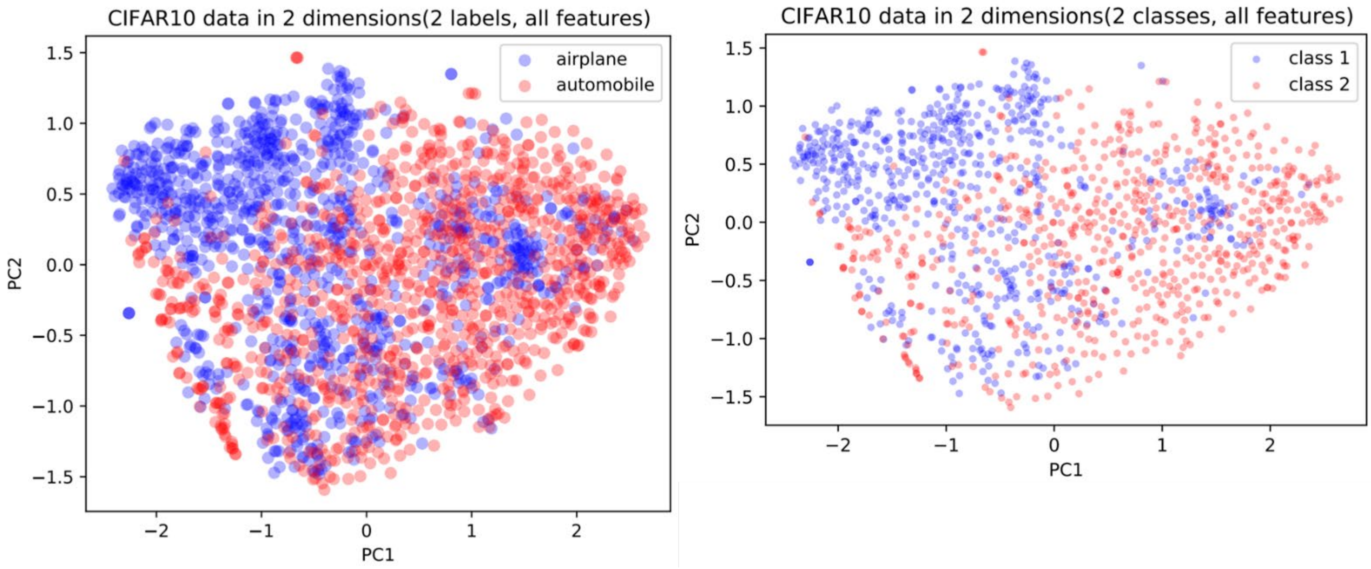


Figure 10: CIFAR10 dataset with 2 classes and all features

Next, instead of all features, we used only 25 features to perform classification with same example set. Examples have been shown in figure 11, with original label (left) and with predicted label (right). From table 6, when all 3072 features were considered, classification accuracy was 85.10% and when only 25 features used, accuracy was 83.64%. Convergence time was significantly less as expected because of reduced amount of data used. So, 79.83% convergence time improvement achieved with 1.72% loss of accuracy when using 25 most important features instead of 3072.

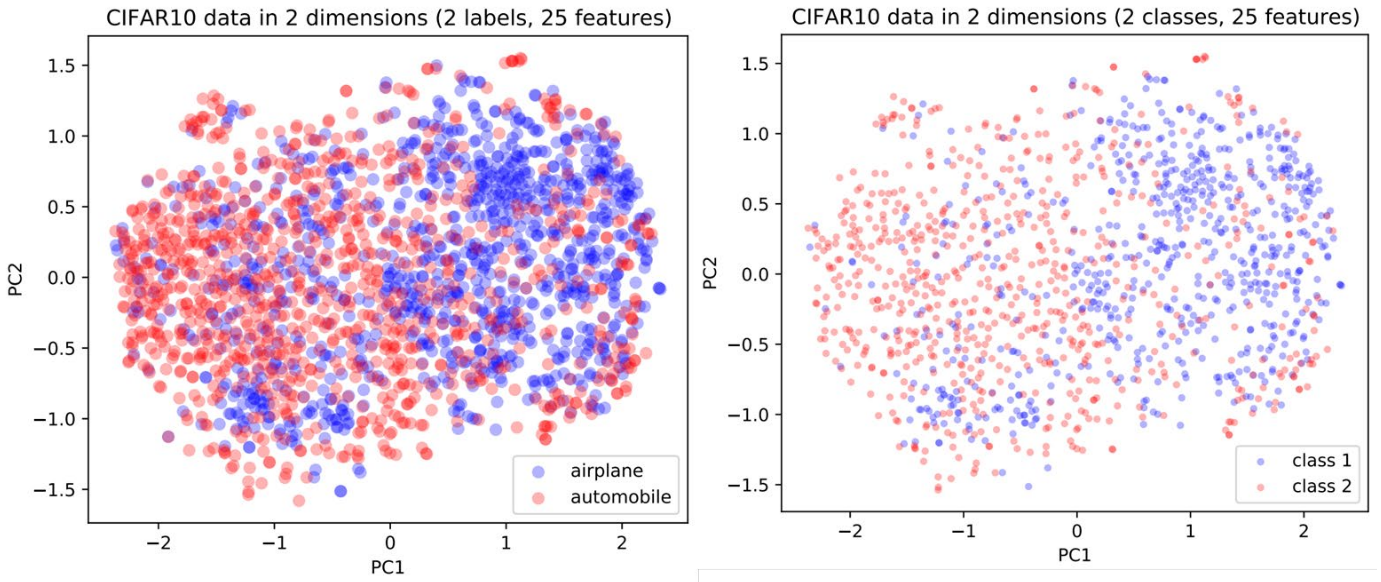


Figure 11: CIFAR10 dataset with 2 classes and 25 features

Figure 12 shows the performance curve for CIFAR10 dataset in classification accuracy and convergence time with different number of features used. We have considered eight different number of features and calculated corresponding classification accuracy and convergence time. Convergence time was decreased with a smaller number of features used every time. Highest classification accuracy, 87.90%, was achieved when top 500 features have been used.

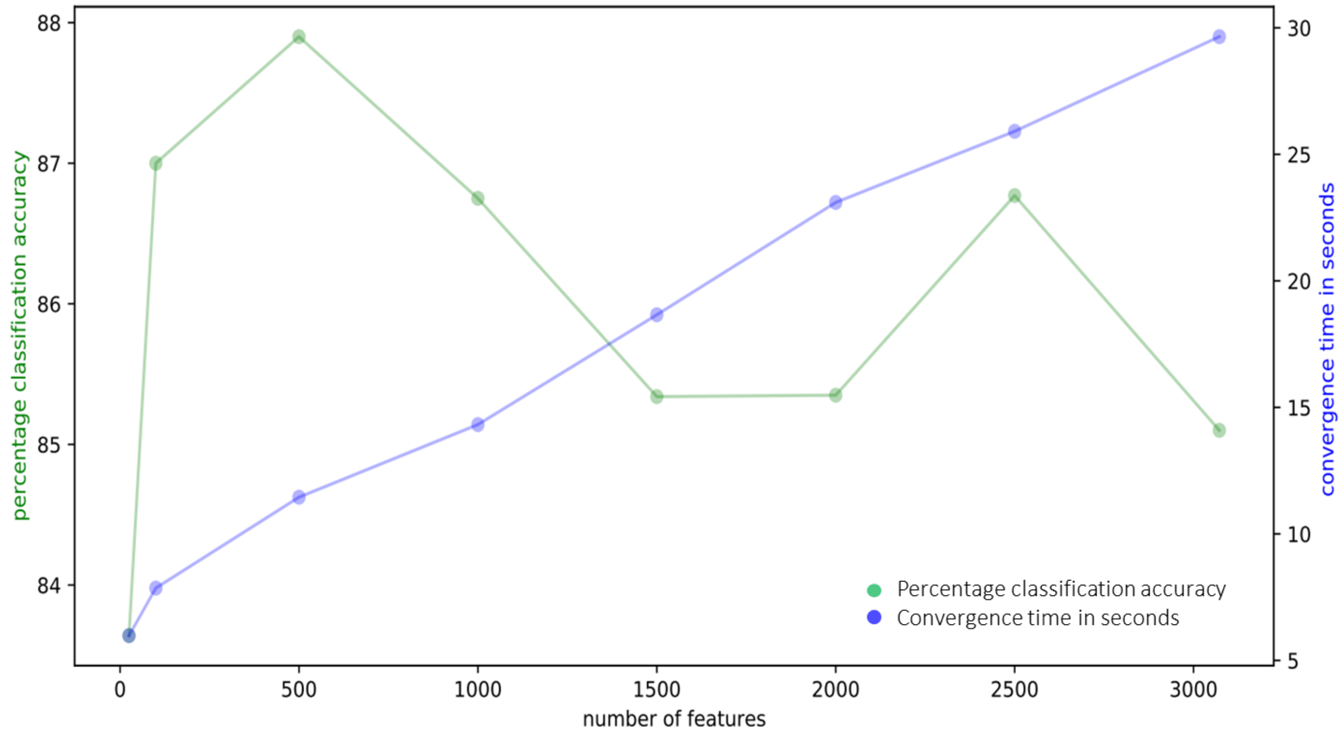


Figure 12: Classification Performance curve for CIFAR10

4.3.2 Feature Selection in classification on MNIST data

In the next section, we used the MNIST handwritten dataset for classification performance evaluation. Because MNIST is a simpler dataset, we were able to achieve higher classification accuracy. We have filtered the dataset for only two classes: handwritten digits for zero and one. Figure 13 shows MNIST test dataset with original labels (left) and predicted labels (right) in two dimensions.

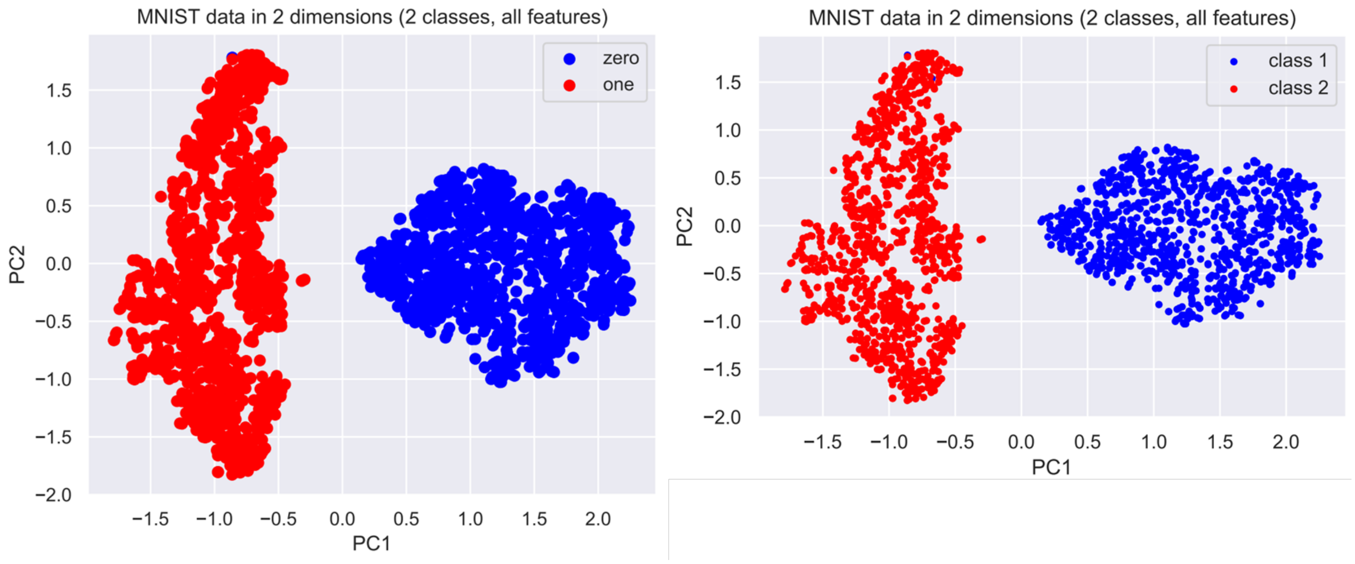


Figure 13: MNIST dataset with 2 classes and all features

Figure 14 shows all the examples of MNIST test dataset for top 25 features using original labels (left) and predicted labels (right). When 784 features considered, prediction accuracy was 99.91% and prediction accuracy was 95.37% when 25 most important features used. So, there is 4.54% decrease in performance with 61.77% improvement in convergence time.

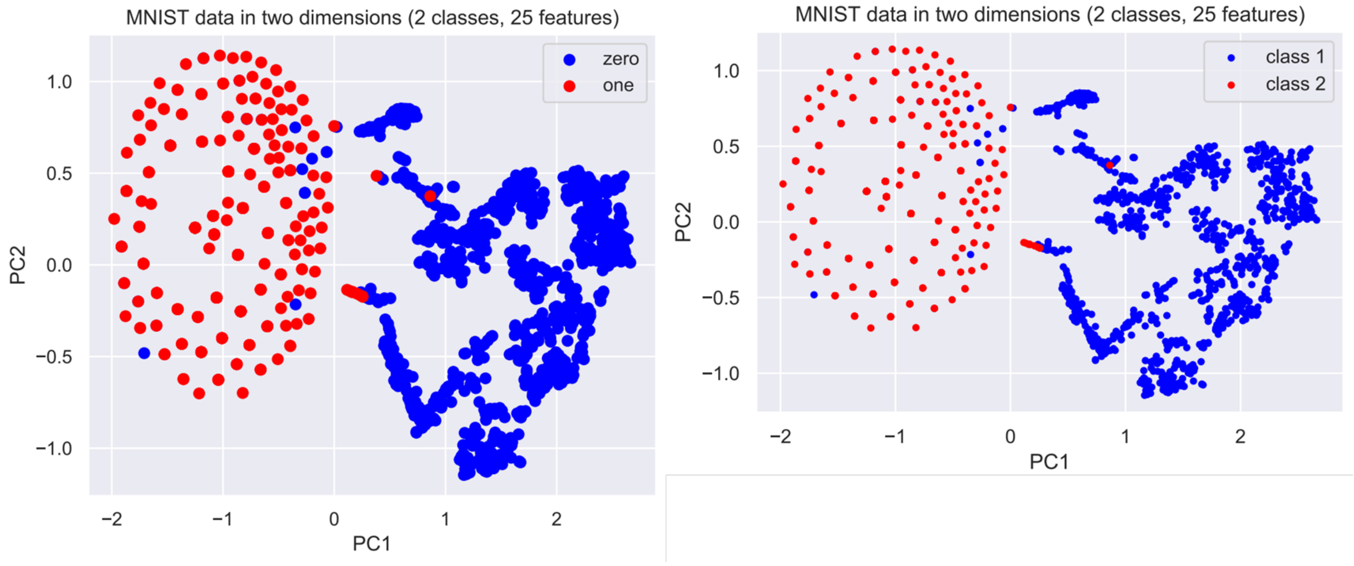


Figure 14: MNIST dataset with 2 classes and 25 features

Figure 15 we have shown the performance curve for MNIST dataset with classification accuracy and convergence time. With different number of features in between 784 and 25, we can see the classification accuracy has been decreased 4.54% at most with a consistent amount of improvement in convergence time. So, it can be concluded that, DNA is capable of sorting features in terms of their importance in the overall dataset for both MNIST and CIFAR10.

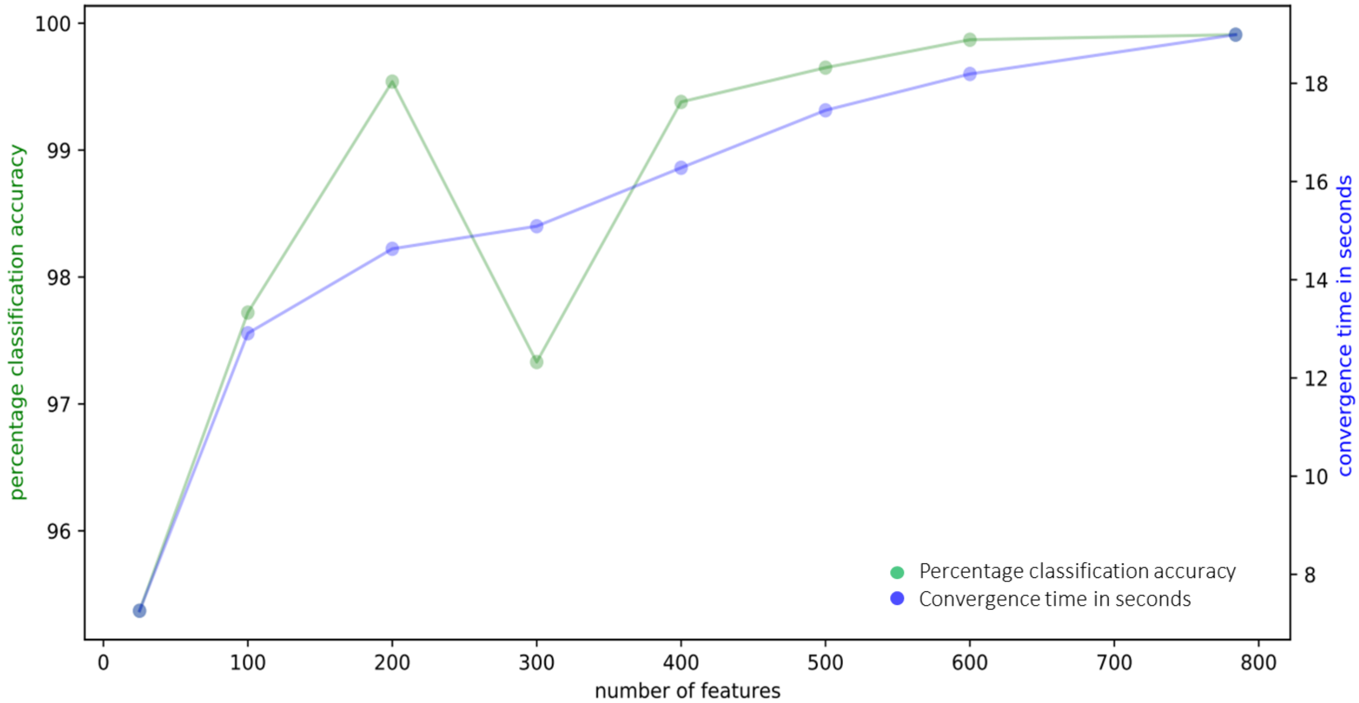


Figure 15: Classification Performance curve for MNIST dataset

4.4 Feature Selection in Unsupervised Learning

In this section, we have used feature importance set obtained from DNA to find clustering accuracy using K-means clustering algorithm. Convergence time and clustering accuracy for each number of features, have been shown in table 7 for both MNIST and CIFAR10 dataset.

Table 4: Clustering Performance on MNIST and CIFAR10 using K-Means

Number of Features		Convergence time (s)		Clustering accuracy (%)	
CIFAR10	MNIST	CIFAR10	MNIST	CIFAR10	MNIST
3072	784	28.61	10.66	69.30	99.85
2500	600	24.30	8.86	66.95	99.90
2000	500	21.10	8.40	67.55	99.86
1500	400	17.23	7.98	68.65	99.91
1000	300	13.46	7.27	69.70	97.91
500	200	10.17	6.52	69.40	99.86
100	100	6.42	6.13	67.05	99.29
25	25	5.67	5.88	67.35	92.00

4.4.1 Feature Selection in Clustering on CIFAR10 data

Figure 16 shows all the examples of CIFAR10 test dataset with all features in the left image and assigned clusters for each example has been shown in the right image. Like the analysis with supervised learning, we have used only examples with two labels: airplane and automobile for clustering performance analysis.

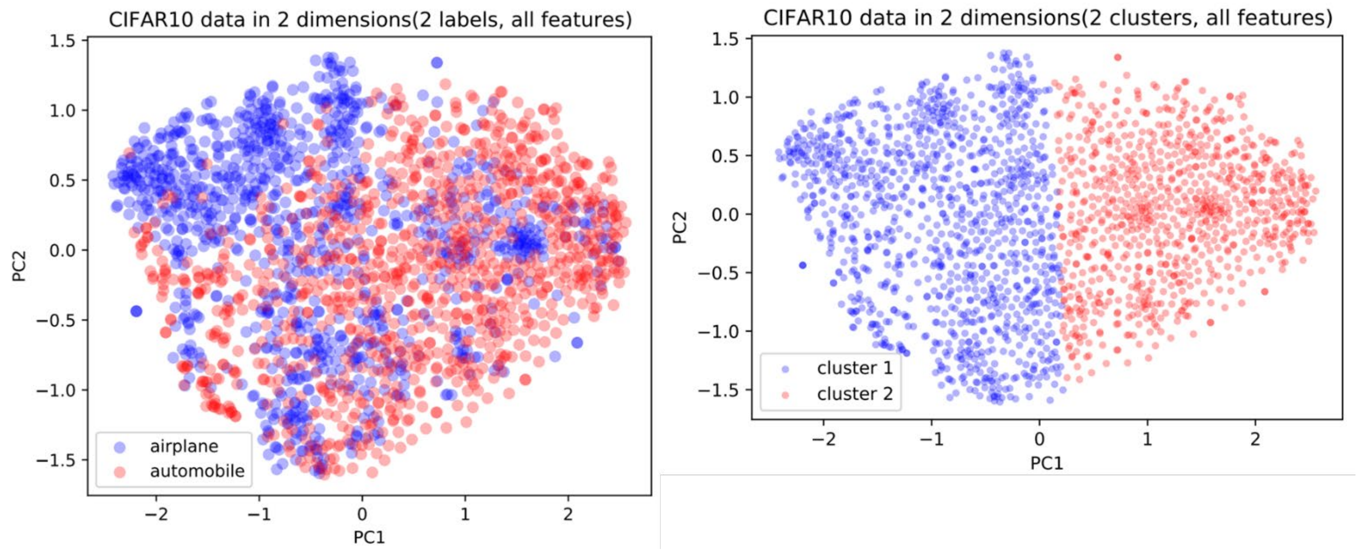


Figure 16: CIFAR10 dataset with 2 clusters and all features

In figure 17, we have shown all examples of CIFAR10 test dataset with 25 features in two dimensions and examples with predicted clusters side by side. When only 25 features used, clustering accuracy was 67.35% which is 2.81% less than clustering accuracy obtained using all features.

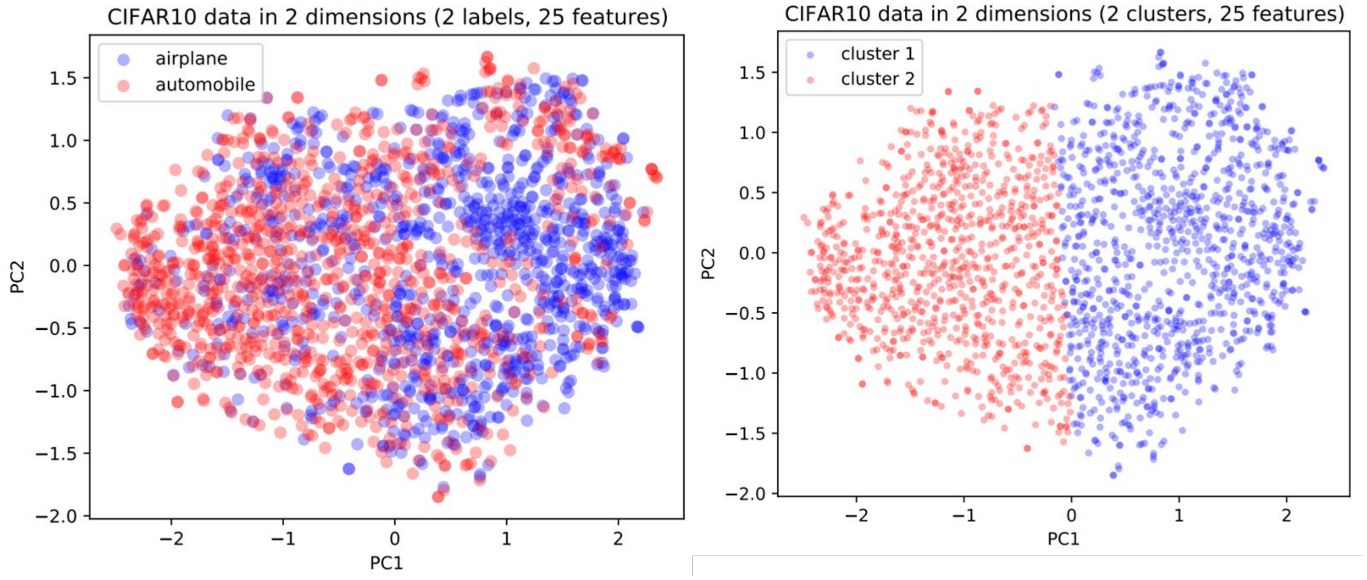


Figure 17: CIFAR10 dataset with 2 clustering and 25 features

Clustering performance has been shown in figure 18 with percentage clustering performance and convergence time with respect to different number of features. Highest performance 69.70% obtained when using 1000 top features. Convergence time has been improved by 81.18% with a consistent manner from 3072 features to 25 features.

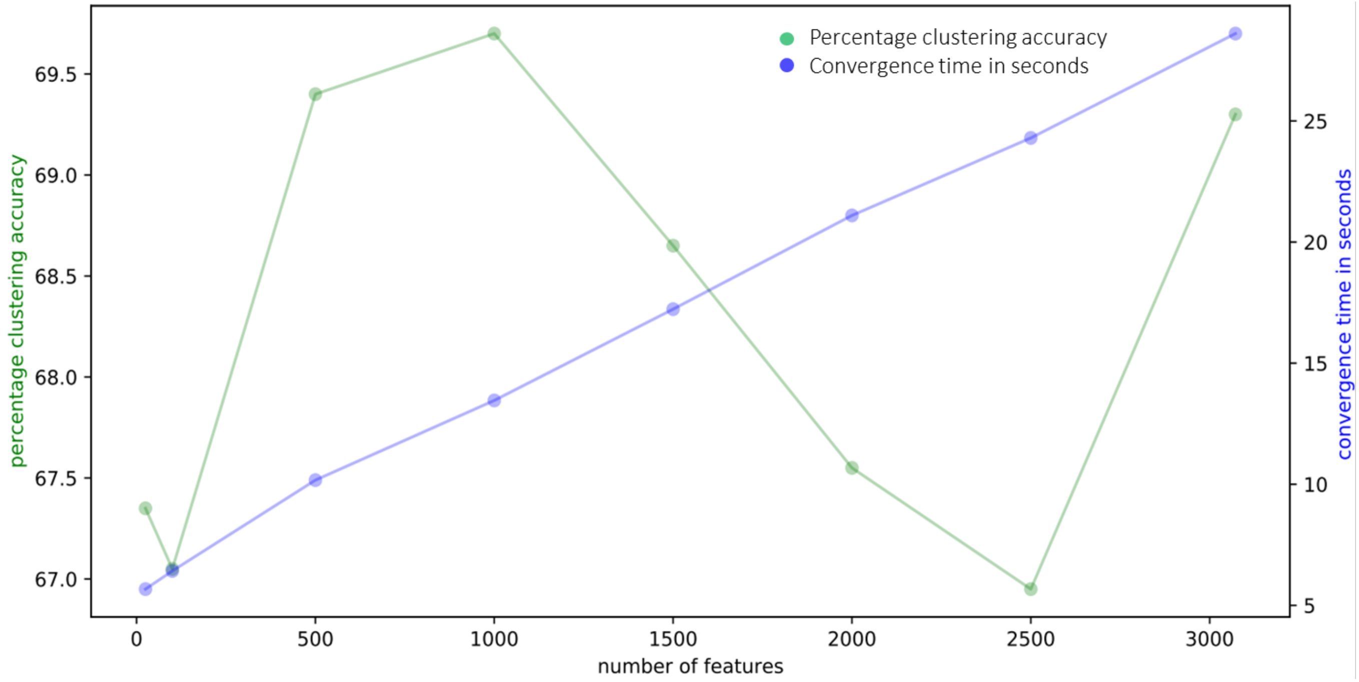


Figure 18: Clustering Performance curve for CIFAR10

4.4.2 Feature Selection in Clustering on MNIST data

Here we used MNIST test data for our clustering performance analysis. In figure 19 we have shown each example with all features in two dimensions (left image) and after clustering using K-Means both clusters shown in right image. Clustering accuracy, obtained using all features, was 99.85%.

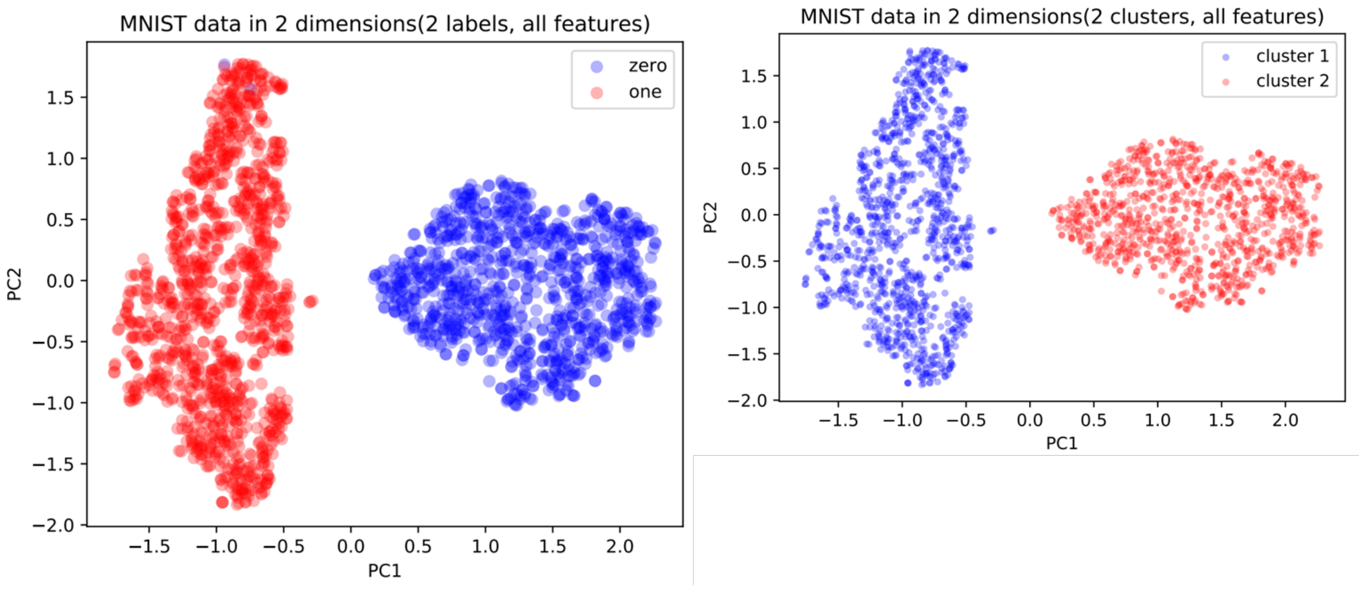


Figure 19: MNIST dataset with 2 classes and all features

In figure 20, we have exhibited MNIST data with 50 features (left) and corresponding cluster prediction in right image. When using 50 features, classification accuracy was 98.89% which is 1.02% less than the accuracy using all 784 features with 44.84% improvement in convergence time of K-Means algorithm.

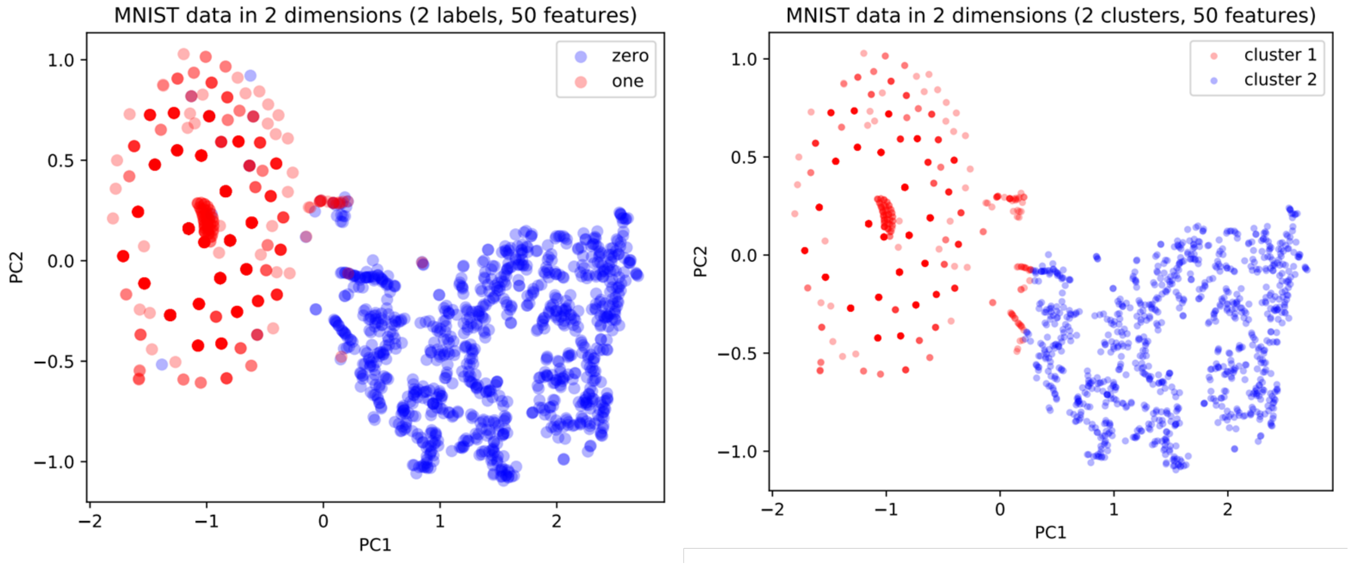


Figure 20: MNIST dataset with 2 classes and 50 features

In figure 21, we have manifested clustering performance curve for MNIST dataset. While determining clustering accuracy, we observed at most 0.56% variation using different number of top features from 784 to 100. We found a sudden decrease of performance when using number of features less than 50.

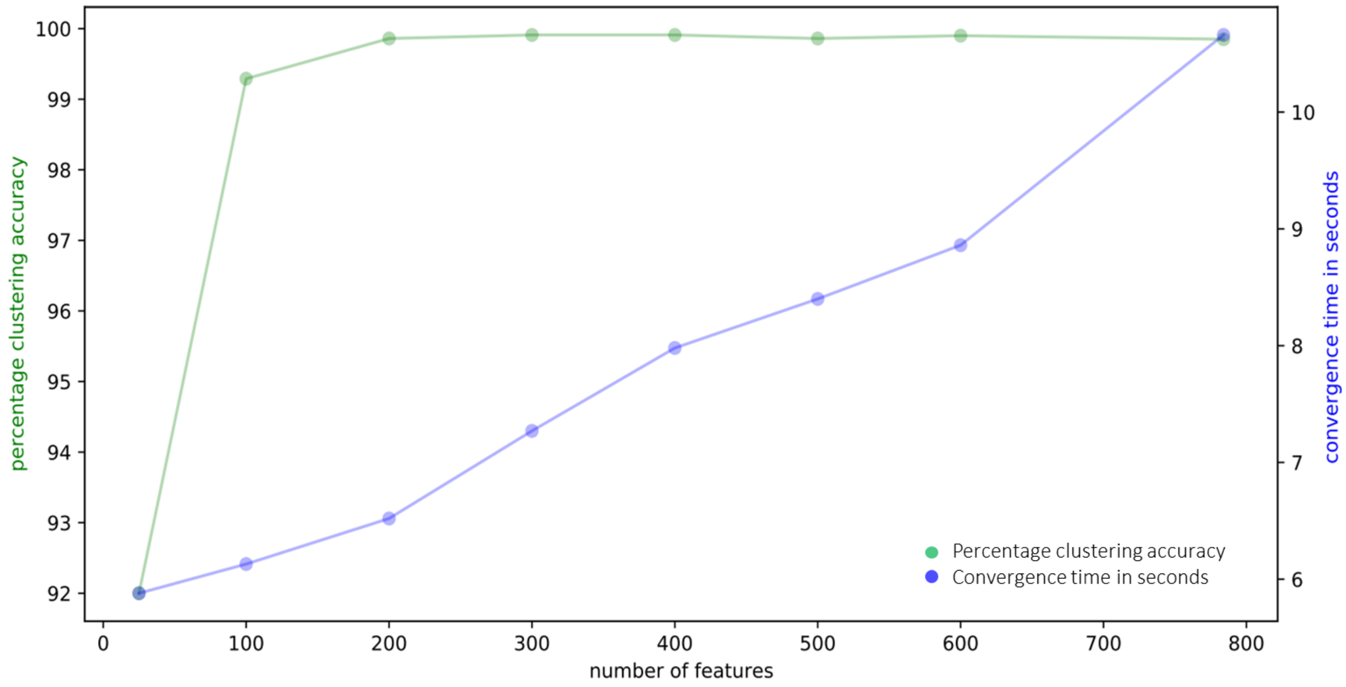


Figure 21: Clustering performance curve for MNIST

4.5 Comparison of Feature Selection Performance with cPCA

In this section, we compared feature selection performance of DNA with that obtained using cPCA. We used the feature importance set derived from both techniques to estimate classification performance with varied numbers of top features in the MNIST test dataset and showed the comparison in figure 22. We can see that, among seven test cases, in four test cases, feature set from DNA provides better classification accuracy than those obtained from cPCA. Four test cases when the number of top features used were: 600, 500, 200 and 25. We excluded test case for 784 features here, because performance is same when all features used with different feature ranking obtained from each algorithm. When using only 25 top features from DNA, classification accuracy was 0.08% higher.

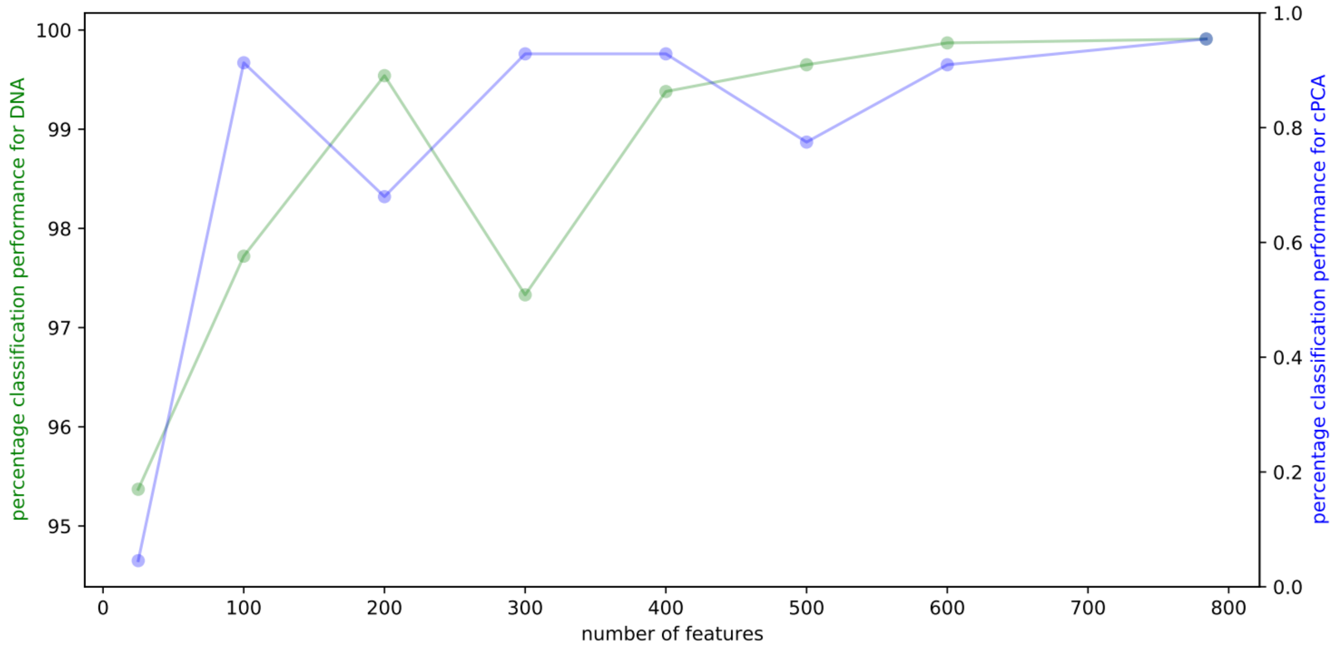


Figure 22: Feature selection performance comparison in classification (DNA vs cPCA)

In figure 23, we have manifested comparison of clustering performance using feature importance set from DNA and cPCA. Similar to the comparison in supervised learning setting, we got better performance using DNA algorithm. Using feature ranking set from both algorithms, among seven test cases, in four cases clustering performance was better when used feature importance set obtained from DNA. Four cases when 600, 500, 400 and 200 top features used. When using 200 top features from DNA, the model provides 1.01% better clustering performance.

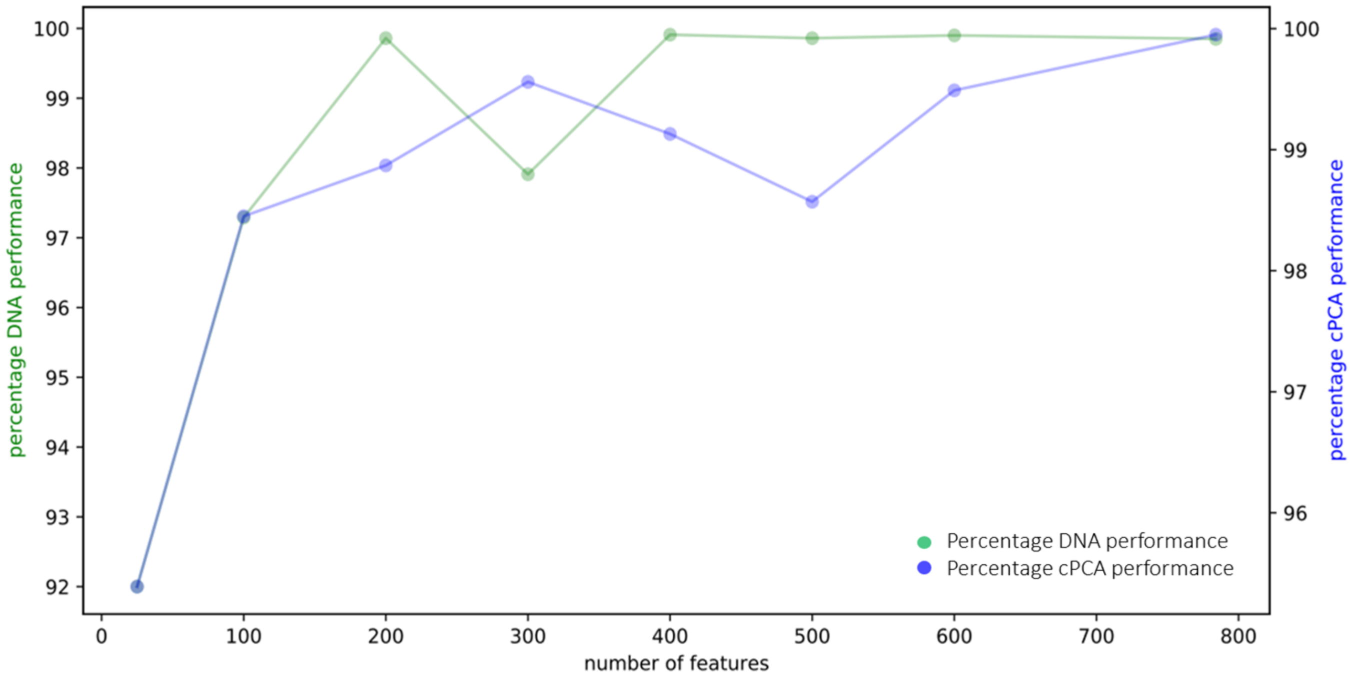


Figure 23: Feature selection performance comparison in clustering (DNA vs cPCA)

We executed all operations one at a time in a single window to establish an unbiased environment for calculating convergence time in supervised and unsupervised learning settings. We used a Dell Precision 5500 laptop with an Intel Core i7 - 10750H processor and 16 GB of RAM. We calculated the difference time by using the python time method from the time class to measure the start and finish times. Convergence time may vary depending on the machine configuration, but the percentage variation in convergence time should be comparable.

CHAPTER V

CONCLUSION AND FUTURE WORKS

Leveraging the advances of the discriminative principal component and the nonnegative matrix decomposition, this paper puts forward a new multi-view learning model, this is terms DNA, to extract the discriminative information of one dataset relative to the other dataset. First, we demonstrated how to extract unique information from one dataset relative to another using DNA. We used Google Trends disease symptoms dataset and took a subset which contains only information about COVID-19 and Flu symptoms. We have demonstrated, using DNA, that COVID-19 symptoms were prevalent at the time of year 2020 with respect to year 2018 and 2019, compared to shared symptoms between COVID-19 and Flu. It can be very useful to early detect possible outbreaks in future. Additionally, we have proven that DNA outperforms other competitive algorithms: cPCA, PCA and NMF, in finding unique discriminative information about COVID-19. Second, we developed a feature selection method based on DNA and demonstrated its performance using supervised and unsupervised learning techniques. We picked a varied number of top features from feature ranking obtained from DNA and calculated classification and clustering accuracy in some standard datasets. We found an insignificant fall in performance with a gradual decrease in convergence time when used smaller feature set because of reduced data size. Other conventional discriminative analysis techniques such as cPCA have also been addressed and results obtained. We compared performance obtained using cPCA with our proposed method and found better performance from our approach.

In the future, our research opens in several directions: (1) develop non-negative dPCA and compare its performance against DNA; (2) understand the connection between eigenvalue decomposition and nonnegative matrix factorization on $C_y^{-1}C_x$, (3) modify DNA algorithm to extract discriminative information from more than two datasets. (4) create new features to obtain better performance in supervised and unsupervised learning, (5) work with more complex datasets like images of cells, tissues from human body (6) work with human sentiment data to obtain discriminative information etcetera.

REFERENCES

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, “Tensor decompositions for learning latent variable models,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 2773–2832, Jan. 2014.

A. Cichocki, R. Zdunek, A. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. John Wiley & Sons, 2009

A. K. Jain and R. C. Dubes. Algorithm for Clustering Data, chapter Clustering Methods and Algorithms. Prentice-Hall Advanced Reference Series, 1988.

Abid, A., Zhang, M.J., Bagaria, V.K. and Zou, J., 2017. Contrastive principal component analysis. arXiv preprint arXiv:1709.06716.

Abid, A., Zhang, M.J., Bagaria, V.K. et al. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nat Commun 9, 2134 (2018).

Akaho, S. A kernel method for canonical correlation analysis. In Proc. Int’l Meeting on Psychometric Society, 2001.

Amaryllis Mavragani and Konstantinos Gkillas. 2020. COVID-19 predictability in the United States using Google Trends time series. Scientific reports 10, 1 (2020).

Amaryllis Mavragani. 2020. Tracking COVID-19 in Europe: infodemiology approach. JMIR public health and surveillance 6, 2 (2020), e18941.

Andrew, Galen, et al. "Deep canonical correlation analysis." International conference on machine learning. PMLR, 2013.

Atina Husnayain, Anis Fuad, and Emily Chia-Yu Su. 2020. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. International Journal of Infectious Diseases 95 (2020), 221–223.

Benton, Adrian, et al. "Deep generalized canonical correlation analysis." arXiv preprint arXiv:1702.02519 (2017).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3. Jan (2003): 993-1022.

Borga, M. (1999). Canonical correlation. Online tutorial. Available online at: http://www.imt.liu.se/~magnus/cca/tutorial

C. Spearman. General intelligence objectively determined and measured. American Journal of Psychology, 15:206–221, 1904.

Cayton, Lawrence. "Algorithms for manifold learning." Univ. of California at San Diego Tech. Rep 12.1-17 (2005).

Chen X, Liu H, Carbonell J (2012) Structured sparse canonical correlation analysis. Proceedings of the 15th international conference on artificial intelligence and statistics, pp 199 207

Chen, W., Ma, H., Yu, D. & Zhang, H. SVD-based technique for interference cancellation and noise reduction in NMR measurement of time-dependent magnetic fields. Sensors 16, 323 (2016).

Ciucci, S., Ge, Y., Durán, C. et al. Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies. Sci Rep 7, 43946 (2017).

Cuilian Li, Li Jia Chen, Xueyu Chen, Mingzhi Zhang, Chi Pui Pang, and Haoyu Chen. 2020. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance 25, 10 (2020).

Cunningham P., Cord M., Delany S.J. (2008) Supervised Learning. In: Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg.

Dash, Manoranjan, and Huan Liu. "Feature selection for classification." Intelligent data analysis 1.1-4 (1997): 131-156.

E. Acar, S. A. Camtepe, M. S. Krishnamoorthy, and B. Yener, Modeling and Multiway Analysis of Chatroom Tensors. Berlin, Germany: SpringerVerlag, 2005, pp. 256–268.

Farquhar, J., Hardoon, D., Meng, H., Shawe-Taylor, J.S., Szedmak, S.: Two view learning: Svm-2k, theory and practice. In: Advances in neural information processing systems. pp. 355–362 (2006).

Garte, S.: The role of ethnicity in cancer susceptibility gene polymorphisms: The example of CYP1A1. Carcinogenesis 19(8), 1329–1332 (Aug 1998).

Girish Chandrashekar, Ferat Sahin, A survey on feature selection methods, Computers & Electrical Engineering, Volume 40, Issue 1, 2014, Pages 16-28, ISSN 0045-7906,

Gönen, Mehmet, and Ethem Alpaydın. "Multiple kernel learning algorithms." The Journal of Machine Learning Research 12 (2011).

Google LLC. 2021. Google COVID-19 Search Trends symptoms dataset. http: //goo.gle/covid19symptomdataset. (Feb. 2021).

H. Zhao, Z. Wang and F. Nie, "A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 4, pp. 629-640, 1 April 2019, doi: 10.1109/TKDE.2018.2842023.

Hardoon D, Shawe-Taylor J (2011) Sparse canonical correlation analysis. Mach Learn 83:331 353

Hariharan B., Malik J., Ramanan D. (2012) Discriminative Decorrelation for Clustering and Classification. In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7575. Springer, Berlin, Heidelberg.

Harshman, R.A.: PARAFAC2: Mathematical and technical notes. UCLA Work. Pap. phonetics, 22, pp. 30–44 (1972).

Hotelling H. 1933 Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441, 498–520.

Hotelling, Harold (December 1936). "Relation between two sets of variates". Biometrika. 28 (3–4): 321–377.

I. Buciu, "Non-Negative Matrix Factorization, a New Tool for Feature Extraction: Theory and Applications," Int'l J. Computers, Comm. and Control, vol. 3, pp. 67-74, 2008.

J.C. Ang, A. Mirzal, H. Haron, H.N.A. Hamed Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection IEEE/ACM Trans Comput Biol Bioinf (2016).

Jia Chen, "Graph Multiview Canonical Correlation Analysis", IEEE Transactions on Signal Processing, vol. 67, no. 11, June 1, 2019.

John, G.H., Kohavi, R. and Pfleger, K., Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, 121-129, 1994.

Jolliffe, Ian. Principal Component Analysis. Wiley Online Library, 2002.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (November 2018), 45 pages.

K. Pearson. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2:559–572, 1901.

Kira, K. and Rendell, L.A., The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of Ninth National Conference on Artificial Intelligence, 129-134, 1992.

Kohavi, R. and Sommetlield, D., Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In: Proceedings of First International Conference on Knowledge Discovery and Data Mining, Morgan Kaufmann, 192-197, 1995.

Koller, D. and Sahami, M., Toward optimal feature selection. In: Proceedings of International Conference on Machine Learning, 1996.

Köppen, Mario. "The curse of dimensionality." 5th online world conference on soft computing in industrial applications (WSC5). Vol. 1. 2000.

Kumar, Vipin, and Sonajharia Minz. "Feature selection: a literature review." SmartCR 4.3 (2014).

L.O. Jimenez and D.A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man and Cybernetics, 28(1):39–54, 1997.

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401:899–91.

Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.

Liu, Xiaoke; (2020) Discriminative principal component analysis for high dimensional classification with applications in NIR spectroscopy. Doctoral thesis (Ph. D), UCL (University College London).

Luo, Fulin, et al. "Semisupervised sparse manifold discriminative analysis for feature extraction of hyperspectral images." IEEE Transactions on Geoscience and Remote Sensing 54.10 (2016): 6197-6211.

M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization," Computational Statistics and Data Analysis, vol. 52, no. 1, pp. 155-173, 2007.

Maria Effenberger, Andreas Kronbichler, Jae Il Shin, Gert Mayer, Herbert Tilg, and Paul Perco. 2020. Association of the COVID-19 pandemic with internet search volumes: a Google Trends^{TM} analysis. International Journal of Infectious Diseases 95 (2020), 192–197.

Md Imrul Kaish, Md Jakir Hossain, Evangelos E. Papalexakis, and Jia Chen. 2021. COVID-19 or Flu? Discriminative Knowledge Discovery of COVID-19 Symptoms from Google Trends Data. In epiDAMIK 2021: 4th epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery. ACM, New York, NY, USA

Melzer, T., Reiter, M., and Bischof, H. Nonlinear feature extraction using generalized canonical correlation analysis in ICANN, 2001.

N. Sidiropoulos, R. Bro, and G. Giannakis, Parallel factor analysis in sensor array processing, IEEE Trans. Signal Process., 48 (2000), pp. 2377–2388.

P. Paatero and U. Tapper, "Positive Matrix Factorization: Anonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values," Environmetrics, vol. 5, no. 2, pp. 111-126, 1994.

P. Paatero, "Least Squares Formulation of Robust nonnegative Factor Analysis," Chemometrics and Intelligent Laboratory Systems, vol. 37, no. 1, pp. 23-35, 1997

P. S. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Proceedings of the 4th International Conference on Knowledge Discovery & Data Mining (KDD'98), pages 9–15, 1998.

Perros, Ioakeim, et al. "SPARTan: Scalable PARAFAC2 for large & sparse data." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.

Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.

R Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of ACM SIGMOD Conference on Management of Data, 1998.

R. Bro, “PARAFAC. Tutorial and applications,” Chemometrics and Intelligent Laboratory Systems, vol. 38, pp. 149–171, 1997.

R. O. Duda and P. E. Hart. Pattern Classification and Scene Analysis, chapter Unsupervised Learning and Clustering. John Wiley & Sons, 1973.

R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945-949.

Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 161–168.

Riedmiller, Martin. "Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms." Computer Standards & Interfaces 16.3 (1994): 265-278.

S. A. Raurale, J. McAllister and J. M. del Rincon, "Real-Time Embedded EMG Signal Analysis for Wrist-Hand Pose Identification," in IEEE Transactions on Signal Processing, vol. 68, pp. 2713-2723, 2020. S. Zhang, X. Zhao, B. Lei Speech emotion recognition using an enhanced kernel isomap for human-robot interaction Int J Adv Robot Syst (2013)

Sadoughi, Farahnaz et al. “Application of Canonical Correlation Analysis for Detecting Risk Factors Leading to Recurrence of Breast Cancer.” Iranian Red Crescent medical journal vol. 18,3 e23131. 1 Mar. 2016.

Sargin, Mehmet Emre, et al. "Audiovisual synchronization and fusion using canonical correlation analysis." IEEE transactions on Multimedia 9.7 (2007).

Seyed Mohammad Ayyoubzadeh, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, Mahnaz Ahmadi, and Sharareh R Niakan Kalhori. 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR public health and surveillance 6, 2 (2020).

Shaham, Uri, et al. "Spectralnet: Spectral clustering using deep neural networks." arXiv preprint arXiv:1801.01587 (2018).

Suthaharan, Shan. "Support vector machine." Machine learning models and algorithms for big data classification. Springer, Boston, MA, 2016. 207-235.

Suvrit Sra; Inderjit S. Dhillon (2006). Generalized Nonnegative Matrix Approximations with Bregman Divergences. Advances in Neural Information Processing Systems 18. Advances in Neural Information Processing Systems. ISBN 978-0-262-23253-1. Wikidata Q77685465.

T. G. Kolda, B. W. Bader, and J. P. Kenny, “Higher-order web link analysis using multilinear algebra,” in Proc. IEEE Int. Conf. Data Min., Houston, TX, Nov. 2005.

T. Huang et al. (Eds.): ICONIP 2012, Part III, LNCS 7665, pp. 382–389, 2012.c Springer-Verlag Berlin Heidelberg 2012

Ting-Kai Sun, Song-Can Chen, Zhong Jin and Jing-Yu Yang, "Kernelized discriminative canonical correlation analysis," 2007 International Conference on Wavelet Analysis and Pattern Recognition, 2007, pp. 1283-1287, doi: 10.1109/ICWAPR.2007.4421632.

V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS-clustering categorical data using summaries. In Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD’99), 1999.

W.S. Torgerson. Multidimensional scaling, I: Theory and method. Psychometrika, 17:401–419, 1952.

White M, Yu Y, Zhang X, Schuurmans D (2012) Convex multiview subspace learning. Adv Neural Inform Process Syst 25:1–9

Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 217–244). American Psychological Association.

Xanthopoulos, Petros, Panos M. Pardalos, and Theodore B. Trafalis. "Linear discriminant analysis." Robust data mining. Springer, New York, NY, 2013. 27-33.

Xu, Chang, Dacheng Tao, and Chao Xu. "A survey on multi-view learning." arXiv preprint arXiv:1304.5634 (2013)

Yu, Shipeng, et al. "Bayesian co-training." Advances in neural information processing systems 20 (2007).

Yvan Saeys, Iñaki Inza, Pedro Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics, Volume 23, Issue 19, 1 October 2007, Pages 2507–2517

Zhang, L., Romero, D., Giannakis, G.B.: Fast convergent algorithms for multi- kernel regression. In: Proc. Statistical Signal Process. Workshop. pp. 1–4. Palma de Mallorca, Spain (June 26-29, 2016).

Zhou, F., Wu, R., Xing, M. & Bao, Z. Eigensubspace-based filtering with application in narrow-band interference suppression for sar. IEEE Geosci. Remote Sens. Lett. 4, 75–79 (2007).

BIOGRAPHICAL SKETCH

Md Imrul Kaish earned his bachelor's degree in Electrical and Electronic Engineering from Islamic University of Technology (IUT), Dhaka, Bangladesh in December 2015. He served as a lecturer at Daffodil International University (DIU), a private university in Bangladesh, from September 2016 to December 2019, prior joining University of Texas Rio Grande Valley (UTRGV). He received Presidential Graduate Research Assistantship (PGRA) for two years master's program at UTRGV in January 2020. He completed his Master of Science in Electrical Engineering from UTRGV in May 2022. His research interests are focused applications of machine learning and dimensionality reduction. He published one conference paper in a workshop held in ACM in August 2021 and secured 3rd place in poster presentation competition in E-week, February 2022. He is an avid learner who wishes to pursue a career in AI-related disciplines. He can be contacted at mdimrul.kaish01@gmail.com.