# The Accuracy of KNN, SVM, neural network, naive Bayes classifier and DBSCAN, kmeans for Determining the Cancer by Gene-expression signatures

Hsieh Cheng-Han, Hsu Ting-Hao, Sun Shih-Yu, Lu Che-Yuan, Huang Chia-Yen

May 2023

## Abstract

Nowadays, using machine to give an early diagnosis of a cancer type is widely studied. In this paper, based on the gene-expression dataset on Synapse.org, we compare the performance of machine learning techniques, including data preprocessing, e.g., Principal components analysis (PCA), Autoencoder (AE), classification, e.g., k-nearest neighbors (KNN), support vector machine (SVM), neural network (NN), and clustering for dealing with unknown types, e.g., k-means, density-based spatial clustering of applications with noise (DBSCAN). The experiment results show that with PCA reducing the dimension of the dataset to 32, using KNN for classifying the known types, and finally use DBSCAN to cluster the unknown types, the accuracy of predicting can reach about 94%, which beats the other methods out.

breast invasive carcinoma (BRCA) colon adenocarcinoma (COAD) kidney renal clear cell carcinoma (KIRC) lung adenocarcinoma (LUAD) prostate adenocarcinoma (PRAD)

## 1 Introduction

Over the years, machine learning methods for early predict various disease are widely used in medical field. [1] [2] From diabetes [3], heart disease [4], to lung adenocarcinoma [5] etc., machine learning methods is universally studied attributed to the promising perspective. With a well-built database of disease and a well-choose combination of machine learning mathods, within seconds, a machine may determine the type of disease that this person may have by collecting the data from the person, and thereby achieve the goal of personalized medicine. As a result, machine learning methods have become a popular tool for medical researchers.

In real application, besides from the konwn cancers in the train data, there're numerous unkown types of cancers. When a new type of cancer appears, how shall the machine results? In ideal situation, this new type of cancer should be reported as an unkonwn type, which can be done by the combination of classification and clustering. However, how to choose the proper machine learning methods is a hot potato, since the different combinations of machine learning methods have different performance.

In this work, the dataset on Synapse.org is used, which contains 20531 RNA sequences and 3 types of cancer in the train data, and 2 new types in test data. To find out which techniques is more suitable to predict cancer by RNA sequence, this work examines the the accuracy of the combinations of some major machine learning techniques, including PCA, AE, KNN, SVM, NN, k-means, DBSCAN.

## 2 Related works

### Data preprocess

**Autoencoder (AE):** An AE is a type of neural network used for unsupervised learning and dimensionality reduction. It is designed to learn efficient representations or encodings of input data by training the network to recon-

struct its own inputs. The AE consists of an encoder and a decoder. The encoder compresses the input data into a lower-dimensional representation, often called the latent space or code. The decoder then aims to reconstruct the original input data from this compressed representation. By learning a compact representation of the input data, AEs can capture important features and discard noise or irrelevant information. This makes them effective in tasks such as image or text classification, where reducing the dimensionality of the data can improve performance.

**Principal components analysis (PCA):** PCA is a statistical technique used for dimensionality reduction and data exploration. It aims to transform a dataset with a large number of variables into a lower-dimensional space while preserving the most important information. PCA accomplishes this by identifying the principal components, which are linear combinations of the original variables. Applications of PCA span various fields, including image and signal processing, finance, genetics, and social sciences. It is particularly useful in scenarios where the dataset is high-dimensional and the interpretation and visualization of the data are challenging.

## Classification algorithm

**k-nearest neighbors (KNN) classification algorithm:** The KNN classification algorithm is a supervised learning method which is first developed by Fix and Hodges [6]. The idea of KNN is based on the idiom, "birds of a feather flock together". By picking the $k$-nearest neighbors of a data point, the unkonwn class label can be determined. Lots of works [7] [8] [9] show the fact that KNN performs well for prediction of diabetes disease.

**support vector machine (SVM):** Given a set of training datas, where each data is labeled as a binary class, such as $0$ and $1$, SVM training algorithm creates a model that assigns new examples to the binay labels by making it a non-probabilistic binary linear classifier. In addition, accroding to [10] [11], SVM can also use a method called kernel trick to effectively perform non-linear classification by implicitly mapping its inputs into a high-dimensional feature space.

**neural network (NN):** Neural network have been used in many fields to deal with intricate datas. With input layer, hidden layer and output layer constructed by neurons, each data in dataset is processed while passing through neurons, layer by layer. After the processing, the outcome can be used to predict.Using back propogation, the accuracy of predictions increase in each training. In order to construct the hidden layer more efficient, NAS(Neural Arcitecture Searching) is used to search suitable structure for hidden layer, increasing the accuracy. According to [12][13], many neural network have been constructed and trained already, with high efficiency and accuracy in prediction of diabetes.

## Clustering algorithm

**k-means:** k-means is a popular and easily implemented clustering method in machine learning and data analysis. With $n$ data points in a dataset, k-means algorithm partition them into $k$ distinct clusters. The algorithm works iteratively and converges to a solution by minimizing the sum of squared distance between the data points and the center of their cluster. K-means is a useful algorithm for clustering data, which is widely applied on many researches, e.g. [14], [15], [16].

**Density-based spatial clustering of applications with noise (DBSCAN):** DBSCAN is a data clustering algorithm proposed by Ester et al. [17] DBSCAN is particularly effective in discovering clusters of arbitrary shapes and handling noise in the data. Unlike k-means need user specify how many clusters, DBSCAN determine the clusters and noise automatically by the density of data points. This characteristic makes it particularly useful when addressing the datasets where the number of cluster is unkonwn. DBSCAN is the major clustering algorithm in machine learning field, since the ability to discover the number of clusters automatically. And the variants of DBSCAN [18] is also developed widely.

Table 1: COMPARISON OF CLASSIFICATION TECHNIQUES. (ARRHYTHMIA DATA SET)

| Method | Search (s) | acc | acc (normalized) | acc (PCA-32) | acc (PCA-32-normalized) | acc (AE-32) | acc (AE-32-normalized) |
|---|---|---|---|---|---|---|---|
| KNN-Brute-Force + DBSCAN | 0.014 | $44.3038 \pm 0$ | $46.8354 \pm 0$ | $38.6076 \pm 0$ | $42.4051 \pm 0$ | $34.8101 \pm 0$ | $44.3038 \pm 0$ |
| SVM (linear) + DBSCAN | 1.048 | $25.9494 \pm 0$ | $35.4430 \pm 0$ | $32.2785 \pm 0$ | $29.7468 \pm 0$ | $31.6456 \pm 0$ | $27.8481 \pm 0$ |
| SVM (polynomial) + DBSCAN | 0.765 | $25.9494 \pm 0$ | $25.3165 \pm 0$ | $1.2658 \pm 0$ | $33.5443 \pm 0$ | $30.3797 \pm 0$ | $30.2785 \pm 0$ |
| SVM (RBF) + DBSCAN | 5.298 | $25.9494 \pm 0$ | $22.7848 \pm 0$ | $0.6329 \pm 0$ | $30.3797 \pm 0$ | $22.1519 \pm 0$ | $34.1772 \pm 0$ |
| SVM (sigmoid) + DBSCAN | 2.298 | $28.4810 \pm 0$ | $20.2532 \pm 0$ | $30.3797 \pm 0$ | $36.0759 \pm 0$ | $22.1519 \pm 0$ | $24.6835 \pm 0$ |
| KNN-Brute-Force + kmeans | 0.014 | $53.1646 \pm 0$ | $50.6329 \pm 0$ | $43.0380 \pm 0$ | $43.0380 \pm 0$ | $36.0759 \pm 0$ | $41.7722 \pm 0$ |
| SVM (linear) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (polynomial) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (linear) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (sigmoid) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |

Table 2: COMPARISON OF CLASSIFICATION TECHNIQUES. (GENE EXPRESSION CANCER RNA-SEQ DATA SET)

| Method | Search (s) | acc | acc (normalized) | acc (PCA-30) | acc (PCA-30-normalized) | acc (AE-30) | acc (AE-30-normalized) |
|---|---|---|---|---|---|---|---|
| KNN-Brute-Force + DBSCAN | 4.317 | $59.0361 \pm 0$ | $66.5663 \pm 0$ | $94.5783 \pm 0$ | $78.3133 \pm 0$ | $64.4578 \pm 0$ | $59.3373 \pm 0$ |
| SVM (linear) + DBSCAN | 131.809 | $61.7470 \pm 0$ | $9.9398 \pm 0$ | $9.9398 \pm 0$ | $10.5422 \pm 0$ | $63.2530 \pm 0$ | $13.5542 \pm 0$ |
| SVM (polynomial) + DBSCAN | 31.221 | $61.7470 \pm 0$ | $1.2048 \pm 0$ | $64.4578 \pm 0$ | $61.1446 \pm 0$ | $63.2530 \pm 0$ | $50.3012 \pm 0$ |
| SVM (RBF) + DBSCAN | 3.988 | $X$ | $X$ | $9.9398 \pm 0$ | $63.2530 \pm 0$ | $40.3614 \pm 0$ | $49.6988 \pm 0$ |
| SVM (sigmoid) + DBSCAN | 0.446 | $X$ | $X$ | $71.6867 \pm 0$ | $65.9639 \pm 0$ | $44.2771 \pm 0$ | $38.2530 \pm 0$ |
| KNN-Brute-Force + kmeans | 4.859 | $74.0964 \pm 0$ | $53.6145 \pm 0$ | $94.8795 \pm 0$ | $84.9398 \pm 0$ | $42.1687 \pm 0$ | $65.9639 \pm 0$ |
| SVM (linear) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (polynomial) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (linear) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| SVM (sigmoid) + kmeans | $X$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $X \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |

# 3 main section

# 4 Experiment result

# 5 Conclusion

# References


[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[2] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[4] M. Learning, "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review," *Adv. Comput. Sci. Technol*, vol. 10, no. 7, pp. 2137–2159, 2017.

[5] L. Huang, L. Wang, X. Hu, S. Chen, Y. Tao, H. Su, J. Yang, W. Xu, V. Vedarethinam, S. Wu *et al.*, "Machine learning of serum metabolic patterns encodes early-stage lung adenocarcinoma," *Nature communications*, vol. 11, no. 1, p. 3556, 2020.

[6] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989. [Online]. Available: http://www.jstor.org/stable/1403797

[7] M. NirmalaDevi, S. A. alias Balamurugan, and U. V. Swathi, "An amalgam knn to predict diabetes mellitus," in *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, 2013, pp. 691–695.

[8] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5.

[9] V. Vijayan and A. Ravikumar, "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus," *International journal of computer applications*, vol. 95, no. 17, 2014.

[10] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[11] M. Hofmann, "Support vector machines-kernels and the kernel trick," *Notes*, vol. 26, no. 3, pp. 1–16, 2006.

[12] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," *Journal of Ambient Intelligence and Humanized Computing*, Apr 2020. [Online]. Available: https://doi.org/10.1007/s12652-020-01963-7

[13] T. Beghriche, M. Djerioui, Y. Brik, B. Attallah, and S. B. Belhaouari, "An efficient prediction system for diabetes disease based on deep neural network," *Complexity*, vol. 2021, p. 6053824, Dec 2021. [Online]. Available: https://doi.org/10.1155/2021/6053824

[14] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k means clustering algorithm for prediction of students academic performance," 2010.

[15] N. Nidheesh, K. Abdul Nazeer, and P. Ameer, "An enhanced deterministic k-means clustering algorithm for cancer subtype prediction from gene expression data," *Computers in Biology and Medicine*, vol. 91, pp. 213–221, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482517303402

[16] M. S. Kadhm, I. W. Ghindawi, and D. E. Mhawi, "An accurate diabetes prediction system based on k-means clustering and proposed classification approach," *International Journal of Applied Engineering Research*, vol. 13, no. 6, pp. 4038–4041, 2018.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[18] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 2014, pp. 232–238.