Xi'an Jiaotong-Liverpool University
西交利物浦大學

Covid-19 CT Image Data Classification Using Principal Component Analysis (PCA) and K-Nearest Neighbour (KNN) Classifier, Naive Bayes Classifier and Multi-layer Perceptron (MLP) Neural Networks

1. Introduction

The rise in mortality rate as the result of the notoriously asymptomatic COVID-19 influenza pandemic has pushed the global community to develop screening tools to rapidly detect mild signs of virus incubation inside human lungs. Multiple studies have demonstrated that, among these tools, Machine Learning algorithms significantly outperform human clinicians in the rapid and accurate diagnosis of COVID-19 symptoms (Chen, *et al*., 2022; Gomez *et al*., 2022). Machine Learning applications relatively excel at detecting indications of early pulmonary infection in COVID-19 CT scan images that are occasionally difficult to discern (Gangloff *et al*., 2021; Zoabi *et al*., 2021). Most unprocessed image datasets tend to be high-dimensional (containing many variables). Scaling up, it would **severely slow down training processes**, as the current generation of GPU struggles to process data points from large quantities of raw image data.

Dimensionality Reduction techniques can be applied to address this issue. Dimensionality Reduction techniques aim to lessen most of the non-essential variables, enabling effective transfer learning while slowly increasing model accuracy (Bharadiya, 2023). One such example is the Principal Component Analysis (PCA). PCA accomplishes Dimensionality Reduction by linearly transforming the original feature space into a smaller set of orthogonal components (the principal components) using the "eigen-decomposition method" (Raschka, 2015). In other words, it finds the key elements by linearly dissecting the most useful variable groups in the data. PCA is useful in maintaining consistent training accuracy, especially if a Machine Learning model is designed to process large data inflow such as picture variables.

This study aims to investigate the effects of reducing the dimensionality of COVID-19 images by applying PCA to extract 20 principal components. For effective comparison, evaluation will be conducted on the three machine learning models: k-Nearest Neighbour (kNN), Naïve Bayes, and Multi Layer Perceptron Neural Networks (MLP-NN). Our main objective is to observe and evaluate the theoretical accuracy increases through applying PCA with a default parameter of 20 features. Additionally, we will determine the extent dimensionality reduction through PCA can enhance classification accuracy while preserving essential variables.

2. Principal Component Analysis Method and Experimental Design

2.1 In-depth Analysis of PCA

Principal Component Analysis (PCA) is a widely used method for simplifying large datasets by reducing the dimensionality while preserving as much variance as possible (Raschka, 2015; Bharadiya, 2023). The core idea behind PCA is to identify patterns in the data that show the biggest variations and linearly transform them. These patterns are called the principal components. By identifying these key directions in which the data varies the most, PCA allows us to represent the data more efficiently.

Intuitively, PCA could be imagined as an oval shape (ellipsoid) containing clusters of datasets. Each side of the oval represents a principal component. If a side is small, it means the data doesn't vary much in that direction.

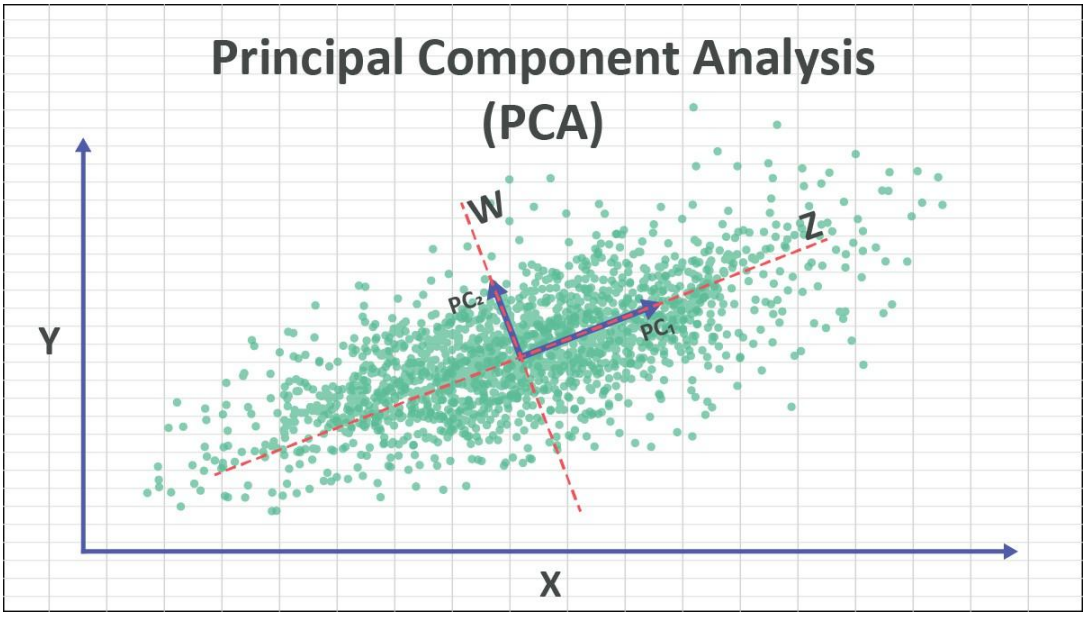

Image 1 - Illustration of How PCA linearly divides the dataset. Source: Principal Component Analysis (PCA) 101 - NumXL

To mathematically find the sides of the oval, we have to perform the following steps:

1. Moving the data to the center by subtracting the mean value from each data point. These values will be known consecutively as the mean vector value x_t and must fulfil this condition*:*

$$x_t = (t = \{n| \in N\}\,;\, \Sigma_{t=1}^{n} x_t x_t^T)$$

2. We could then calculate the covariance matrix (C) along with its eigenvalues and eigenvectors (U). The covariance matrix is calculated adhered to the formula:

$$C = \frac{1}{n} \Sigma_{t=1}^{n} x_t x_t^T$$

3. Assume that the dimension of the dataset h = Σx_t , x_t will then be converted into vector s_t by:

$$s_t = U^T x_t \quad (1)$$

4. Keep in mind that x_t are plural (not only one). To aggregate the entirety of x_t we can slightly modify the equation to:

$$s_t(n) = U_n^T x_t \quad (2)$$

   - "n" represents the dimension (h) of the dataset

PCA is needed for reducing data dimensionality as it eliminates redundancy by combining similar features into fewer, more distinct ones, making the data more compact. This improves computational efficiency for faster processes and reduces memory usage and training time (Bharadiya, 2023). PCA is also intended to help with the issue of overfitting (citation), where too many features cause a model to yield maximum training accuracy, but perform poorly on test accuracy. By focusing on the most important data points, PCA assists models to generalise better and reduces overfitting.

2.2 Experiment Design

In this experiment, we utilize the COVID-19 CT images which consist of 746 figures (397 tested negative and 349 tested positive). In the initialization stage, PCA will be applied to reduce the dimensionality of the dataset to 20 principal components (n_components = 20). We expect that using the reduced feature set would lead to an increase in the model efficiency and speed without significantly compromising the classification accuracy.

Our experimental setup consists of comparing three different machine learning models: K-Nearest Neighbour (KNN) classifier, Naive Bayes classifier, and Multi-Layer Perceptron Neural Network (MLP-NN) Neural Networks, on 2 versions of the dataset:

1. Original dataset
2. PCA-applied dataset

Extra note: both datasets will be equally separated into training and test sets.

3. Experimental Results with Analysis

Table 1: Three model performances using the 20 new features and original features

Training Performance		
Models	Original features	20 new features (PCA)
kNN	**77.48%**	**82.57%**
Naive Bayes	**70.00%**	**72.00%**
MLP-NN	**87.02%**	**100.00%**
Test Performance		
kNN	**70.24%**	**74.80%**

Naive Bayes	**69.44%**	**72.12%**
MLP-NN	**70.00%**	**78.00%**

In general, Table 1 shows the difference in percentages between machine learning algorithms before and after the application of PCA. There is a trend of improvement that can be identified consistently in all 3 algorithms. All improvements can be summarised as follows:

1. Training phase: MLP-NN exhibit a dramatic accuracy increase (20%) followed by kNN (roughly 5%), and Naive-Bayes (2%)
2. Testing phase: MLP-NN demonstrated a respectable increase (roughly 8%) followed by kNN (around 4.5%) and Naive-Bayes (2.5% relative improvement).

By large, we can deduce that MLP-NN outperforms both kNN and Naive Bayes in overall performance.

The table also demonstrates that the MLP-NN, when using PCA, achieved perfect accuracy during training. However, a sharp decline in accuracy (-22%) was observed in the testing period, indicating that the MLP-NN likely experiencing overfitting. We suspect that there are discrepancies present between training and test set due to feature over-removal. This implication contrasts with one of the primary objectives of applying PCA (section 2.1: reducing overfitting)

Table 2: Ranking 20 new features using PCA and all data

PCA	Accuracy using k-NN	Accuracy using Naive Bayes	Accuracy using MLP-NN
Feature 1	57.37%	64.88%	65.15%
Feature 2	67.56%	66.76%	66.76%
Feature 3	70.78%	68.36%	66.08%
Feature 4	69.97%	64.34%	66.29%
Feature 5	67.56%	64.61%	67.67%

Feature 6	67.56%	64.62%	66.95%
Feature 7	64.88%	64.88%	68.76%
Feature 8	68.90%	65.95%	72.39%
Feature 9	70.51%	65.95%	73.99%
Feature 10	69.71%	67.83%	74.26%
Feature 11	69.97%	71.31%	75.28%
Feature 12	72.12%	70.51%	77.68%
Feature 13	73.99%	73.19%	76.60%
Feature 14	73.73%	72.12%	77.48%
Feature 15	73.99%	73.19%	76.60%
Feature 16	75.34%	**74.26%**	77.21%
Feature 17	**75.60%**	72.92%	75.34%
Feature 18	75.07%	71.05%	76.68%
Feature 19	75.07%	70.78%	76.14%
Feature 20	74.80%	72.12%	**78.08%**

The feature ranking results provide insights into which feature dimensions capture the most relevant information for COVID-19 CT image classification. Similar to Table 1, MLP-NN outperforms both k-NN and Naive Bayes classifiers in terms of accuracy across most of the features, with its highest accuracy being 78.08% with feature 20. k-NN shows a pattern of performing slightly lower than MLP-NN across all features, with its highest accuracy being 75.60% in feature 17. Naive Bayes on the other hand, has the lowest performance among all three models, with its best accuracy being 74.26% when using feature 16.

An interesting observation can be seen in Feature 13 and 15, where the accuracy rates of all three models are identical. Features 1-7 generally have a lower accuracy rate for all models, suggesting that these features may not capture enough variance for classification as some of the higher-ranked features (these less informative features could be considered for exclusion in scenarios where computational efficiency is important). In contrast to that, features 12-20 tend to have higher accuracies across all models, suggesting that they are the most informative among the PCA-derived features.

From these observations, it can be seen that MLP-NN appears to be less sensitive to individual feature variations, maintaining a relatively high accuracy across the features. This suggests that MLP- NN has a stronger ability to generalize from reduced features. The Naive Bayes classifier shows more variance in its performance with different features, which could be due to its reliance on feature independence assumptions. It performs best with feature 16, indicating that this feature aligns well with the assumptions of Naive Bayes. The k-NN classifier shows quite variability in its accuracies, which could be due to its sensitivity to the chosen feature space in terms of distance calculations. It achieved its best accuracy in features 16 and 17, indicating that these features might provide better separation between classes in the reduced feature space.

4. Conclusion

The application of PCA alongside ML algorithms in identifying COVID-19 CT images offers some advantages and limitations. Our experiment has proven that PCA is useful in heightening model accuracy and boosting training time, allowing for greater efficiency when working with high- dimensional data. However, the drawback is the potential loss of critical information, resulting in the algorithms overfitting. Moreover, PCA can introduce uncertainty in feature ranking because the transformation obscures the interpretability of the original features in relation to their individual contribution to classification.

For recommendations, raising the number of features (n_components > 20) could potentially help retain more of the crucial classification information. Future work might also explore alternative dimensionality reduction methods, such as SVD or autoencoders, to assess whether they provide better results for COVID-19 classification. Additionally, leveraging advanced neural network architectures, such as Convolutional Neural Networks (CNNs), which are specifically designed for image data, could offer greater robustness in handling high-dimensional CT data without the need for as much feature reduction. These steps could contribute to further refining the accuracy and efficiency of COVID-19 prediction models.

References

Bharadiya, J. (2023). *A tutorial on principal component analysis for dimensionality reduction in machine learning.* Research Gate. https://www.researchgate.net/profile/Jasmin-Bharadiya-4/publication/371306692_A_Tutorial_on_Principal_Component_Analysis_for_Dimensionality_Reduction_in_Machine_Learning/links/647e1fc72cad460a1bf88e90/A-Tutorial-on-Principal- Component-Analysis-for-Dimensionality-Reduction-in-Machine-Learning.pdf

Chen, J. *et al.* (2022). Machine learning techniques for CT imaging diagnosis of novel coronavirus pneumonia: A Review. Neural Computing and Applications, 36(1), 181–199. https://doi.org/10.1007/s00521-022-07709-0

Gangloff, C., Rafi, S., Bouzillé, G., Soulat, L., & Cuggia, M. (2021). Machine learning is the key to diagnose COVID-19: A proof-of-concept study. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86735-9

Gomes, R. *et al*. (2022). A comprehensive review of machine learning used to combat COVID-19. *Diagnostics*, *12*(8), 1853. https://doi.org/10.3390/diagnostics12081853

Raschka, S. (2015). *Principal Component Analysis in 3 Simple Steps*. Sebastian Raschka, PhD. https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#:~:text=The%20classic%20approach%20to%20PCA,each%20element%20represents%20the%20covariance

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-020- 00372-6