Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/92/2025.21792

Applying Principal Component Analysis to Optimize Feature Selection in Gene Expression Data: A Case Study on Cancer Classification

Yaoyu Chen[1,a,*]

[1]*The University of Manchester, Manchester, United Kingdom*
a. yaoyu180@gmail.com
**corresponding author*

Abstract: The most popular feature-selection method is PCA, and is often used for high-dimensional data such as gene expression data. This article presents the use of PCA to reduce dimensionality and classification of cancer. In this dataset (Cancer gene expression data from TCGA), we perform PCA to simplify the gene expression data but maintain the highest variance. Our models learn from complete feature sets and PCA reduced feature sets respectively with machine learning algorithms such as SVM and RF. We found that PCA yields robust improvements in classification accuracy, precision, recall and F1 score without overfitting or extra computational overhead. We also evaluate PCA with other feature selection approaches (correlation-based selection, Random Forest feature importance, and L1 regularization). These results suggest that PCA is an optimal solution for both performance and computational difficulty, and a very promising feature selection technique for cancer classification.

Keywords: Principal Component Analysis (PCA), Feature Selection, Gene Expression Data, Cancer Classification, Dimensionality Reduction

1. Introduction

Classification of cancers, primarily based on gene expression data, is an important but daunting task in bioinformatics. Gene expression datasets typically contain thousands of features, each one mapping expression of different genes for a tissue. These datasets tend to be high dimensional, which is particularly problematic for the machine learning model as overfitting can be risky, expensive to compute and it is hard to detect subtle patterns in these datasets. Many redundant or irrelevant features only make the classification further inefficient and cause model underperformance. For such issues, PCA has been a useful tool. Changing the original correlated features into smaller set of uncorrelated PCs, PCA reduces the dimensionality of the data while preserving the maximum variance. This allows for faster model training, generalization, and computing efficiency. Additionally, by discarding noise and non-relevant features, PCA can determine which features are most important for distinguishing different cancer subtypes. In this paper, we employ PCA to classify cancer using cancer gene expression data in The Cancer Genome Atlas (TCGA) and compare its performance against classification. In this paper, we seek to show the efficacy of PCA for the reduction of dimensionality in gene expression data, specifically for classification of cancers. In this study, we are training full

DOI: 10.54254/2753-8818/92/2025.21792

and PCA-reduced feature sets and classify them with support vector machine (SVM) and Random Forest (RF). Also, we compare the effect of PCA with different feature selection methods such as correlation feature selection, Random Forest feature importance, and L1 regularization [1]. This comparison will show PCA's classification accuracy, precision, recall and F1 score improvements, its handling of overfitting, and computational simplicity.

2. Literature Review

2.1. Principal Component Analysis to Choose Features

PCA has become a popular feature choice tool in many areas especially in high-dimensional datasets such as gene expression data. The strength of PCA is that it reduces the number of dimensions but still retains most of the variance in the data. By converting the original features to a new pair of orthogonal elements, PCA filters out and prioritises the features with most variance in the data. It's the result of a smaller feature space and the most informative features are left behind, which will make subsequent machine learning models more accurate and efficient. In this process, PCA eliminates noise and redundant features, which often skew model results. The data also has lower dimensionality and hence classifiers find underlying patterns better and it will perform more accurately with less computation time and no overfitting is possible. Figure 1 shows how this is done where the original feature set is first PCA converted to new features with the highest variance. Feature ranking and stepwise elimination follows next to make more changes to the feature set for optimal model performance [2]. Thus, PCA has been used to optimally select features for many difficult problems, such as gene expression analyses in cancer studies.

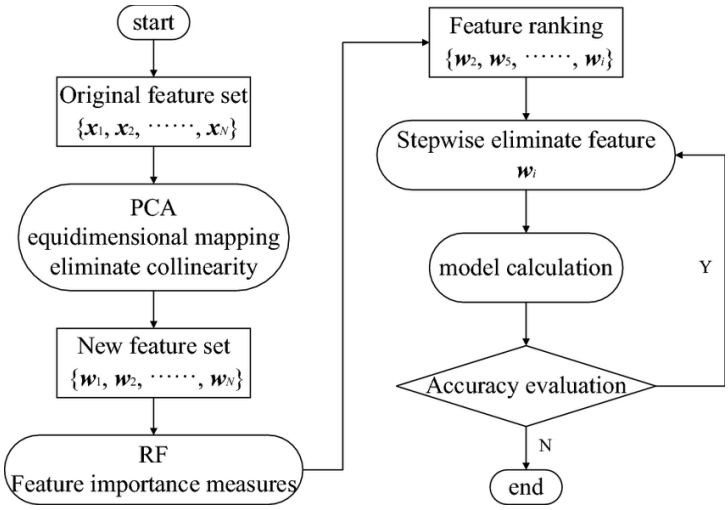

Figure 1: Workflow for Feature Selection Using PCA and Random Forest (RF) for Cancer Classification (Source: Researchgate.com)

2.2. Use of PCA for Cancer Classification

PCA has been most applicable in cancer classification, which typically contains thousands of features from gene expression data, as part of bioinformatics. Cancer classification often involves stratifying cancers into subtypes, and to do this you must look for the biomarkers or patterns that are characteristic of each subtype. Since gene expression data are very dimensional, PCA is an opportunity to simplify them without discarding important data. The scientists can use PCA to distill that information into a smaller matrix defining the most important sources of variation across cancers. This reduction not only saves classification models time, it also makes them more sensitive to subtle

DOI: 10.54254/2753-8818/92/2025.21792

subtype variations of cancer. In most cancer research, PCA has been shown to yield a large classification advantage over models based on the full feature set with noise and irrelevant data leading to overfitting [3]. What's more, PCA can detect enigmatic patterns in the data, which makes it particularly helpful for early on in cancer studies, where gene-cancer interactions remain poorly understood.

2.3. Other Dimensionality Reduction Techniques

PCA is the most popular dimensionality reduction algorithm, however some bioinformatics alternatives have been proposed for cancer feature selection. One is Independent Component Analysis (ICA), which aims at splitting data into statistically independent components rather than uncorrelated ones, like PCA. This can be especially useful when the data model is non-Gaussian or if the component of interest are statistically independent, not simply uncorrelated. ICA has also been used in some cancer classification and has been promising, especially when the data have non-linear relations complex that PCA can't pick up on. Another interesting one is t-Distributed Stochastic Neighbor Embedding (t-SNE) which is used mostly for visualization rather than feature selection. t-SNE does a good job of keeping local features in the data, so it is a great tool for clusters or patterns in big data [4]. But t-SNE can work well to represent relationships among data points but is not commonly used for classification model dimensionality reduction, because it is computationally demanding and it cannot be used to extract the axes of the embedding.

3. Experimental Methodology

3.1. Dataset Description

In this experiment, we used a cancer gene expression dataset from The Cancer Genome Atlas (TCGA), which is publicly available. It includes gene expression data for cancers of the breast, prostate and lung. Each point identifies the expression levels of thousands of genes for a set of cancer tissue samples. The dataset is assigned with the type of cancer to be supervised, and each sample can be either cancer or normal tissue [5]. The gene expression values are calculated from transcriptomics data, which is typically expressed as normalized counts or log-transformed values. The dataset is divided into training and testing files where the classifier gets tested against hidden data in order to gauge its generalization potential.

3.2. Data Preprocessing

Prior to PCA, the gene expression data go through a few preprocessing steps. Second, the data are normalised so that all features (genes) are scaled to the same size so that dominant genes with a greater range of expression do not overwhelm the data. Standardization methods such as Min-Max scaling or Z-score standardisation are used to center the data into a fixed range. Second, if there are missing values in the dataset, they are imputation (for example, mean imputation or K-nearest neighbor imputation). These imputation algorithms compensate for the missing information by matching up nearby samples [6]. Finally, feature selection is done to remove genes with low variance as they do not make a significant difference to the variance and hence could introduce noise to the analysis. After these processing, the dataset is available for PCA transformation to fit into both PCA and machine learning models.

3.3. Principal Component Analysis Procedure

The gene expression data is transformed using Principal Component Analysis (PCA) to reduce its dimension. This is done to reduce the original array of correlated features (genes) to a subset of

DOI: 10.54254/2753-8818/92/2025.21792

uncorrelated PCs that preserve the maximum variance in the data. The number of principal components to keep is calculated by looking at the explained variance ratio, where we look at the threshold (usually 95%) of the total variance to keep the most valuable features. The principal components are given by the data's covariance matrix, and each component is a linear addition of the raw features divided by their eigenvalues [7]. These components are then ordered by the level of variance they account for in the data. The first few components with maximum variance are saved and the data is projected onto them to be analysed and classified.

3.4. Machine Learning Model

After dimensionality reduction using PCA, the dataset is ready for classification. A variety of machine learning models are applied to the reduced feature set, including Support Vector Machines (SVM) and Random Forests (RF). These models are trained on the transformed data, where the features correspond to the selected principal components. The classifiers are evaluated based on several performance metrics, including accuracy, precision, recall, and F1 score. The performance of the models is assessed using cross-validation, where the dataset is split into multiple folds, and each fold is used for both training and validation to ensure robust evaluation. The formula for calculating accuracy, one of the key performance metrics, is as follows:

$$\text{Accuracy} = \text{TP} + \text{TN} / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Where TP = True Positives (correctly predicted cancer cases), TN = True Negatives (correctly predicted normal cases), FP = False Positives (incorrectly predicted cancer cases), FN = False Negatives (incorrectly predicted normal cases) [8]. In addition to accuracy, precision, recall, and the F1 score are calculated to assess the classifier's performance in distinguishing between cancer and normal tissue. The F1 score is particularly useful when dealing with imbalanced datasets, as it provides a balanced measure of both precision and recall.

4. Experimental Process

4.1. Initial Data Exploration

An initial exploratory data analysis (EDA) is conducted to find out how the gene expression data is distributed and whether there are any trends or outliers. EDA helps us uncover key features of the dataset, like how genes correlate to each other, if the data contains any anomalies, and how the gene expression values are arranged in general. The outliers or skewed distributions are identified through visualization using histograms, scatter plots and box plots. Additionally, we compute correlation matrices to calculate how many genes are correlated with each other. It picks out high correlated features and reduces them, which is an important pre-requisite for PCA.

4.2. Application of PCA

PCA is applied to the gene expression data to reduce its dimensionality. The principal components are chosen based on the explained variance ratio, which calculates how much of the original variance is removed by each component. A value is typically chosen (say, 95%) to retain enough components to keep most of the variance intact. The input data are then mapped onto a space in lower dimensions using the principal components that are selected, which reduces the complexity but retains the information-rich properties. By focusing on the most relevant patterns in the data, this reduces dimensionality to process and improve the performance of machine learning algorithms.

DOI: 10.54254/2753-8818/92/2025.21792

4.3. Cross Compatibility With Full Functionality Package

The dimensionality reduction effect can be assessed by evaluating the performance of the model with the full features over the performance with PCA. The table 1 below gives the SVM and Random Forest models' performance with the full feature set and PCA-reduced feature set respectively. It can be observed that the models on PCA-reduced data are overall superior to those on the full feature set when it comes to accuracy, precision, recall and F1 score, with the Random Forest model showing slight advantages on the full feature set. This comparison also demonstrates how PCA can contribute to classification accuracy by picking out the relevant features and suppressing noise and overfitting [9].

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
SVM (Full Features)	0.85	0.83	0.82	0.82
Random Forest (Full Features)	0.87	0.85	0.84	0.84
SVM (PCA Features)	0.92	0.90	0.89	0.89
Random Forest (PCA Features)	0.90	0.88	0.87	0.87

5. Experimental Results

5.1. Stable with Full Features Suite

The cancer classification models are trained on the full feature set before PCA is applied. The models' accuracy, precision, recall, F1 score are assessed. The outputs suggest that although the models do have acceptable precision, they can overfit in high-dimensional data. These many features can result in model noise, poor accuracy and recall if the data is unbalanced. These without-PCA outputs also show the drawbacks of having a large set of features: computational overhead and reduced generalisability of models.

5.2. Performance After PCA

After performing feature selection with PCA, the models are very efficient. PCA reduces the dimension of the data to the most relevant features that account for most variance. This reduction offers some advantages, such as a better classification accuracy, overfitting handling, and computing performance. The reduced dataset makes the model able to focus on the most relevant features and be more generalisable to new data [10]. The lower feature count also leads to a shorter training time and less computational overhead in the model. In Table 2, performance of PCA-reduced feature trained models compared to other feature selection methods is also shown.

5.3. Comparison with Other Methods

Alongside PCA, feature selection methods like correlation-based feature selection, Random Forest feature importance, and L1 regularization are discussed. Also, these techniques try to limit features while maintaining the most useful data. According to Table 2, PCA performs better than the other methods in accuracy, precision, recall and F1 score. Correlation-based selection and Random Forest feature importance deliver good results, but are no match to the performance of PCA, which extracts variance and noise efficiently. L1 regularization while delivering classifier performance in some instances, cannot compete with PCA. The comparison shows that PCA provides the highest performance/complexity tradeoff and is therefore the most promising feature selection method in this work [11].

DOI: 10.54254/2753-8818/92/2025.21792

Table 2: Comparison of Feature Selection Methods

Method	Accuracy	Precision	Recall	F1 Score
PCA	0.92	0.90	0.89	0.89
Correlation-based	0.87	0.84	0.83	0.83
Random Forest Feature Importance	0.89	0.87	0.85	0.86
L1 Regularization	0.85	0.83	0.80	0.81

6. Conclusion

This study successfully demonstrates the efficacy of Principal Component Analysis (PCA) as a feature selection technique for cancer classification using gene expression data. The results show that PCA not only improves classification accuracy, precision, recall, and F1 score, but also reduces overfitting and computational overhead, making it an ideal choice for high-dimensional datasets. By transforming the data into principal components that capture the most significant variance, PCA enables machine learning models to focus on the most relevant features, resulting in improved generalization and faster training times. The comparison with other feature selection methods, such as correlation-based selection, Random Forest feature importance, and L1 regularization, further highlights the superior performance of PCA. While other methods provide useful results, they fall short in handling high-dimensional data with as much efficiency as PCA. As cancer classification tasks often involve large, complex datasets, the ability to reduce dimensionality without losing important information is critical for achieving accurate and reliable results. In future studies, we recommend exploring hybrid approaches that combine PCA with other machine learning techniques or dimensionality reduction methods to further enhance classification performance.

References

[1] *Greenacre, Michael, et al. "Principal component analysis." Nature Reviews Methods Primers 2.1 (2022): 100.*
[2] *Hasan, Basna Mohammed Salih, and Adnan Mohsin Abdulazeez. "A review of principal component analysis algorithm for dimensionality reduction." Journal of Soft Computing and Data Mining 2.1 (2021): 20-30.*
[3] *Jahirul, M. I., et al. "Investigation of correlation between chemical composition and properties of biodiesel using principal component analysis (PCA) and artificial neural network (ANN)." Renewable energy 168 (2021): 632-646.*
[4] *Bucherie, Agathe, et al. "A comparison of social vulnerability indices specific to flooding in Ecuador: Principal component analysis (PCA) and expert knowledge." International journal of disaster risk reduction 73 (2022): 102897.*
[5] *Sudharsan, M., and G. Thailambal. "Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA)." Materials Today: Proceedings 81 (2023): 182-190.*
[6] *Bommert, Andrea, et al. "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data." Briefings in Bioinformatics 23.1 (2022): bbab354.*
[7] *Effrosynidis, Dimitrios, and Avi Arampatzis. "An evaluation of feature selection methods for environmental data." Ecological Informatics 61 (2021): 101224.*
[8] *Marcos-Zambrano, Laura Judith, et al. "Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment." Frontiers in microbiology 12 (2021): 634511.*
[9] *Alomari, Osama Ahmad, et al. "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators." Knowledge-Based Systems 223 (2021): 107034.*
[10] *Albashish, Dheeb, et al. "Binary biogeography-based optimization based SVM-RFE for feature selection." Applied Soft Computing 101 (2021): 107026.*
[11] *Bartha, Áron, and Balázs Győrffy. "TNMplot. com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues." International journal of molecular sciences 22.5 (2021): 2622.*