SPU 53rd ANNIVERSARY

NATIONAL AND INTERNATIONAL SRIPATUM UNIVERSITY CONFERENCE 2023

The 18th National and The 8th International Sripatum University Conference

27 OCTOBER 2023

“หนังสือประมวลบทความ PROCEEDINGS”

การประชุมวิชาการระดับชาติ ครั้งที่ 18
และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8
มหาวิทยาลัยศรีปทุม

เรื่อง “วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน”

Organized by Sripatum University in cooperation with
• The University of Palermo, Italy • Sholokhov Moscow State University for the Humanities,Russia
• Universidad de Colima, Mexico • University of Taipei, Taiwan
• Institut Teknologi Sepuluh Nopember, Indonesia • The Joint Graduate School of Energy and Environment
• The Social Science Research Association of Thailand • Lawyers Council Under the Royal Patronage
• Thai Federation on Logistics • The Institute of Internal Auditors of Thailand • Prachachuen Research Network
• Journal Network of Social Sciences and Humanities • Program Management Unit for Human Resources
& Institutional Development, Research and Innovation (PMU-B)

หนังสือประมวลบทความ (Proceedings)
การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการ ระดับนานาชาติ ครั้งที่ 8
มหาวิทยาลัยศรีปทุม ออนไลน์ ประจำปี 2566
เรื่อง วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน
(Research and Innovations to Sustainable Development)

วันศุกร์ที่ 27 ตุลาคม 2566

รวบรวมโดย

คณะกรรมการพิจารณาผลงาน

การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8 ประจำปี 2566

ออกแบบปกโดย งานกราฟิกและศิลปกรรม มหาวิทยาลัยศรีปทุม

จัดรูปเล่มโดย โรงพิมพ์ มหาวิทยาลัยศรีปทุม

หนังสือประมวลบทความ (Proceedings)

การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8 มหาวิทยาลัยศรีปทุม ออนไลน์

เรื่อง วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน

The Proceedings of the 18th National and the 8th International Sripatum University Conference : Research and Innovations to Sustainable Development

วันที่: 27 ตุลาคม 2566

Date: 27 October 2023

ISBN (e-book): 978-974-655-469-5

ข้อมูลทางบรรณานุกรมของหอสมุดแห่งชาติ

หนังสือประมวลบทความการประชุมวิชาการระดับชาติ ครั้งที่ 18 และระดับนานาชาติ ครั้งที่ 8 มหาวิทยาลัยศรีปทุม ออนไลน์ เรื่อง การวิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน.-- พิมพ์ครั้งที่ 18.-- กรุงเทพฯ: มหาวิทยาลัยศรีปทุม, 2566.

3383 หน้า.

1. การประชุมวิชาการ. 2. บทความวิจัย. 3. บทความวิชาการ. I. ชื่อเรื่อง.

060

เจ้าของ

มหาวิทยาลัยศรีปทุม

จัดทำโดย

ศูนย์ส่งเสริมการวิจัยและการประกันคุณภาพการศึกษา มหาวิทยาลัยศรีปทุม

สถานที่จัดพิมพ์และจัดทำรูปเล่ม

โรงพิมพ์ มหาวิทยาลัยศรีปทุม

2410/2 ถนนพหลโยธิน แขวงเสนานิคม เขตจตุจักร กรุงเทพฯ 10900 โทร. 02 579 1111 ต่อ 1552

สารบัญ

สารบัญบทความ (ต่อ)

หน้า

Comparing Two Feature Reduction Techniques: Principal Component Analysis and Analysis of Variance for Classification Enhancement in Machine Learning

Kamolwan Khongjeen

School of Mathematics, Institute of Science, Suranaree University of Technology

E-mail: kamolwan.kho@gmail.com

Jessada Tanthanuch

School of Mathematics, Institute of Science, Suranaree University of Technology

E-mail: jessada@g.sut.ac.th

Eckart Schulz

School of Mathematics, Institute of Science, Suranaree University of Technology

E-mail: eckart@math.sut.ac.th

ABSTRACT

Machine learning classification serves as a vital tool for addressing a diverse array of real-world challenges, automating decision-making, improving efficiency, and providing insights from data. Within this tapestry of machine learning, feature reduction constitutes an important tool, wielding the potential to profoundly shape the performance, resilience, and interpretability of classification models. This research embarks on an exploration of Analysis of Variance (ANOVA) and Principal Component Analysis (PCA) in the realm of feature selection and synthesis. The investigation is focused on a dataset drawn from Köklü, comprising a wide range of 106 distinctive features representing the shapes and colors of five rice varieties. Subsets of various sizes of this dataset are used to comparatively evaluate the performance of dimensionality reduction by ANOVA feature selection and PCA feature generation, using five distinct machine learning techniques: K-Nearest Neighbors, Decision Tree, Random Forest, Multilayer Perceptron, and Support Vector Machine. The results show that six features created by PCA suffice to obtain good classification accuracies for all models except the decision tree. Feature reduction by Principal Component Analysis shows noticeably better classification accuracies as compared to ANOVA feature selection, particularly for small-sized datasets.

Keywords: feature synthesis, ANOVA, PCA, classification

1. Introduction

Classification through machine learning has wide-ranging applications that significantly impact efficiency, accuracy, decision-making, and innovation across industries, making it a fundamental and indispensable aspect of the modern world. However, it is important to be aware of the potential weak points and challenges associated with this approach. Machine learning models heavily depend on the quality and quantity of

data and computational resources, which can make them impractical for deployment in data-constrained environments. Feature synthesis techniques aim to enhance the quality of input features used in machine learning models by creating new features that capture meaningful information or relationships within the data. These techniques can potentially improve the performance of classification models (Patel, 2021). Here are the steps to apply feature synthesis techniques for improving classification in machine learning:

1. Understand the relationships between features and the target variable.

2. Process the data by performing tasks such as data cleaning, handling missing values, and scaling features.

3. Perform feature selection.

4. Conduct exploratory data analysis (EDA) to identify potential feature relationships and patterns in the data.

5. Identify relationships and interactions between features.

6. Engage in feature engineering by creating new features or transforming existing ones to capture relevant information.

7. Perform feature selection again.

8. Train models, evaluate them, tune model parameters, and repeat the process to fine-tune their performance.

In this research, two feature reduction techniques, namely Analysis of Variance (ANOVA) and Principal Component Analysis (PCA), are investigated and compared. In order to assess their performances on large as well as small data sets, the rice data set freely available at https://www.muratkoklu.com/datasets/ (Köklü *et al.*, 2021) is chosen, as it is large-scale and feature-rich. It contains 75,000 data samples and comprises 106 distinct features representing the shapes and colors of rice from five different rice varieties: Arborio, Basmati, Ipsala, Jasmine, and Karacadag. The reduced feature sets obtained from ANOVA and PCA techniques are employed in training five different machine learning techniques: K-Nearest Neighbors (KNN), Decision Tree (Dtree), Random Forest (Rforest), Multilayer Perceptron (MLP), and Support Vector Machine (SVM). The accuracies of all these models are then compared and analyzed.

2. Research Objective

To assess and compare the effectiveness of two feature reduction methods, ANOVA and PCA, in lowering the dimensionality of a feature-rich dataset.

3. Literature Review

3.1 Machine Learning Algorithms

3.1.1 K-Nearest Neighbors (KNN)

The *K*-nearest neighbors (KNN) algorithm is a non-parametric supervised learning algorithm that works by finding the *K* most similar points to a new data point, and then assigning the new data point to the same

class as the majority of its *K* nearest neighbors. The choice of the hyperparameter *K* is crucial. A higher *K* reduces sensitivity to noise but may decrease accuracy, while a lower *K* increases sensitivity to noise but may enhance accuracy (Sun, Du, & Shi, 2018).

3.1.2 Decision Trees (Dtree)

The decision tree algorithm is a versatile supervised learning method for classification and regression tasks. It constructs a tree structure with nodes testing data features and leaf nodes representing class labels or predicted values. Notably, it's a non-parametric algorithm, making it applicable to diverse data types without assuming any specific data distribution (Patel & Prajapati, 2018).

3.1.3 Random Forest (Rforest)

A random forest algorithm is used for classification and regression tasks. The process involves building multiple decision trees during training. For classification, it selects the most commonly chosen class among the trees, while for regression, it takes the mean of individual tree predictions. The random forest algorithm is a supervised learning algorithm, which means it is trained on a dataset of labeled data. The algorithm works by creating a large number of decision trees, each of which is trained on a different bootstrap sample of the training data. At each node of a decision tree, a random subset of features is considered for splitting the data. This helps reduce correlation and overfitting risks (Schonlau & Zou, 2020).

3.1.4 Multilayer Perceptron (MLP)

A multilayer perceptron is a type of artificial neural network that is composed of multiple layers of interconnected neurons. Each neuron in a layer is connected to all of the neurons in the next layer The MLP is a feedforward network, meaning that the information flows in one direction, from the input layer to the output layer. MLP excels in classification, regression, and clustering tasks. It is trained with backpropagation, iteratively adjusting neuron connection weights to minimize errors. Its simplicity and efficiency make it a popular choice for machine learning, even with large datasets (Taud & Mas, 2018).

3.1.5 Support Vector Machine (SVM)

A support vector machine is a supervised machine learning technique applicable to both classification and regression tasks. Its core principle involves identifying the optimal hyperplane that effectively segregates data points belonging to two distinct classes. In this context, by hyperplane one means a shifted subspace of codimension one in high-dimensional space, a concept that generalizes the notion of plane in three-dimensional space. The SVM algorithm endeavors to locate the hyperplane with the maximum margin, which refers to the distance between the hyperplane and the nearest data points from each class. These nearest data points are commonly referred to as support vectors. When data consists of more than two classes, in a one-versus-all setting, a collection of individual SVMs is built, each separating one class from the union of the remaining classes (Namgyal, 2021).

3.2 Application of ANOVA in Feature Selection

3.2.1 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to analyze and compare the means of two or more groups or treatments to determine if there are statistically significant differences among them. Instead of comparing groups one pair at a time (which can lead to increased chances of Type I errors), it allows for a more comprehensive analysis. ANOVA assumes that the data within each group or treatment are normally distributed and have equal variances. It also assumes that observations are independent. ANOVA decomposes the total variability in the data into two components: variation between groups (explained variance) and variation within groups (unexplained or residual variance). It then assesses whether the explained variance is significantly larger than the residual variance. If ANOVA indicates that there are significant differences among groups, post-hoc tests can be conducted to identify which specific group(s) differ from each other. ANOVA is a widely used method in various fields, including experimental research, social sciences, and manufacturing (Paolella, 2018).

3.2.2 Application of ANOVA in Feature Selection

Feature selection plays a crucial role in the realms of machine learning and data analysis, with its primary objective being the identification and retention of the most pertinent and informative features, while simultaneously reducing or eliminating the influence of irrelevant or redundant ones. ANOVA serves as a valuable statistical technique for feature selection, enabling the assessment of the individual features or variables' significance in predicting a given outcome or dependent variable. It is worth noting that ANOVA-based feature selection is most applicable when dealing with scenarios involving continuous target variables and either categorical or numerical features. However, it can also be used given discrete target variables given numerical or continuous data. ANOVA is versatile and finds utility in various facets of feature selection, including filter-based feature selection, multivariate analysis, feature ranking, feature subset selection, enhancement of model performance, and the facilitation of a clear understanding of how features contribute significantly to the final outcome (Tripathy & Sharaff, 2023).

3.3 Application of PCA in Feature Selection

3.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used in data analysis and machine learning. It aims to transform a high-dimensional dataset into a lower-dimensional one while retaining the most important information. PCA accomplishes this by identifying the principal components (PCs), linear combinations of the original variables, in such a way that the first PC accounts for the highest variance in the data, the second PC for the second-highest, and so forth. Importantly, these PCs are orthogonal to each other, indicating that they are uncorrelated. This orthogonality property ensures that each PC captures unique information, effectively reducing redundancy in the data. PCA relies on the calculation of eigenvalues and eigenvectors derived from the covariance matrix of the original data. Eigenvalues quantify the variance explained by each PC, while eigenvectors determine the direction of each PC within the original feature space (Gray, 2017).

3.3.2 Application of PCA in Feature Synthesis

PCA can be employed in feature synthesis in multiple ways to decrease data dimensionality and elevate the quality of features utilized for modeling and analysis. It assists in the ranking of features according to their contributions to the primary principal components. PCA can also be harnessed to construct fresh features that encapsulate a blend of information from the original features. These novel features represent linear combinations of the initial variables and may occasionally offer superior informativeness or lower noise levels compared to individual features. This proves particularly advantageous when confronted with high-dimensional datasets, as it can enhance computational efficiency and mitigate the risk of overfitting in machine learning models (Mohammed et al., 2016). Furthermore, PCA does not require knowledge of the target variable, and may thus be considered an unsupervised technique.

4. Research Methodology

4.1 Research Hypotheses

PCA is more effective than ANOVA in performing feature reduction for classification by machine learning with datasets of various sizes.

4.2 Research Framework

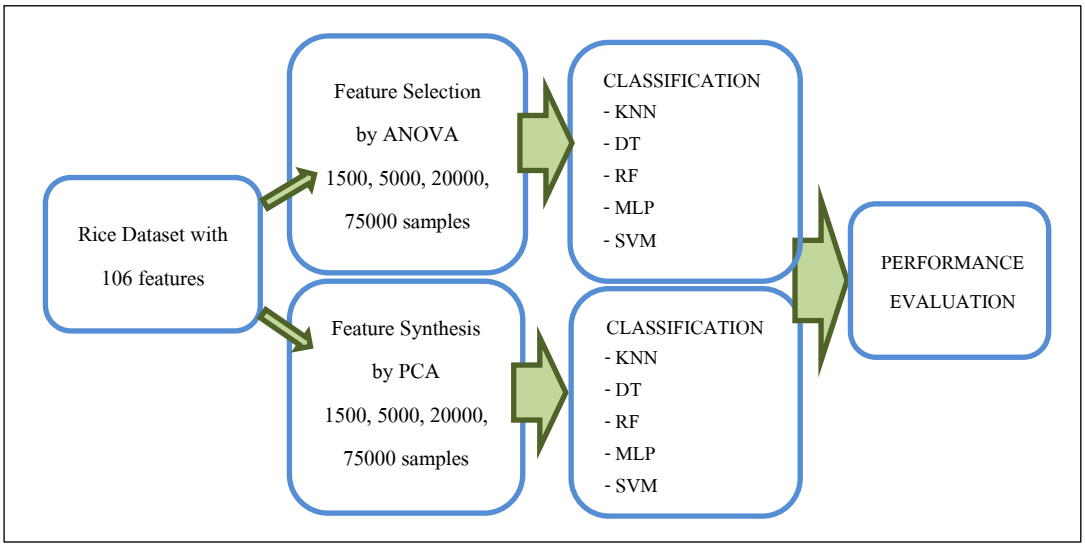


Figure 1 Research Framework

4.3 Research Design

In this research, ANOVA and PCA are utilized for selecting and synthesizing features from 106 distinct attributes that represent various aspects of shape and color of rice from five different rice varieties: Arborio, Basmati, Ipsala, Jasmine, and Karacadag, as shown in Figure 2. In order to evaluate the performance on smaller sized datasets, sub-datasets of 1,500, 5,000 and 20,000 samples each were extracted from the original 75,000-sample dataset by random, in a manner to still represent each variety equally to obtain balanced data. The feature-reduced datasets obtained by ANOVA and PCA were then employed with five different machine learning techniques: K-Nearest Neighbors (KNN), Decision Tree (Dtree), Random Forest (Rforest), Multilayer Perceptron

(MLP), and Multiclass Support Vector Machine (SVM). The accuracies of these models were compared and analyzed using 10-fold cross validation.

4.4 The Data and Tools Used in the Research

The rice dataset is sourced from https://www.muratkoklu.com/datasets/, provided by Köklü (2021). There are 75,000 data samples of 106 distinct features of rice grains. Feature synthesis and the classification by

(a) (b) (c) (d) (e)

Figure 2 Images of rice grains (a) Arborio (b) Basmati (c) Ipsala (d) Jasmine (e) Karacadag

the five machine learning algorithms were done in the Python programming language, using the Scikit-learn version 1.2.1 Python Libraries together with the Intel(R) Extension for Scikit-learn. The most important parameters used in each machine learning model are listed in Table 1, all other parameters remain at their default values. The various programs were running under the Linux operating system on an AMD Ryzen 4500U CPU and 40GB of RAM.

Table 1 Parameters used in each machine learning model

KNN	Dtree	Rforest	MLP	SVM
K=5 Euclidean distance	Criterion="gini"	n_estimators=100 Criterion="gini"	hidden_layer_sizes=(50,1) maxiter=200	Linear Kernel Penalty C=1

5. Research Findings

Figure 3 shows the accuracy results from 10-fold cross validation as a function of the number of features used. Results are displayed for each of the four dataset sizes. The graphs in the left column reflect the results of ANOVA feature selection, choosing from one up to 40 of the top-ranked features. The graphs in the right column show the results from PCA feature synthesis, selecting from 1 up to 20 of the highest-ranked features.

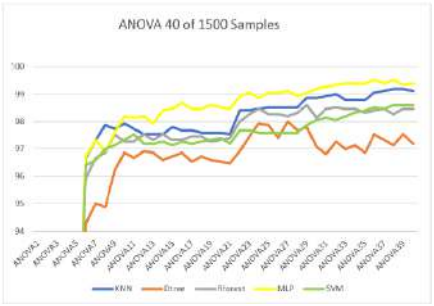
Figure 4 shows performance comparisons of ANOVA versus PCA feature selection for each of the machine learning models individually. The dataset consisting of 1500 samples was chosen here, as its small size serves to highlight the phenomena to be observed.

Finally, Table 2 lists the results for the smallest (1,500 samples) and the largest (75,000 samples) data set numerically, to allow for a more detailed analysis.

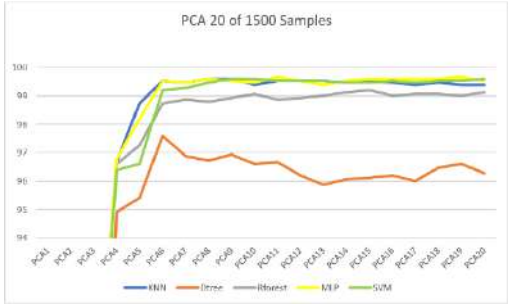
6. Discussion

The results show that except for the decision tree model, using only the first six PCA-generated features already yields accuracies close to the highest accuracies achievable. When using ANOVA feature selection, on the other hand, accuracies are initially noticeably lower as compared to PCA selection. As the number of ANOVA features increases, the accuracies increase only slowly, to eventually surpass the accuracies of PCA selection.

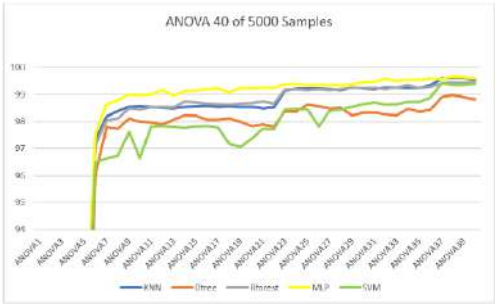
The decision tree model is an exception to this. Figure 4(b) shows that an increase of the number of PCA components beyond ten leads to a noticeable drop in accuracy for the 1500-sample dataset. Such a drop can also be observed with the larger datasets, albeit to a lesser degree. It should be noted that the training of a decision tree already involves the ranking of the features, and PCA synthesis may interfere with this ranking process.



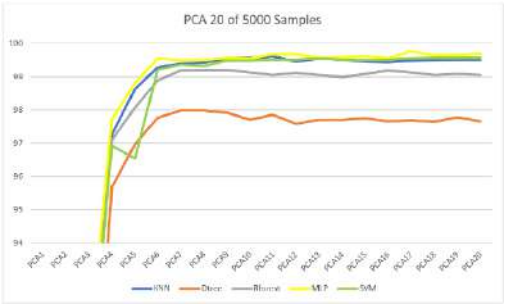
(a)



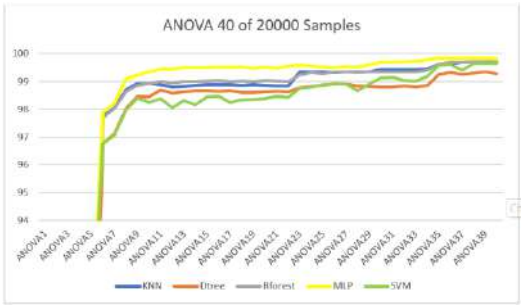
(b)



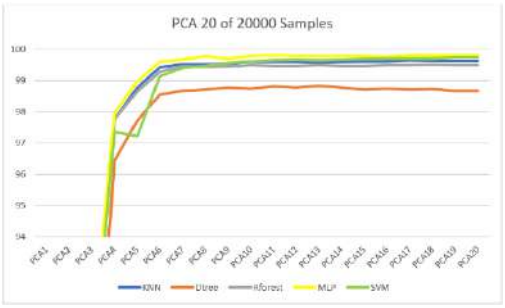
(c)



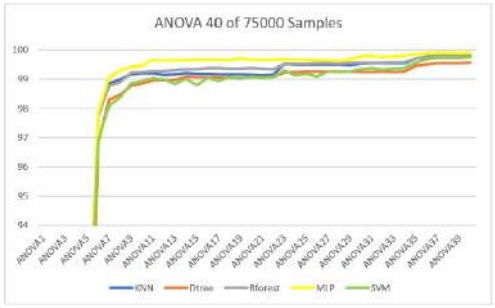
(d)



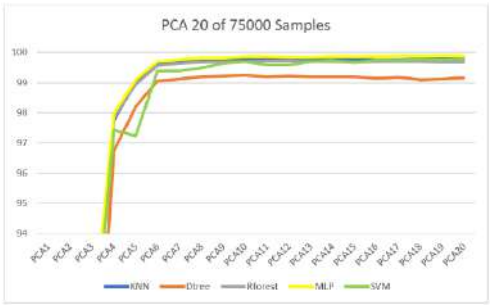
(e)



(f)



(f)



(g)

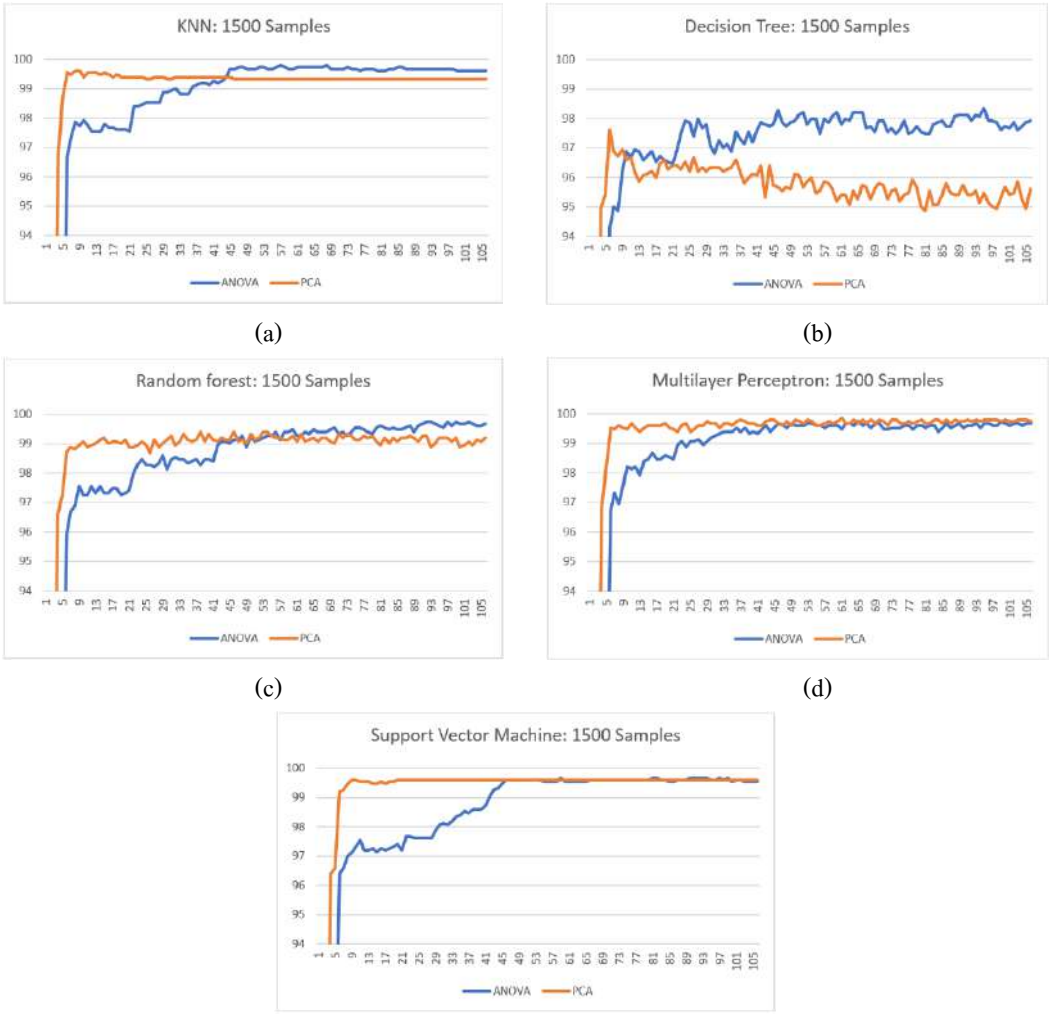
Figure 3 Performance evaluation of the various classification models.

(a) ANOVA, 1-40 top-ranked features, 1,500 samples (b) PCA, first 1-20 generated features, 1,500 samples

(c) ANOVA,1-40 top-ranked features, 5,000 samples (d) PCA, first 1-20 generated features, 5,000 samples

(e) ANOVA, 1-40 top-ranked features, 20,000 samples (f) PCA first 1-20 generated features, 20,000 samples

(g) ANOVA, 1-40 top-ranked features, 75,000 samples (h) PCA, first 1-20 generated features, 75,000 samples



(a) (b)

(c) (d)

(e)

Figure 4 Classification performances of ANOVA and PCA (1,500 samples) by machine learning algorithm.
(a) K-Nearest Neighbor (b) Decision Tree (c) Random Forest
(d) Multilayer Perceptron (e) Support Vector Machine

Mazlan *et al.* (2020) have performed feature reduction on geomagnetic data and found by ANOVA that choosing two of nine features was sufficient. On the other hand, in case of PCA, four of the nine generated features should be used. However, the authors used statistical techniques for feature selection and did not evaluate their selection results through training of a machine learning model.

In Figure 4, a slight drop in accuracy for the 1500-sample dataset can be observed with the KNN model when the number of PCA components moves beyond 9. In fact, inspection of Table 2 shows that at only 8 or 9 PCA components, accuracy is indeed higher than when all 106 features are used. This means that PCA feature synthesis may at times help increase accuracies, a phenomenon which has also been observed in (Reddy *et. al*., 2020) with different datasets.

Table 2 Accuracies by 10-fold cross validation.

1,500-sample dataset (left) and 75,000-sample dataset (right)

1,500 samples						75,000 samples					
No of features	**KNN**	**DT**	**RF**	**MLP**	**SVM**	**No of features**	**KNN**	**DT**	**RF**	**MLP**	**SVM**
ANOVA1	86.00±2.6	83.20±3.11	86.33±2.46	87.27±2.39	86.47±2.94	ANOVA1	86.69±0.35	87.51±0.32	88.20±0.3	88.14±0.33	88.11±0.46
PCA1	51.73±4.26	45.67±4.79	50.87±4.08	54.07±4.38	53.60±3.95	PCA1	49.90±0.38	46.90±0.35	55.26±0.54	55.39±0.35	55.18±0.73
ANOVA2	87.47±2.56	82.67±2.92	87.13±3.14	89.20±2.1	88.47±2.17	ANOVA2	87.72±0.27	83.61±0.32	88.03±0.29	88.91±0.26	88.82±0.33
PCA2	78.87±2.09	72.67±2.55	78.13±3.17	79.87±2.79	79.2±2.45	PCA2	79.31±0.51	73.95±0.55	78.21±0.57	81.86±0.51	81.20±0.56
ANOVA3	87.20±2.84	82.33±3.3	86.60±2.56	89.33±2.75	88.87±2.89	ANOVA3	87.61±0.25	83.68±0.23	87.81±0.2	88.92±0.2	88.89±0.33
PCA3	86.40±2.86	81.8±2.97	86.40±2.95	86.40±2.31	86.07±1.92	PCA3	89.45±0.23	86.28±0.34	89.95±0.21	90.56±0.29	89.59±0.33
ANOVA4	87.33±2.86	82.80±4.09	87.13±2.63	89.00±2.72	88.93±2.55	ANOVA4	87.94±0.25	84.16±0.37	87.86±0.35	89.05±0.27	88.86±0.29
PCA4	96.73±1.75	94.93±2.19	96.60±1.35	96.80±0.78	96.40±1.24	PCA4	97.75±0.17	96.75±0.14	97.86±0.18	97.99±0.15	97.43±0.16
ANOVA5	87.60±2.31	82.73±3.46	87.13±2.37	88.4±2.53	87.60±2.59	ANOVA5	88.07±0.28	84.17±0.36	87.98±0.4	89.17±0.3	88.83±0.39
PCA5	98.73±0.81	95.40±1.62	97.27±1.38	98.20±1.16	96.60±1.28	PCA5	99.02±0.1	98.21±0.2	98.94±0.12	99.1±0.12	97.23±0.41
ANOVA6	96.67±1.23	94.27±1.77	95.93±1.41	96.73±1.53	96.40±1.31	ANOVA6	97.75±0.12	96.83±0.15	97.75±0.16	97.87±0.15	96.89±0.15
PCA6	99.53±0.6	97.60±1.0	98.73±1.28	99.53±0.67	99.20±0.58	PCA6	99.61±0.06	99.06±0.13	99.57±0.04	99.71±0.06	99.39±0.12
ANOVA7	97.33±1.55	95.0±1.53	96.67±1.12	97.33±1.37	96.60±1.35	ANOVA7	98.85±0.11	98.29±0.16	98.77±0.15	99.07±0.14	98.10±0.25
PCA7	99.47±0.58	96.87±1.74	98.87±1.12	99.47±0.5	99.27±0.63	PCA7	99.71±0.07	99.11±0.11	99.65±0.07	99.77±0.07	99.39±0.13
ANOVA8	97.87±1.29	94.87±1.49	96.87±1.52	96.93±1.64	97.00±1.44	ANOVA8	98.99±0.09	98.47±0.16	98.90±0.12	99.30±0.11	98.37±0.18
PCA8	99.60±0.61	96.73±1.59	98.80±1.39	99.60±0.33	99.47±0.58	PCA8	99.75±0.06	99.19±0.08	99.68±0.08	99.83±0.05	99.49±0.34
ANOVA9	97.73±1.24	96.20±1.27	97.53±1.16	97.60±1.2	97.13±1.03	ANOVA9	99.17±0.12	98.78±0.12	99.23±0.1	99.43±0.1	98.85±0.13
PCA9	99.60±0.61	96.93±1.67	98.93±1.08	99.53±0.43	99.60±0.53	PCA9	99.77±0.06	99.22±0.07	99.7±0.07	99.82±0.05	99.65±0.08
ANOVA10	97.93±1.09	96.87±1.46	97.27±1.25	98.20±0.95	97.33±1.12	ANOVA10	99.21±0.11	98.85±0.14	99.25±0.11	99.47±0.06	98.94±0.13
PCA10	99.40±0.63	96.60±1.72	99.07±1.4	99.47±0.5	99.60±0.53	PCA10	99.77±0.06	99.23±0.07	99.70±0.06	99.86±0.07	99.7±0.07
ANOVA15	97.80±1.19	96.73±1.62	97.33±1.15	98.47±1.19	97.13±1.37	ANOVA15	99.19±0.14	99.07±0.09	99.35±0.1	99.66±0.1	98.8±0.47
PCA15	99.53±0.6	96.13±1.76	99.20±1.02	99.60±0.33	99.47±0.4	PCA15	99.77±0.06	99.19±0.1	99.69±0.06	99.86±0.05	99.66±0.31
ANOVA20	97.60±1.31	96.53±2.04	97.33±1.26	98.53±0.83	97.40±1.09	ANOVA20	99.15±0.12	99.07±0.09	99.39±0.12	99.66±0.13	99.09±0.11
PCA20	99.40±0.63	96.27±2.09	99.13±0.95	99.53±0.79	99.60±0.44	PCA20	99.78±0.07	99.16±0.11	99.69±0.07	99.88±0.03	99.79±0.05
ANOVA30	98.87±0.85	97.07±0.8	98.13±0.78	99.20±0.93	98.07±0.96	ANOVA30	99.53±0.08	99.24±0.11	99.57±0.11	99.77±0.08	99.35±0.12
PCA30	99.33±0.67	96.33±1.61	99.13±0.9	99.67±0.45	99.60±0.44	PCA30	99.78±0.08	99.13±0.06	99.70±0.07	99.90±0.03	99.82±0.06
ANOVA40	99.13±0.73	97.20±1.26	98.47±0.67	99.40±0.47	98.60±0.76	ANOVA40	99.81±0.06	99.56±0.11	99.81±0.07	99.89±0.05	99.75±0.08
PCA40	99.40±0.63	96.13±1.76	99.33±0.79	99.67±0.45	99.60±0.44	PCA40	99.79±0.07	99.11±0.06	99.74±0.06	99.91±0.04	99.83±0.06
All 106	99.33±0.67	97.73±0.95	99.73±0.61	99.67±0.33	99.60±0.44	All 106	99.79±0.07	99.61±0.1	99.87±0.04	99.90±0.05	99.83±0.06

Overall, listed in order of increasing accuracies, the K-nearest neighbour, the support vector machine and the multilayer perceptron models show best performance when employing PCA feature reduction. In case of the K-NN and SVM methods and the 1500-sample dataset, highest accuracy is achieved with only 9 PCA components. This points to the feasibility of PCA dimensionality reduction for smaller-sized datasets.

7. Suggestions

This study has shown that PCA feature synthesis can be successfully employed for feature reduction given a dataset with about one hundred features. However, many applications of machine learning to biology and

medicine deal with huge datasets containing thousands of features, and it would thus be informative to evaluate how ANOVA and PCA methods can deal with such types of datasets.

The graphs in figures 4 (a), (c) and (e) show that in the range of 9-21 ANOVA components, accuracies are essentially unchanged. While ANOVA ranks features in terms of their significance for classification, it does not exhibit relationships among the features themselves. Thus, it may be worthwhile to study correlation among the various features to allow for additional feature reduction.

8. Acknowledgements

The authors wish to thank the School of Mathematics and Geoinformatics at the Institute of Science, Suranaree University of Technology, Thailand, for its support. K. Khongjeen graciously acknowledges financial support through the Development and Promotion of Science and Technology Talents Project (DPST scholarship).

9. References

Gray, V. (2017). *Principal Component Analysis: Methods, Applications, and Technology*. Nova Science Publishers, Incorporated.

Köklü, M., Cinar, I. & Taspinar, Y.S. (2021). Classification of rice varieties with deep learning methods. *Computers and Electronics in Agriculture,* 187, 106285.

Mazlan, A.D.N., Hairuddin, M.A., Tahir, N., Khirul Ashar, N.D., & Jusoh, M.H. (2020). Comparative analysis of PCA and ANOVA for assessing the subset feature selection of the geomagnetic disturbance storm time. *Journal of Electrical and Electronic Systems Research,*17. https://doi.org/10.24191/jeesr.v17i1.002

Mohammed, S.B., Khalid, A., Osman, S.F.E., & Helali, R. G. M. (2016). Usage of principal component analysis (PCA) in AI applications. *International Journal of Engineering Research & Technology (IJERT)*, 5(12), 372-375. ISSN: 2278-0181.

Namgyal, J. (2021). *Multiclass Support Vector Machines for Diabetic Retinopathy Diagnosis* [Master's thesis, Suranaree University of Technology].

Paolella, M. S. (2018). *Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.

Patel, H., & Prajapati, P. (2018). Study and analysis of decision tree-based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6, 74-78. https://doi.org/10.26438/ijcse/v6i10.7478.

Patel, H. (2021, August 30). *What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. Towards Data Science*. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Reddy G.T., Reddy M.P.K., Lakshamanna, K., Kaluri, G, & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access,* 8, 54776-54788.

Sun, J., Du, W., & Shi, N. (2018). A Survey of the k-Nearest Neighbors (kNN) Algorithm. *Information Engineering and Applied Computing*, 1, 10.18063/ieac.v1i1.770.

Schonlau, M., & Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3–29. https://doi.org/10.1177/1536867X20909688

Taud, H., & Mas, J. (2018). Multilayer Perceptron (MLP). In M. Camacho Olmedo, M. Paegelow, J. F. Mas, & F. Escobar (Eds.), *Geomatic Approaches for Modeling Land Change Scenarios* (pp. 27-32). Springer.

Tripathy, G., & Sharaff, A. (2023). AEGA: Enhanced feature selection based on ANOVA and extended genetic algorithm for online customer review analysis. *Journal of Supercomputing*, 79, 13180–13209.