

---

# Exploring the Role of Dimensionality Reduction in Enhancing Machine Learning Algorithm Performance

## Abstract

1 In this study, we delve into the pivotal role of dimension reduction techniques  
2 in influencing the performance of machine learning algorithms for heart disease  
3 prediction. Through a comprehensive exploration of a dataset encompassing crucial  
4 features such as age, sex, chest pain type, blood pressure, cholesterol levels,  
5 and more, we investigate the impact of different techniques—namely, Principal  
6 Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), and  
7 Linear Discriminant Analysis (LDA) on classification algorithm effectiveness. The  
8 classification algorithms considered were Logistic Regression, Support Vector  
9 Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes, and Deep Neural  
10 Network (DNN). We used K-fold cross validation to train and validate the classification  
11 algorithms. The performance of these algorithms was assessed using a range  
12 of key metrics including accuracy, F1-score, precision, recall, and specificity. The  
13 results reveals that Linear Discriminant Analysis consistently emerged as a potent  
14 method, remarkably enhancing algorithm performance across all assessed metrics.  
15 We also identified Naive Bayes and Logistic Regression as standout algorithms,  
16 demonstrating remarkable resilience and reliability across diverse scenarios. These  
17 findings collectively shed light on the intricate interplay between dimension reduction  
18 techniques and algorithm selection, offering critical insights for crafting more  
19 accurate and robust strategies in the prediction of heart disease.

## 20 1 Introduction

21 Machine learning represents a dynamic field of computational methods created to simulate human  
22 intelligence through learning from the surrounding context. Machine learning techniques have  
23 demonstrated their effectiveness across a wide array of domains, encompassing pattern recognition,  
24 computer vision, aerospace engineering, finance, entertainment, computational biology, as well  
25 as applications within the realms of biomedical and medical fields [1]. In the rapidly evolving  
26 landscape of machine learning, the explosion of data availability has brought both opportunities and  
27 challenges. While vast datasets hold the potential to unveil hidden insights and patterns, they also  
28 introduce complexities that can hinder the performance of machine learning algorithms. One critical  
29 challenge is the curse of dimensionality, where high-dimensional data spaces can lead to increased  
30 computational demands, overfitting, and reduced generalization ability [2].

31  
32 The "curse of dimensionality" has garnered significant attention in basic research due to its  
33 implications for increased data storage and computing costs [2]. Dimensionality reduction refers to  
34 the process of transforming high-dimensional data into a lower-dimensional representation while

35 preserving essential characteristics. By reducing the number of features or variables, dimensionality  
36 reduction methods aim to simplify the data, improve computational efficiency, and enhance the  
37 interpretability of models [3, 4]. These techniques encompass both feature selection, which identifies  
38 the most informative attributes, and feature extraction, which constructs new features that capture the  
39 essence of the original data [5]. This field is particularly challenging and has become a focal point for  
40 scholars due to its complexity. Finding effective ways to reduce feature dimensions while preserving  
41 essential information has become a hot and difficult area of research within these domains [6]. Over  
42 the years, various techniques have been proposed and studied to enhance the performance of machine  
43 learning algorithms by reducing the data's dimensionality while preserving essential information  
44 [7]. The utilization of Machine Learning classifier models in the medical sector is steadily increasing  
45 [8]. These models have demonstrated significant utility in effectively diagnosing diverse medical and  
46 clinical datasets [9].

47

48 The authors in [10] investigated impact of dimensionality reduction techniques on machine learning  
49 models for cancer prediction using gene expression data was explored. Principal Component  
50 Analysis (PCA), PCA with a kernel, and autoencoder were employed to reduce RNA sequencing  
51 data's dimensionality. Neural network and support vector machine classifiers were trained and tested  
52 using original, dimensionally reduced, and cancer-relevant data. The results demonstrated that  
53 dimensionality reduction enhances classifier performance, with the autoencoder outperforming PCA  
54 and PCA with a kernel. This study highlights the potential of dimensionality reduction in improving  
55 machine learning models on high-dimensional data in cancer research .

56

57 The authors in [11] investigated the potential of machine learning dimensionality reduction methods,  
58 including principal component analysis (PCA), kernel PCA (KPCA) with polynomial kernel function,  
59 latent semantic analysis (LSA), Gaussian random projection (GRP), sparse random projection (SRP),  
60 multidimensional scaling (MDS), Isomap, and locally linear embedding (LLE), to enhance risk  
61 stratification models for chest pain patients in the emergency department (ED). The data of 795  
62 patients presenting with chest pain at Singapore General Hospital between September 2010 and July  
63 2015 were analyzed. These methods were used in combination with logistic regression to create  
64 prediction models. The multidimensional scaling algorithm demonstrated the best performance  
65 with an AUC of 0.901. While the models outperformed existing clinical scores in ROC analysis,  
66 the improvement in predicting 30-day major adverse cardiac events (MACE) was only marginal.  
67 Moreover, the black box nature of these models made them challenging to interpret in clinical  
68 practice. Further investigation is needed to explore their practical clinical implementation.

69

70 The authors in [12] focused on a study aimed to enhance the prediction of Diabetic Retinopathy, a  
71 significant cause of global vision loss, by employing machine learning techniques. It undertook a  
72 comprehensive approach by addressing data preprocessing, dimensionality reduction, and classifier  
73 selection. The researchers collected a Diabetic Retinopathy dataset from the UCI repository and  
74 initially normalized it using the StandardScalar technique. Principal Component Analysis (PCA) was  
75 then applied to extract essential features, followed by the implementation of the Firefly algorithm  
76 for further dimensionality reduction. Subsequently, a Deep Neural Network Model was utilized for  
77 disease classification. This approach sought to improve prediction accuracy, accounting for often  
78 overlooked data preprocessing and dimensionality reduction aspects. However, the study's scope  
79 might be limited by the specific dataset employed, potentially affecting generalizability, and the  
80 performance of the Firefly algorithm and the chosen classifier may vary across diverse datasets or  
81 scenarios.

82

83 Machine learning algorithms play a vital role in diverse fields by enabling predictions and pattern  
84 discovery from vast datasets. However, the curse of dimensionality presents challenges such as  
85 increased computational complexity and potential overfitting. Dimensionality reduction techniques  
86 have emerged as effective tools to mitigate these issues. This research aims to investigate the role of  
87 dimensionality reduction in enhancing the performance of machine learning algorithms using various  
88 dimension reduction methods such as PCA, Kernel PCA and LDA. This proposed project will deploy  
89 popular machine learning classification models for the original data and reduced data and compare  
90 the perform of different machine algorithms developed from each of the data. The data used will  
91 involve a discrete response variable (whether an individual has heart disease or not) for the purpose

92 of supervised learning and 13 features

93

94 The main objective of this research paper is to investigate and assess the impact of dimensionality  
95 reduction techniques on the performance of machine learning classification algorithms for heart  
96 disease. In pursuit of this overarching goal, the study encompasses several specific objectives.  
97 These objectives include conducting dimensionality reduction on heart disease classification data  
98 using techniques like PCA, kernel PCA, and LDA. Additionally, the study involves the training  
99 and validation of machine learning classification algorithms on both the original and reduced  
100 datasets. Furthermore, the evaluation of machine learning classification algorithm performance,  
101 using metrics such as precision, recall, F1-score, sensitivity, specificity, and accuracy, on both  
102 the original and reduced data, is a critical aspect of this research. Ultimately, the study aims to  
103 demonstrate that the employed dimensionality reduction techniques do not significantly degrade the  
104 performance of machine learning classification algorithms in the context of heart disease classification.  
105

106 In this work we will delve into the pivotal role of dimensionality reduction techniques in enhancing  
107 the performance of machine learning algorithms. We will focus on prediction of heart disease from a  
108 selected dataset using various machine learning classifiers such as logistic regression, kernel support  
109 vector machines, Naive Bayes, K nearest neighbours and deep neural networks. By mitigating the  
110 curse of dimensionality, we will explore how these machine learning algorithms deployed to reduced  
111 data will contribute to more accurate predictions, faster training times, and improved generalization  
112 to new data. We will examine a variety of dimensionality reduction methods, ranging from classical  
113 linear techniques such as Principal Component Analysis (PCA), kernel PCA and linear discriminant  
114 analysis.  
115

## 116 2 Data and Methods

### 117 2.1 Data

118 The dataset used in this study was obtained from Kaggle and is available at [13]. The dataset consists  
119 of 14 features, encompassing patient attributes such as age, sex, blood pressure, cholesterol levels,  
120 and exercise test results, including variables like chest pain type, Electrocardiogram results, and  
121 thallium stress test outcomes. The response variable, 'Heart Disease,' indicates the presence (1)  
122 or absence (0) of heart disease. Given the binary classification nature of the problem, supervised  
123 machine learning classification algorithms were used to predict the presence of heart disease."

Table (1) shows the description of the variables.

Table 1: Description of Variables

Variable Name	Description	Type
Age	Age of the patient	Continuous
Sex	Sex of the patient (0 = Female, 1 = Male)	Discrete
Chest pain type	Type of chest pain experienced	Discrete
BP	Blood Pressure	Continuous
Cholesterol	Serum Cholesterol levels	Continuous
FBS over 120	Fasting Blood Sugar > 120 mg/dL (1 = True, 0 = False)	Discrete
EKG results	Electrocardiogram results	Discrete
Max HR	Maximum Heart Rate achieved during exercise	Continuous
Exercise angina	Exercise-induced angina (1 = Yes, 0 = No)	Discrete
ST depression	ST segment depression induced by exercise	Continuous
Slope of ST	Slope of the ST segment during exercise	Discrete
Number of vessels fluro	Number of major vessels colored by fluoroscopy	Discrete
Thallium	Thallium stress test results	Discrete
Heart Disease	Presence of heart disease (1 = Yes, 0 = No)	Discrete

124

125 **2.2 Dimension Reduction Techniques**

126 In this project we explored three dimension reduction techniques. Principal Component Analysis,  
127 Kernel Principal Component Analysis and Linear Discriminant Analysis.

128 **2.3 Supervised Machine Learning Classification Algorithms**

129 In this project logistic regression, K nearest neighbours, support vector machine with radial basis  
130 function, Naive Bayes and deep neural network were trained on both the original and reduced data.

131 **2.4 Performance Evaluation Metrics**

132 We used accuracy, precision, recall, specificity, and F1-score to evaluate the performance of the  
133 classification models. These metrics are defined as follows:

- 134 1. **Accuracy:** Accuracy measures the ratio of correctly predicted instances to the total instances  
135 in the dataset.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

136 If the class label of a record in a dataset is positive, and the classifier predicts the class label  
137 for that record as positive, then it is called a true positive. If the class label of a record in  
138 a dataset is negative, and the classifier predicts the class label for that record as negative,  
139 then it is called a true negative. If the class label of a record in a dataset is positive, but the  
140 classifier predicts the class label for that record as negative, then it is called a false negative.  
141 If the class label of a record in a dataset is negative, but the classifier predicts the class label  
142 for that record as positive, then it is called a false positive.

- 143 2. **Precision:** Precision measures the ratio of true positive predictions to the total predicted  
144 positives. It is a measure of how many of the predicted positive instances are actually  
145 positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- 146 3. **Recall (Sensitivity or True Positive Rate):** Recall measures the ratio of true positive  
147 predictions to the total actual positives. It is a measure of how many of the actual positive  
148 instances were correctly predicted.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- 149 4. **F1 score:** The F1-score is the harmonic mean of precision and recall. It provides a balance  
150 between precision and recall.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- 151 5. **Specificity (True Negative Rate):** Specificity measures the ratio of true negative predictions  
152 to the total actual negatives. It's a measure of how many of the actual negative instances  
153 were correctly predicted.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

154 **2.5 Cross Validation and Model Performance Metrics**

155 In this project, K-fold cross-validation was used in training and validation of classification algorithms.  
156 Cross-validation is a technique crucial in machine learning for assessing a model performance on  
157 unfamiliar data. It involves partitioning the available data into distinct subsets, or folds. One fold  
158 is reserved as a validation set, while the rest serve as training data. This cycle is repeated multiple

159 times, with different folds acting as the validation set each time. The outcomes from these iterations  
 160 are then averaged to yield a more reliable measure of the model performance.  
 161 The core objective of cross-validation is to counteract overfitting, where a model excels on the training  
 162 data but falters on new, unseen data. By evaluating the model on multiple validation sets, cross  
 163 validation provides a more realistic estimate of the model generalization performance, i.e., its ability  
 164 to perform well on new, unseen data. The model performance metrics in this project will be reported  
 165 in terms of the mean validation accuracy, mean validation F1 score, mean validation recall, mean  
 166 validation precision, mean validation specificity and mean validation sensitivity.

### 167 **3 Results and Discussion**

168 In this work, we have explored different dimension reduction techniques and trained various machine  
 169 learning classification algorithms using both the reduced and original data. A train test split with  
 170 0.8/0.2 on the data gave 54 samples of the validation set. Due to relatively small size of the validation  
 171 set which is susceptible to significance variance leading to less reliable validation scores. We adopted  
 172 a k-fold cross validation with k=5 to train and validate machine learning classification algorithms for  
 173 both original and reduced data.

#### 174 **3.0.1 Performance of classifiers with the original data**

175 In this section, we discuss the results of experimentation with the original data using logistic  
 176 regression, K nearest neighbours, support vector machines, Naive Bayes and Deep Neural Networks.  
 177 The table (2) below shows the performance of the algorithms as averaged in K-fold cross validation  
 (mean validation-MV) scores using the training and validation method. From table (2). Naive Bayes

Table 2: Classifier Performance based on mean validation for original data(%)

<b>Algorithm</b>	<b>Accuracy.</b>	<b>F1-score</b>	<b>Precision.</b>	<b>Recall.</b>	<b>Specificity.</b>
Logistic Regression	84.07	81.16	84.80	78.29	88.36
SVM	65.19	53.95	65.56	46.30	80.71
KNN	62.69	50.66	61.14	44.35	78.14
Naive Bayes	85.19	82.78	84.76	81.68	87.52
DNN	76.05	64.57	67.01	63.13	75.07

178 achieved the highest accuracy at 85.19%, closely followed by Logistic Regression at 84.07%. These  
 179 models seem to perform better in overall classification. Naive Bayes has the highest F1-score of  
 180 82.78%, followed by Logistic Regression at 81.16%. This suggests that Naive Bayes and Logistic  
 181 Regression have better trade-offs between precision and recall. Naive Bayes and Logistic Regression  
 182 exhibit high precision values, with Naive Bayes at 84.76% and Logistic Regression at 84.80%.  
 183 Naive Bayes leads in recall with a score of 81.68%, and Logistic Regression follows closely with  
 184 78.29%. Logistic Regression has the highest specificity at 88.36%, implying that it is effective at  
 185 identifying negative cases. Logistic Regression demonstrates balanced performance in terms of  
 186 precision, recall, and specificity. Its accuracy and F1-score are also notably high. This algorithm is  
 187 simple yet effective and can serve as a baseline model for many classification tasks. These results  
 188 suggest that both Naive Bayes and Logistic Regression are strong candidates for further evaluation  
 189 and potential deployment in real-world applications. However, the choice between the two depends  
 190 on the specific requirements of the task and the dataset characteristics. While the SVM, KNN, and  
 191 DNN algorithms show lower performance in comparison to Naive Bayes and Logistic Regression,  
 192 it's worth noting that model performance can be influenced by hyperparameter tuning, dataset size,  
 193 and data preprocessing techniques. Further experimentation and optimization might lead to improved  
 194 results for these models.  
 195

#### 196 **3.0.2 Performance of classifiers with PCA reduced data with 5 features**

197 We performed a PCA on the data and reduced the number of features to 5. Table (3) shows the  
 198 performance metrics of the 5 classification algorithms. In general, the accuracy and F1-scores show  
 199 relatively consistent patterns between PCA-reduced and original data for most algorithms. The  
 200 Logistic Regression algorithm's performance remains stable, with only a slight drop in accuracy  
 201 and F1-score when using PCA-reduced data. Similarly, Naive Bayes retains its high accuracy and

Table 3: Performance Metrics (mean validation) of Classification Algorithms based on PCA reduced data %

Algorithm	Accuracy.	F1-score	Precision.	Recall.	Specificity.
Logistic Regression	82.96	80.09	83.84	77.25	86.83
SVM	84.07	76.21	74.70	78.00	78.80
KNN	81.48	76.83	85.98	70.44	89.77
Naive Bayes	82.22	78.66	82.72	75.54	87.16
DNN	76.48	74.75	74.85	75.20	79.45

202 F1-score. However, SVM, KNN, and DNN experience more noticeable drops in accuracy and  
 203 F1-score when using PCA-reduced data. Precision and recall values also show variations. While  
 204 some algorithms like Logistic Regression and Naive Bayes maintain similar precision and recall  
 205 between the two datasets, others like SVM and KNN exhibit trade-offs. SVM's precision decreases  
 206 while recall increases with PCA-reduced data. On the other hand, KNN's precision improves, but  
 207 recall drops significantly. This suggests that the choice of algorithm can have different impacts when  
 208 using PCA-reduced data. The specificity values for most algorithms remain relatively stable between  
 209 the two datasets. However, KNN notably improves in specificity when using PCA-reduced data,  
 210 suggesting better performance in correctly identifying negative cases. Comparing the results from  
 211 PCA-reduced and original data, it's evident that the impact of dimensionality reduction on algorithm  
 212 performance varies. While some algorithms maintain their performance, others experience changes in  
 213 accuracy, precision, recall, and specificity. The decision to use PCA-reduced data should be carefully  
 214 considered based on the specific algorithm's behavior and the desired trade-offs in performance  
 215 metrics.

### 216 3.0.3 Performance of the Classifiers using Kernel PCA reduced data using 5 Features

217 We performed Kernel PCA and reduced the data to 5 features, the five components were selected based  
 218 on the eigen values of the resulting covariance matrix. We trained and validated the classification  
 algorithms using the kernel reduced data From table (4) we observe that Kernel PCA-reduced data

Table 4: Performance Metrics (mean validation) of Classification Algorithms based on kernel PCA reduced data %

Algorithm	Accuracy.	F1-score	Precision.	Recall.	Specificity.
Logistic Regression	82.96	78.93	83.14	77.33	87.20
SVM	77.78	75.85	74.37	77.66	77.76
KNN	80.37	76.56	79.58	74.06	85.34
Naive Bayes	81.48	78.05	81.45	75.80	85.17
DNN	76.67	74.24	75.84	73.08	80.39

219 shows a slight drop in accuracy and F1-scores across all algorithms compared to the original data.  
 220 Logistic Regression, Naive Bayes, and KNN maintain their accuracy quite well, but SVM and DNN  
 221 exhibit more noticeable decreases. Precision values are generally maintained or slightly reduced with  
 222 Kernel PCA-reduced data. Recall, on the other hand, is somewhat affected, with SVM and KNN  
 223 experiencing drops in recall. This suggests that Kernel PCA might influence recall more than precision  
 224 for certain algorithms. Specificity values mostly remain consistent between the original and Kernel  
 225 PCA dataset, with a slight drop for SVM and DNN in the Kernel PCA-reduced data. Comparing  
 226 the results from Kernel PCA-reduced and original data, we observe that dimensionality reduction  
 227 using Kernel PCA has a varying impact on algorithm performance. While some algorithms maintain  
 228 their performance well, others experience decreases in accuracy, F1-score, and recall. Precision and  
 229 specificity tend to be more stable across the board. The decision to use Kernel PCA should be made  
 230 based on the specific algorithm's behavior and the desired balance between performance metrics.  
 231 Kernel PCA can be effective in capturing complex patterns in the data, but the trade-offs should be  
 232 carefully considered.  
 233

### 234 3.0.4 Performance of classifiers with LDA reduced data with 1 feature

235 In this section, we present the experimental results obtained after training and validation classification  
 236 algorithm using the LDA reduced data and compare the performance of the classification algorithms

on both the LDA reduced data and original data. From table (5) comparing the two sets of results, we

Table 5: Performance Metrics (mean validation) of Classification Algorithms based on LDA reduced data %

Algorithm	Accuracy.	F1-score	Precision.	Recall.	Specificity.
Logistic Regression	85.56	83.58	85.74	81.73	88.25
SVM	84.81	82.96	84.49	81.73	86.92
KNN	85.93	82.85	90.19	77.47	92.61
Naive Bayes	84.81	82.49	85.32	80.06	88.26
DNN	78.09	82.52	85.70	80.17	89.40

237 observe that, LDA-reduced data with one feature generally outperforms the original data in terms  
238 of accuracy and F1-score for all algorithms. This indicates that the single feature extracted through  
239 LDA contains more discriminative information than the original dataset's features. Precision and  
240 recall values also show improvements with LDA-reduced data for most algorithms. This suggests  
241 that the single LDA feature better separates the classes, leading to higher precision and recall rates.  
242 LDA-reduced data maintains or improves specificity values for all algorithms, indicating a better  
243 ability to correctly identify negative cases. Comparing the results from LDA-reduced data and the  
244 original data, it's evident that the LDA transformation to one feature has significantly enhanced the  
245 performance of the classification algorithms. The single feature extracted through LDA captures  
246 meaningful discriminatory information, leading to improved accuracy, F1-score, precision, recall,  
247 and specificity across the board. The success of LDA in improving classification performance  
248 demonstrates its efficacy in feature extraction for dimensionality reduction. This result suggests that  
249 the LDA-reduced feature space is better suited for separating classes compared to the original data's  
250 feature space. In summary, the results strongly indicate the advantages of LDA in enhancing the  
251 discrimination between classes. Leveraging LDA-reduced data can lead to more accurate and reliable  
252 classification models.  
253

254 Furthermore, Linear Discriminant Analysis (LDA) emerges as a powerful tool in enhancing the perfor-  
255 mance of classification algorithms, particularly evident when reducing data to a single feature. LDA  
256 operates on the principle of maximizing class separation, ensuring that the derived feature captures  
257 the most discriminative information between instances with and without heart disease. This singular  
258 focus on relevant information not only simplifies the classification task through dimensionality  
259 reduction but also elevates the quality of the extracted feature.

260 The efficacy of LDA in discarding less relevant features and noise contributes to the enhanced  
261 discrimination observed in the results. This singular feature, carefully crafted by LDA, encapsulates  
262 essential characteristics of the data pertinent to the classification problem. Beyond its role in feature  
263 extraction, LDA promotes improved generalization by emphasizing the most relevant information.  
264 This characteristic is vital for models to perform well on new, unseen data, leading to heightened  
265 accuracy, F1-score, precision, recall, and specificity during validation or testing.

266 Moreover, LDA proves effective in handling imbalances between classes, providing a balanced  
267 and informative representation. Its supervised dimensionality reduction design specifically targets  
268 differences between classes, aligning with the inherent nature of binary classification tasks like  
269 heart disease prediction. In summary, the success of LDA with one feature lies in its ability to  
270 streamline and refine the dataset, ensuring that the derived feature encapsulates crucial discriminatory  
271 information, thereby enhancing the overall effectiveness of classification algorithms.

## 272 4 Conclusion

273 In this research, we compared the performance of classification algorithms across four different  
274 scenarios: using the original data, data reduced by Principal Component Analysis (PCA), Kernel  
275 Principal Component Analysis (KPCA), and Linear Discriminant Analysis (LDA). We examined the  
276 performance of five algorithms: Logistic Regression, Support Vector Machine (SVM), K-Nearest  
277 Neighbors (KNN), Naive Bayes, and Deep Neural Network (DNN). The original data served as the  
278 baseline for comparison. It demonstrated varying performance across different algorithms. Naive  
279 Bayes showcased strong accuracy, precision, and recall, while SVM and KNN struggled with low  
280 accuracy and F1-scores. PCA-reduced data exhibited consistent patterns across most algorithms.  
281 While accuracy and F1-scores remained relatively stable for Logistic Regression and Naive Bayes,

282 SVM, KNN, and DNN experienced decreases. PCA demonstrated a trade-off between reducing  
283 dimensionality and maintaining algorithm performance. Kernel PCA-reduced data displayed mixed  
284 results. While some algorithms maintained performance, others experienced drops in accuracy,  
285 F1-score, and recall. KPCA's impact on performance was varied and algorithm-specific. LDA-  
286 reduced data consistently outperformed the original data across all algorithms. The single LDA  
287 feature captured valuable class separation information, leading to improvements in accuracy, F1-score,  
288 precision, recall, and specificity. LDA showcased its ability to enhance discrimination between classes.  
289 Across all scenarios, Naive Bayes consistently exhibited strong overall performance. Its balanced  
290 accuracy, precision, and recall made it a dependable choice. Logistic Regression also maintained  
291 competitive performance in most cases, while SVM, KNN, and DNN faced challenges, particularly  
292 in scenarios involving dimensionality reduction. LDA-reduced data consistently yielded the best  
293 results among the dimensionality reduction techniques. Its ability to enhance class separation led to  
294 improved performance in all metrics for all algorithms. Naive Bayes and Logistic Regression stood  
295 out as the best performing algorithms with LDA-reduced data. While challenges exist, these findings  
296 provide valuable insights for selecting suitable algorithms and data transformation techniques based  
297 on the specific characteristics of the problem domain. Additionally, a combination of hyperparameter  
298 tuning, model selection, and feature engineering can further optimize algorithm performance and  
299 facilitate better decision-making in real-world applications.

300 **References**

- 301 [1] Issam El Naqa and Martin J Murphy. *What is machine learning?* Springer, 2015.
- 302 [2] S Vellingiri, SJPCS Alagumuthukrishnan, et al. A review of dimensionality reduction tech-  
303 niques for efficient computation. *Procedia Computer Science*, 165:104–111, 2019.
- 304 [3] Nandakishore Kambhatla and Todd K Leen. Dimension reduction by local principal component  
305 analysis. *Neural computation*, 9(7):1493–1516, 1997.
- 306 [4] Christopher JC Burges et al. Dimension reduction: A guided tour. *Foundations and Trends® in*  
307 *Machine Learning*, 2(4):275–365, 2010.
- 308 [5] Miguel A Carreira-Perpinán. A review of dimension reduction techniques. *Department of*  
309 *Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, 9:1–69, 1997.
- 310 [6] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review.  
311 *Complex & Intelligent Systems*, 8(3):2663–2693, 2022.
- 312 [7] Claudio Turchetti and Laura Falaschetti. A manifold learning approach to dimensionality  
313 reduction for modeling data. *Information Sciences*, 491:16–29, 2019.
- 314 [8] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on*  
315 *machine learning*, pages 63–71. Springer, 2003.
- 316 [9] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on*  
317 *multiple classifier systems*, pages 1–15. Springer, 2000.
- 318 [10] Md Faisal Kabir, Tianjie Chen, and Simone A Ludwig. A performance analysis of dimensionality  
319 reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*,  
320 3:100125, 2023.
- 321 [11] Nan Liu, Marcel Lucas Chee, Zhi Xiong Koh, Su Li Leow, Andrew Fu Wah Ho, Dagang  
322 Guo, and Marcus Eng Hock Ong. Utilizing machine learning dimensionality reduction for  
323 risk stratification of chest pain patients in the emergency department. *BMC medical research*  
324 *methodology*, 21:1–13, 2021.
- 325 [12] Thippa Reddy Gadekallu, Neelu Khare, Sweta Bhattacharya, Saurabh Singh, Praveen Ku-  
326 mar Reddy Maddikunta, In-Ho Ra, and Mamoun Alazab. Early detection of diabetic retinopathy  
327 using pca-firefly based deep learning model. *Electronics*, 9(2):274, 2020.
- 328 [13] Author/ Organization. Title of webpage. [https://www.kaggle.com/datasets/  
329 utkarshx27/heart-disease-diagnosis-datase](https://www.kaggle.com/datasets/utkarshx27/heart-disease-diagnosis-datase), 2023. 2023.