A Performance Analysis of Dimensionality Reduction Algorithms in Machine Learning Models for Cancer Prediction

Abstract

Developments in technology facilitate the use of machine learning methods in medical fields. In cancer research, the combination of machine learning tools and gene expression data has proven its ability to detect cancer patients. However, processing such high-dimensional and complex data is still a challenge. This paper analyzed the impact different dimensionality reduction techniques have on machine learning models used for cancer prediction. Dimensionality reduction techniques such as principal component analysis (PCA), PCA with a kernel, and autoencoder were utilized to reduce the dimensionality of the RNA sequencing data. Two machine learning classifiers, namely neural network and support vector machine, were trained and tested using the original, dimensionally reduced, and cancer-relevant data. Various metrics, such as accuracy, precision, recall, F-Measure, receiver operating characteristic curve, and area under the curve, were used to assess the performance of classifiers. The results showed that dimensionality reduction positively affects the performance of the classifiers. Additionally, autoencoder performed better than PCA and PCA with a kernal. These findings indicate the potential of dimensionality reduction in improving the analytical results of machine learning classification models on high-dimensional data.

Keywords: Machine Learning, Cancer Prediction, Dimensionality Reduction, RNA-seq Expression, Prostate Cancer

1. Introduction

Prostate cancer is indicated by the National Cancer Institute (NCI) as the most common and second deadliest cancer among men in the United States [1]. According to the American Cancer Society (ACS), 1 in 8 men will be diagnosed with prostate cancer during his entire lifetime, whereas 1 in 41 men will die from

the disease [2]. Generally, cancers in the early stage are easier to treat. Therefore, early detection of prostate cancer is the key to improving patients recovery chances.

The requirement for early diagnosis leads to the use of new technologies such as machine learning (ML) and gene expression data. Comparing to clinical data, gene expression data is different because of its high dimensionality and complexity. Current ML techniques are unable to process such data accurately. In order to obtain promising results, many researchers choose to perform gene selection to reduce the data size before using ML tools. However, gene selection techniques ignore the biological and statistical relevances between genes [3]. Dimensionality reduction is another way to reduce the data size. By creating a representation of the original data in a lower dimension, dimensionality reduction techniques are able to keep the relevances between genes while reducing the dimensionality and complexity of the data. Since the majority of cancer research so far used gene selection, we want to explore the feasibility of dimensionality reduction techniques in cancer research like predicting prostate cancer patients.

In this study, gene expression data of prostate cancer patients, along with the clinical variable "Gleason score," were utilized as predictors, and the sample type (cancer / non-cancer) was used as the target variable. Three dimensionality reduction approaches, representing three types of dimensionality reduction methods, namely linear methods, sublinear methods and nonlinear methods, were combined with two ML classifiers to form six dimensionality reduction-based ML classification techniques. The results using such techniques were compared with ones using the original data, and data with only cancer-relevant genes and the clinical variable.

The contributions of this study are listed as follows:

- More than 20,000 genes were considered in this study.
- Principal component analysis (PCA) and autoencoder were used to study the performance of linear, sublinear and nonlinear dimentionality reduction techniques.
- Neural network (NN) and support vector machine (SVM) were used in this study to examine the performance of both linear and nonlinear ML classification models.

- Hyperparameter searches were conducted to obtain the optimal ML models.
- 10-Fold cross-validation was used during each hyperparameter search to ensure the generalizability of models.
- Prediction models were tested on original, reduced and cancer-relevant data to observe the impact of dimensionality reduction techniques on performance of models.

This paper consists of five sections following the introduction. Related works are discussed in Section 2. Section 3 delineates the methodology, including data characteristics, dimensionality reduction techniques, imbalance treatment, ML techniques for classification, model construction, and performance measures. The environmental setup and results are shown and discussed in Section 4. Section 5 is our discussion section. Finally, Section 6 summarizes this paper, where we conclude our study and suggest potential future research directions.

2. Related Work

Gene expression techniques such as RNA sequencing have become a vital tool in cancer research and therapy. Developments in RNA sequencing technologies significantly improved the effectiveness of RNA-seq data in cancer research, especially in fields such as cancer biomarker identification, cancer evolution analysis, and personalized medicine [4]. In addition, several studies used RNA-seq data for early cancer diagnosis, subtype classification, and cell analysis [5][6][7][8].

Machine learning (ML) allows computers to learn and make predictions for unknown data. ML is usually categorized under data science, which describes "the systematic study of the organization, properties, and data analysis" [9]. ML techniques can be classified into two divisions - supervised and unsupervised. Classification is a supervised learning technique with the goal of predicting a finite set of categorical classes without an explicit order [10]. ML is a powerful tool that allows users to predict future events using data about past or current events and discover relationships between inputs and desired outputs that cannot be detected by humans [11]. A few recognized classification techniques are decision tree, Naive Bayes, neural networkand support vector machine. In recent years, ensemble techniques such as random forest and XGBoost have been proven superior to traditional methods because of their scaling ability and interpretability [10][12][13]. In the medical field, ML has become widely adopted by researchers

as a technique to detect various types of diseases. Techniques such as decision tree, Naive Bayes and random forest were used by researchers to detect heart disease and kidney disease [14][15][16], while support vector machine was used to detect coronary artery disease and cardiovascular disease [17][18]. Researchers also applied ensemble techniques such as bagged decision tree and XGboost to detect diabetes [19][20]. Although neural network is also used to detect diseases, it is usually used on tasks that require imaging data. Researchers applied convolutional neural network, a type of neural network that specializes in image classification, to predict Alzheimer's disease and Parkinson's disease [21][22][23]. These papers utilized various ML techniques to detect different types of diseases accurately. However, unlike other diseases, high-dimensional data like RNA-Seq is widely used in cancer research. Processing such data still poses a challenge to ML techniques. A common way to solve this issue is to utilize data reduction before applying ML.

Data reduction is needed to handle high dimensional data efficiently. Feature selection and dimensionality reduction are two popular data reduction techniques. Feature selection reduces the data size by removing irrelevant features. Generally, feature selection techniques are divided into supervised category and unsupervised category. Supervised data selection such as SHapley Additive exPlanations (SHAP) algorithm use the target variable(s) to remove irrelevant variables. Unsupervised data selection such as univariate feature selection generally rely on statistical methods to remove redundant features. Feature selection techniques are widely used in ML tasks such as classification, regression, and clustering [24]. In ML-related cancer research, feature selection is proven to be able to improve analytical results by removing unnecessary genes [25][26]. However, unsupervised feature selection methods are not reliable because the possibility of removing relevant features or selecting irrelevant features [27]. In contrary, dimensionality reduction techniques compress input data to obtain a representation of it in a lower dimension. A widely used dimensionality reduction technique in cancer research is the principal component analysis (PCA). Authors in [28] and [29] applied PCA to reduce data size and obtained strong analytical results. Another dimensionality reduction technique is the autoencoder. Autoencoder (AE) was utilized to reduce data size and extract functional genes by authors in [30][31][32]. Feature selection seems to be the preferred data reduction choice for most researchers. However, we want to explore how incorporating dimensionality reduction methods would change the performance of ML models in this paper.

3. Materials and Methods

This study examined the impact dimensionality reduction techniques have on prostate cancer prediction. The methodology we used include the following 5 stages: data gathering, data preprocessing and separation, model construction, model training, and model evaluation. Fig. 1 provides an overview of the workflow of the model.

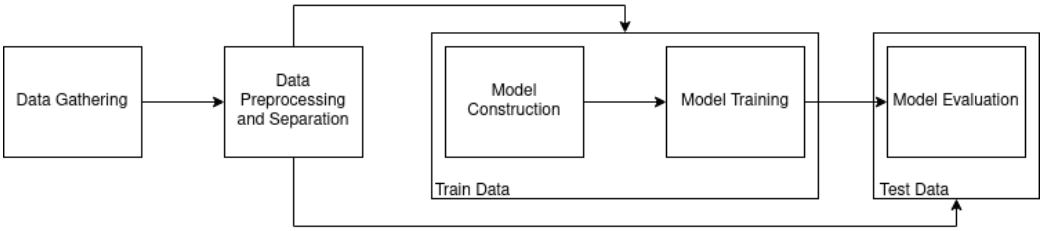


Figure 1: Model Overview.

3.1. Data Description

We obtained RNA-seq data and clinical information of prostate cancer patients from the National Cancer Institute Genomic Data Commons (GDC). $Log_2(x+1)$ normalized illumina Hi-Seq RNA sequencing data was merged with clinical variables retrieved from the GDC. RNA-seq and clinical data were merged together based on their corresponding sample IDs. Samples without RNA-seq information were removed. The merged data set consisted of 550 samples, 497 of which were primary tumor samples (cancer patients), and 52 of which were solid tissue standard samples (non-cancer individuals). The only one metastatic tumor sample was considered as a cancer patient in this study. Thus, cancer samples counted were 498, and non-cancer samples counted were 52. All 20,531 genes and the clinical variable "Gleason score" were assessed as predictors of tissue types (cancer or non-cancer).

3.2. Data Preprocessing

Data preprocessing has been recognized as a fundamental stage of any ML model [33]. Dimensionality reduction, the main focus of this paper, is also considered as a data preprocessing technique. The data preprocessing techniques we implemented will be discussed in this section.

3.2.1. Dimensionality Reduction Techniques

The main focus of this study was the impact dimensionality reduction techniques have on the performance of prostate cancer prediction. Dimensionality reduction is the process of transforming high-dimensional data into a low-dimension space while maintaining the integrity of the original data. The advantages dimensionality reduction offers include performance boost, reduction in computation time, and computational resource requirements. In this study, we chose three different dimensionality reduction techniques representing different approaches. They were PCA, PCA with a kernel function, and autoencoder.

Principal component analysis is a widely used dimensionality reduction technique. PCA reduces the dimension of input data by projecting it onto a smaller space through linear operations including the calculation of the covariance matrix of the given data set, the calculation of the eigenvectors and corresponding eigenvalues from the covarience matrix, and the ranking and selection of eigenvectors based on eigenvalues. The selected eigenvectors are the principal components. The maximum number of principal components cannot exceed the original dimension [34].

However, PCA can only be implemented on continuous data because it relies on the linear relationship between each dimension. This limits its usefulness on high dimensional nominal data or complex data with nominal features. Implementation of kernels such as radial basis function (RBF) were introduced to solve this limitation [35]. In this study, both PCA and RBF-PCA were used to reduce the prostate cancer data set to a lower dimension. The number of principal components was set to 250 as it provided 90% explained variance. Explained variance is a measure that is used to show the discrepancy between a reduced data set and its original data. Higher percentages of explained variance indicates a stronger strength of association between them. It also means that predictions made based on them will be better [36].

Autoencoder is a neural network that is designed to compress and reconstruct input data. Similar to a neural network, an autoencoder also consists of layers of nodes. An autoencoder can be divided into an encoder and a decoder. The encoder takes in an input vector and "press" it to a latent layer in a lower dimension. The decoder tries to reconstruct the original input from the compressed representation produced by the encoder. Different from a neural network, the label of an

autoencoder during its training stage is the input vector itself. Due to this feature, autoencoder is sometimes described as a self-learning algorithm. Autoencoders can extract both linear and nonlinear relationships embedded in the input data itself. The stack-by-stack dimensionality reduction method used by autoencoders also reduces loss of information [32][37].

In this study, the autoencoder used consisted of 5 hidden layers, including the latent layer. The structure of the autoencoder is shown in Fig. 2. The dimension of the latent layer was set to 250, the same as the number of principal components. The loss function we used during the training of the autoencoder was the mean squared error (MSE). After training, the loss of information for the autoencoder was 0.5 in term of MSE. Because the focus of this paper was to study how reducing the dimension of data would impact the performance of cancer prediction, although the autoencoder was trained with both an encoder and a decoder, only the encoder was used after training.

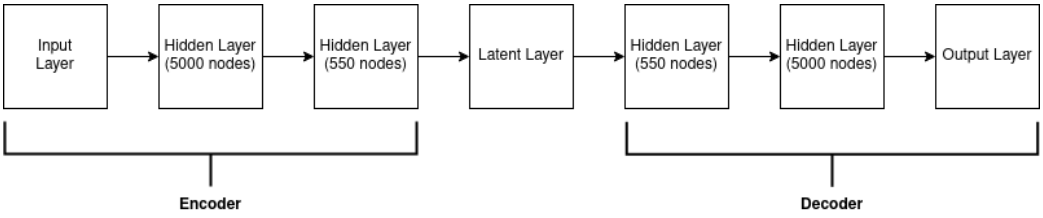


Figure 2: Structure of the autoencoder.

Comparing both techniques, PCA has some advantages over autoencoder in terms of real-world usage. Because autoencoder is a form of neural network, it requires training before applying. This could cause problems to users who don't have access to high performance computers. PCA, on the other hand, can be applied to a given data set without training. This feature allows users with less capable computers to improve the performance of classifiers by reducing the dimension of a given data set. However, the performance difference between nonlinear techniques like autoencoder and linear ones like PCA is currently unclear.

3.2.2. Oversampling

To offset potential bias the model could have toward the majority class, which in this case was the cancer class, the training set of each group was oversampled to artificially increase the number of non-cancer samples [38]. The synthetic minority over-sampling technique (SMOTE) was used in this study. SMOTE randomly

generates new minority samples by creating new samples around existing minority samples. New samples are located between two existing samples to prevent new samples being exact copies of existing samples but also not too different from them [39]. Comparing to other oversampling techniques, SMOTE works better with continuous data [40].

3.3. Classification Techniques

In ML, classification is a typical task and has been widely used in various application domains [41]. The objective of the classification model is to predict qualitative or categorical outputs which assume values in a finite set of classes (e.g., Yes, No or Red, Green, Blue, etc.) without an explicit order [10]. Cancer prediction can be treated as a classification task because it involves classifying patients according to whether they may have cancer or not. In this study, Neural network (NN) and support vector machine (SVM) were used as classification models.

3.3.1. Neural Network

Neural network is a stack of layers of nodes. Each node, apart from ones in the input layer, is connected to all nodes in the previous layer and has a weight associated to it. An activation function is also attached to each node in all layers except the input layer. The activation of a node depends on the result of its activation function, which is calculated using the weighted sum of all nodes in the previous layer. During the training process, the network learns by adjusting nodes weights to correctly predict the label of the input tuples. NN can discover the nonlinear relationships between inputs and outputs. However, NN takes longer to train and is computationally more expensive [42][34].

Being a parametric method, selecting the optimal hyperparameters is required to maximize the performance of NN. Some important hyperparameters include: number of layers, number of nodes in a layer, activation function and learning rate. Sometimes a dropout rate is also attached to each hidden layer. Dropout is the process to increase the generalizability of a network by artificially deactivating some nodes during the training process [43]. Because little is known about how to calculate the optimal values of these hyperparameters, random search was used to find the optimal values. The list of searched hyperparameters and the searched range are shown in Table 1. The optimal values of each hyperparameter are listed in Table 3 in the result section.

Hyperparameter	Search Range
Number of Layers	(Max: 5, Min: 1)
Number of Nodes Per Layer	(Max: $1/2$ of previous layer[a], Min: $1/2$ of Max)
Activation Function	(Sigmoid, RELU, ELU, Swish, Tanh)
Learning Rate	(0.0005, 0.001, 0.0015)
Dropout Rate	(Max: 0.3, Min: 0.1)

Table 1: Searched NN Hyperparameters.

[a]Max in the first layer of NN for original set was set to 1/6 of the previous layer due to limitation on computational resource.

3.3.2. Support Vector Machine

Support vector machine (SVM) is a ML algorithm for both linear and nonlinear data. SVM maps the input data to a higher dimension where linear separation is feasible. Within the new dimension, it searches for the linear optimal hyperplane, a decision boundary that divides samples into two classes using support vectors. SVM is considered highly accurate and less likely to overfit. SVM also can produce a compact description of the learned model. However, SVM suffers from slow training time [34][11].

SVM natively only works with linearly separable data. For nonlinear separable data sets, kernels can be applied to SVM models by creating a feature space that consists of observations of original samples and performing classification in the feature space [44]. SVM is sensitive to kernel choice. Thus, a grid search was implemented to find the best kernel for a given data set. The list of searched kernels is shown in Table 2. The optimal kernels are listed in Table 3 in the result section.

Hyperparameter	Search Range
Kernel	(Linear, Poly, RBF, Sigmoid)

Table 2: Searched SVM Hyperparameters.

3.4. Model Construction

The detailed structure of the model that are shown in Fig. 1 is illustrated in Fig. 3 and Fig. 4. Fig. 3 covers the data preprocessing and separation stage, while Fig. 4 covers the classification stage, which includes model construction, training,

and evaluation stages.

In the data preprocessing stage, the cancer data gathered from the GDC was used to generate the 5 input data sets. One set was not modified, denoted as the original set, while the other four sets were modified using different techniques. An autoencoder was used to dimensionally reduce the original data to 250 dimensions to generate the autoencoder set. The PCA set was generated by applying PCA to the original data. Similarly, the RBF-PCA set was generated by applying PCA with RBF kernel to the original data. Both PCA and RBF-PCA sets contained 250 principal components. The cancer-relevant set, denoted as the CR set, contained only the gleason score and 36 genes that are relevant to prostate cancer [45].

The input data sets were then divided into train sets and test sets through a stratified split to ensure both cancer and non-cancer samples are presented. The ratio between samples in a train set and a test set was 80:20. The train sets were oversampled using SMOTE to artificially increase the number of non-cancer samples before being used in the classification stage. The test sets were not oversampled to simulate real-world data. The train sets and test sets were then sent to the classification stage.

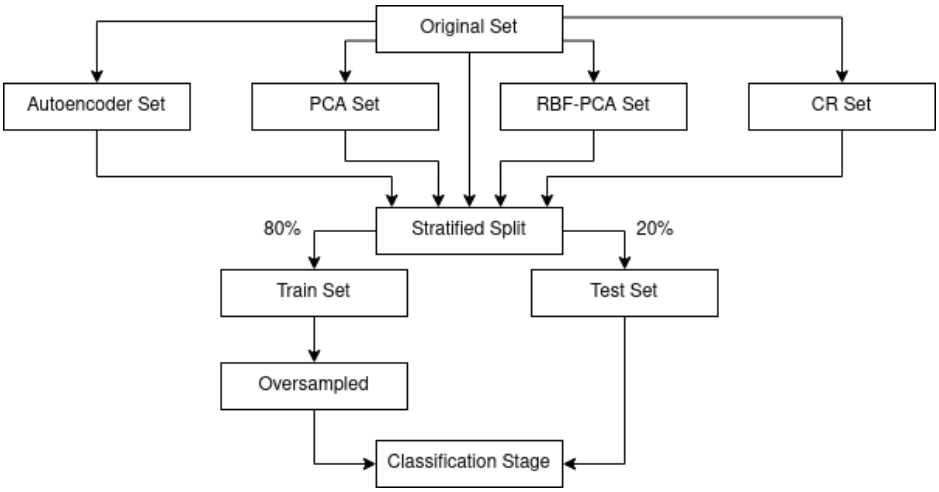


Figure 3: Data Preprocessing Stage.

In the classification stage, the train sets were first used for searching the optimal hyperparameters of both classification techniques. Because we only had around 400 samples in each train set, 10-fold cross-validation was used during

searching to ensure all samples went through the model, thus to offset the model's potential bias towards certain samples caused by using small size data set.

After the optimal hyperparameters were found, they were used to compile the classification models. The train sets were then used to train the compiled models. The performance of the trained models were assessed based on the metrics in Section 3.5 using the test sets. The train-test process was repeated 10 times to ensure the generalizability of the results. After finishing 10 iterations of this process, the mean value of each metric was calculated and exported together as the performance of the model.

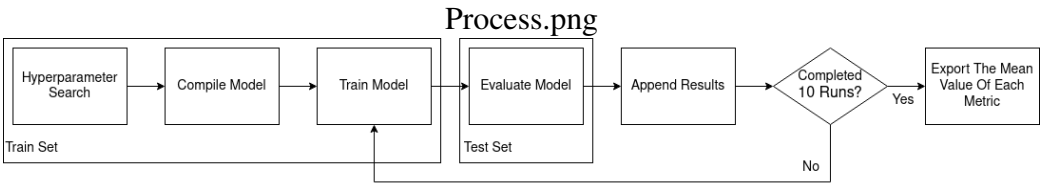


Figure 4: Classification Stage.

3.5. Performance Measures

Several measures were used to evaluate the performance of the classification models, including accuracy, precision, recall, F-Measure, receiver operating characteristic (ROC) curve, and Area Under Curve (AUC).

Accuracy indicates the ratio of correct predictions over all examined samples. Precision measures the ratio of correctly predicted positive samples over all predicted positive samples. Recall measures the ratio of correctly predicted positive samples over all positive samples. F-Measure represents the harmonic mean between precision and recall. ROC curve consists of two axes, one indicates the true positive rate (TPR), the other indicates the false positive rate (FPR). AUC shows the area under the ROC curve. The formulas of the performance measures are shown below.

$$\text{Accuracy: } \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision: } \frac{TP}{TP+FP}$$

$$\text{Recall: } \frac{TP}{TP+FN}$$

$$\text{F-Measure: } \frac{2*Precision*Recall}{Precision+Recall}$$

$$\text{TPR: } \frac{TP}{TP+FN}$$

$$\text{FPR: } \frac{FP}{FP+TN}$$

$$\text{AUC: } \frac{S_p - n_p(n_n+1)/2}{n_p * n_n}$$

Here, TP denotes the number of correctly predicted positive samples, TN denotes the number of correctly predicted negative samples, FP denotes the number of incorrectly predicted positive samples, and FN denotes the number of incorrectly predicted negative samples. For AUC, S_p denotes the sum of the ranks of all positive samples, whereas n_p and n_n denote the number of positive and negative samples respectively [46][47].

The values of all metrics, apart from ROC, range from 0.0 to 1.0. In most metrics, values above 0.9 are considered outstanding, between 0.7 and 0.9 are considered acceptable, and below 0.7 are considered undesirable. In a classification problem, an AUC value of 1.0 suggests perfect prediction, 0.5 suggests random prediction, and any value less than 0.5 suggests poor prediction [10]. A value of 0.5 for AUC indicates that the corresponding ROC curve will fall on the diagonal line. A ROC curve above the diagonal line shows the model's ability to distinguish different classes [48].

All metrics were used to examine the performance of the trained models. In addition, F-Measure was also used during the hyperparameter search.

4. Results

4.1. Environmental Setup

We carried out our analysis on Penn States ROAR supercomputer clusters (ROAR) and a third-party cloud-based environment called vast.ai. The computational node on ROAR ran on Intel(R) Xeon(R) Processor @ 2.8 GHz. It provided a RAM of size 256 GB. The virtual machine on vast.ai ran on AMD Ryzen

5 3600X Processor @ 3.6 Ghz and NVIDIA(R) Titan RTX Graphics Processing Unit (GPU) @ 1770 MHz and frame buffer of size 24 GB. It provided a RAM of size 32 GB. Our analysis was implemented using Python 3.8.3. The Scikit learn 1.0.1 was used for implementing SVM classification model, grid search, ROC plot and AUC score. Matplotlib 3.5.1 was used to export ROC plots. Tensorflow 2.8.0 was used for the implementation of neural network. Keras-tuner 1.1.0 was used for the implementation of random search for neural network models. Imbalanced-learn 0.9.0 was used for the implementation of SMOTE oversampling technique. Packages on both environments were of the same version to ensure consistency.

4.2. Optimal Hyperparameters

We set the search algorithm to search 100 sets of NN hyperparameters randomly. Only the kernel choice was searched for SVM, and the grid search was utilized to ensure every kernel was used during searching. The model's F-Measure score was used to determine optimal hyperparameters. The optimal hyperparameters of both NN and SVM for each input set are shown in Table 3 based on the searched criteria depicted earlier in Table 1 and Table 2.

Hyperparameter	Original Set	Autoencoder Set	PCA Set	RBF-PCA Set	CR Set
Learning Rate	0.0005	0.0015	0.0005	0.001	0.0015
Number of Layers	5	2	4	4	2
Layer 1 Nodes	2982	94	94	110	17
Layer 1 Dropout Rate	0.14	0.18	0.14	0.14	0.26
Layer 1 Activation	Swish	Swish	Swish	Swish	RELU
Layer 2 Nodes	951	39	31	55	4
Layer 2 Dropout Rate	0.22	0.18	0.26	0.14	0.18
Layer 2 Activation	ELU	Tanh	Tanh	Sigmoid	ELU
Layer 3 Nodes	803		15	23	
Layer 3 Dropout Rate	0.3		0.14	0.14	
Layer 3 Activation	ELU		Tanh	Sigmoid	
Layer 4 Nodes	245		15	7	
Layer 4 Dropout Rate	0.18	n/a	0.26	0.14	n/a
Layer 4 Activation	ELU		Tanh	RELU	
Layer 5 Nodes	202				
Layer 5 Dropout Rate	0.1		n/a	n/a	
Layer 5 Activation	Swish				

(a) Optimal NN Hyperparameters.

Hyperparameter	Original Set	Autoencoder Set	PCA Set	RBF-PCA Set	CR Set
Kernel	Linear	Linear	Linear	Linear	Poly

(b) Optimal SVM Hyperparameters

Table 3: Optimal Hyperparameters.

4.3. Results of Models

The results of each trained classification model of each input set are shown from Table 4 to Table 8. Comparing Table 5 to Table 7 with Table 4, we can see that the performance of the classification models increased significantly when using dimensionally reduced data sets, especially when we look at their AUC scores. The reason for this is that dimensionality reduction techniques project the original data onto a lower dimensional space. By reducing the number of dimensions, the relationships between each dimension become less complex. Thus, discovering relationships between input data and desired output labels becomes easier, which in turn increases the performance of the classification models that rely on these relationships.

Classification Model	Accuracy	Precision	Recall	F-Measure	AUC
NN	0.944	0.979	0.962	0.970	0.785
SVM	0.955	0.970	0.980	0.975	0.885

Table 4: Performance of models on original set

As shown in Table 4, SVM had overall the best performance on the CR set, scoring higher than NN in all metrics but precision.

Classification Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**AUC**
NN	0.978	0.986	0.990	0.988	0.943
SVM	0.970	0.980	0.987	0.983	0.925

Table 5: Performance of models on Autoencoder set

Classification Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**AUC**
NN	0.966	0.982	0.980	0.981	0.891
SVM	0.965	0.981	0.980	0.981	0.891

Table 6: Performance of models on PCA set

Classification Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**AUC**
NN	0.944	0.956	0.982	0.969	0.888
SVM	0.964	0.970	0.990	0.980	0.935

Table 7: Performance of models on RBF-PCA set

Comparing Table 5, 6 and 7, we can see that, of all the three dimensionality reduction techniques used, the autoencoder provided the best result. We think this is because PCA, even with a kernel, relies on the linear relationship between each dimension, whereas autoencoder can extract nonlinear relationships. By examining these tables, it seems that NN acheived better performance than SVM on data sets reduced by nonlinear dimensionality reduction methods like autoencoder, while SVM provided nearly as good or better performance on data sets reduced by linear dimensionality reduction methods like PCA.

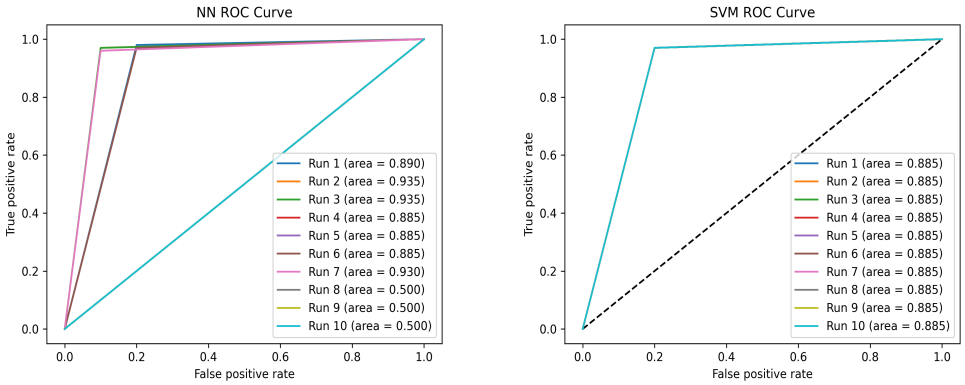
Classification Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**AUC**
NN	0.935	0.980	0.950	0.965	0.730
SVM	0.902	0.943	0.949	0.946	0.690

Table 8: Performance of models on CR set

As shown in Table 8, NN had the best score in all metrics on the CR set. Comparing Table 8 with Table 4, we can see that the performance of the classification

models was worse when using only cancer-relevant genes. We think this is because apart from these identified prostate cancer-relevant genes, there are more genes that are relevant to prostate cancer to some degrees.

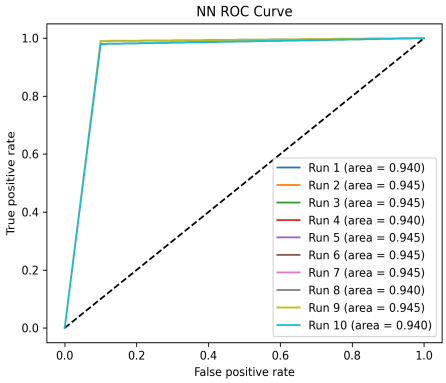
The ROC curve of each trained classification model of each input set is shown in Fig. 5 to Fig. 9.



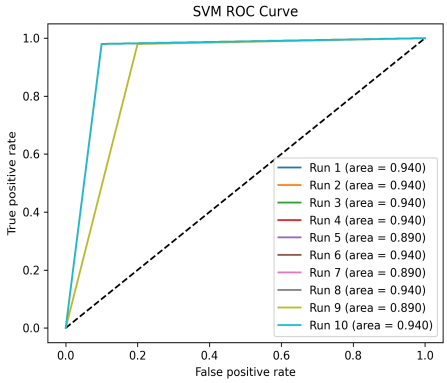
(a) NN ROC (b) SVM ROC

Figure 5: Original set ROC Curve.

As shown in Figure 5, we can see that SVM performed stably, while the performance of NN varied a lot. Although 7 out of 10 iterations of NN showed the same or better performance than SVM, the other 3 iterations showed really bad performance. This could be a result of overfitting during training processes.

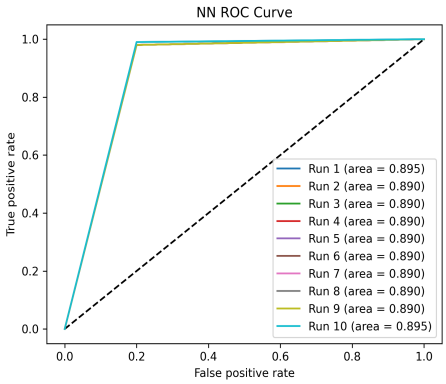


(a) NN ROC.

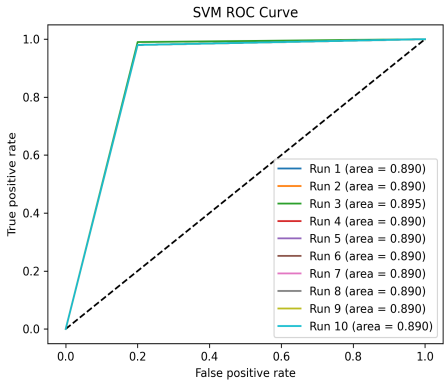


(b) SVM ROC.

Figure 6: AE set ROC Curve.

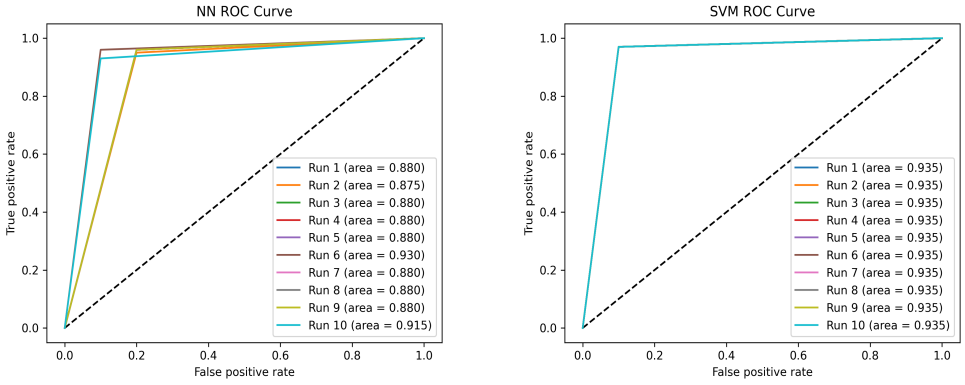


(a) NN ROC



(b) SVM ROC

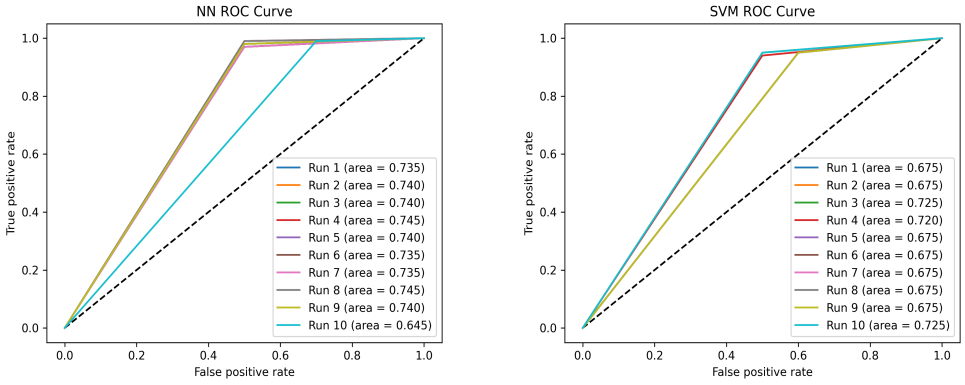
Figure 7: PCA set ROC Curve.



(a) NN ROC (b) SVM ROC

Figure 8: RBF-PCA set ROC Curve.

Comparing Figure 6, 7 and 8, we can see that autoencoder provided the best performance. We think this is because autoencoder is able to produce a reduced data set that maintains the original nonlinear relationships between features. We can also see that the performance of RBF-PCA, when paired with SVM, was comparable to autoencoder. We think this is because RBF-PCA can produce linearly separable data sets, which SVM prefers. Additionaly, when comparing these figures with Figure 5, we can see that none of all classifiers showed really bad performance. We believe this is because dimensionality reduction techniques can reduce the complexity of data, thus can reduce the possibility of overfitting during classification processes.



(a) NN ROC (b) SVM ROC

Figure 9: CR set ROC Curve.

Comparing Figure 9 with other figures, we can see that the performance of classifiers using only cancer-relevant features were a lot worse than ones using all features or reduced features. We believe this is a result of excluding other undiscovered relevant features.

5. Discussion

In this study, we investigated the performance of three types of dimensionality reduction methods on ML models used for prostate cancer prediction. Based on the results, we see that cancer prediction models performed better on dimensionally reduced data sets. This implies that dimensionality reduction can increase performance of ML models when using high dimensional data. Additionally, we see that autoencoder provided greater performance increase over PCA and RBF-PCA. This implies that nonlinear dimensionality reduction methods like autoencoder provide greater performance boost over linear and sublinear ones on high dimensional continuous data such as gene expression data. Furthermore, we see that our cancer prediction models performed better when we applied all predictors instead of only cancer relevant ones. This implies that there are more genes that contribute to predictions apart from the 36 cancer relevant genes.

We analyzed the performance of our models using 6 types of metrics. Because our data was highly imbalanced, with 1 to 10 ratio between non-cancer

samples and cancer samples, relying solely on accuracy to measure the performance was unreliable since a model could achieve high validation accuracy by simply predicting every sample to be a cancer sample. To fix this problem, apart from using SMOTE to generate new non-cancer samples, we also included other metrics that can reflect the performance of models even when using imbalanced data. Since our objective was to predict patients with prostate cancer, we chose precision, which measures the ratio of correctly predicted positive samples over all predicted positive ones, and recall, which measures the ratio of correctly predicted positive samples over all positive ones. We also included the F-measure, the score of which is only high when both precision and recall are high. Additionally, we included ROC curves and AUC scores. We believe that the inclusion of these three metrics provided an explicit presentation of the performance of our models.

Despite what we had done so far, this study still has some limitations. First, we only investigated three dimensionality reduction algorithms. Although we believe that the algorithms we chose are representative of three types of dimensionality reduction techniques, we will explore other algorithms in our future studies. Second, although our ML models showed high performance in predicting cancer patients, such data-driven cancer prediction techniques will be affected by the collection time of data. This could impact the practicality of using ML cancer prediction models in the real world. In our future studies, we will collaborate with medical professionals to test our models in a real clinical scenario.

6. Conclusion and Future Work

Prostate cancer is the most common and second deadliest cancer among men in the United States. Damages caused by this cancer can be reduced to a minimum if it is detected and treated at its early stage. Due to the high dimensionality of RNA-seq data, current ML techniques could not handle them accurately and efficiently. Feature selection and dimensionality reduction are implemented to reduce the complexity and increase the performance of ML models of such data. The main focus of this paper was to analyze the impact dimensionality reduction methods have on ML models used to detect prostate cancer. This study showed that incorporating dimensionality reduction techniques into ML models can significantly increase their performance. Additionally, the autoencoder performed better for this particular data set because it can extract the nonlinear relationships

from a given data set.

In this study, we examined the impact dimensionality reduction techniques have on the performance of ML models. Future studies will include their impact on running time and computational resource usage. In addition, we will conduct an investigation comparing the impact dimensionality reduction techniques and feature selection methods have on similar data sets.

References

[1] N. C. Institute, Prostate cancer-patient version, last accessed 29 July 2022. URL https://www.cancer.gov/types/prostate

[2] A. C. Society, Key statistics for prostate cancer, last accessed 7 August 2022. URL https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html

[3] A. Bhola, A. K. Tiwari, Machine learning based approaches for cancer classification using gene expression data, Machine Learning and Applications: An International Journal (MLAIJ) 2 (3/4) (2015).

[4] M. Hong, S. Tao, L. Zhang, L.-T. Diao, X. Huang, S. Huang, S.-J. Xie, Z.-D. Xiao, H. Zhang, Rna sequencing: new technologies and applications in cancer research, Journal of hematology & oncology 13 (1) (2020) 1–16.

[5] J. Wang, D. C. Dean, F. J. Hornicek, H. Shi, Z. Duan, Rna sequencing (rna-seq) and its application in ovarian cancer, Gynecologic oncology 152 (1) (2019) 194–201.

[6] P. Sharma, N. S. Sahni, R. Tibshirani, P. Skaane, P. Urdal, H. Berghagen, M. Jensen, L. Kristiansen, C. Moen, P. Sharma, et al., Early detection of breast cancer based on gene-expression patterns in peripheral blood cells, Breast Cancer Research 7 (5) (2005) 1–11.

[7] J. Eswaran, A. Horvath, S. Godbole, S. D. Reddy, P. Mudvari, K. Ohshiro, D. Cyanam, S. Nair, S. A. Fuqua, K. Polyak, et al., Rna sequencing of cancer reveals novel splicing alterations, Scientific reports 3 (1) (2013) 1–12.

[8] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, et al., Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer, Nature communications 8 (1) (2017) 1–12.

[9] V. Dhar, Data science and prediction, Communications of the ACM 56 (12) (2013) 64–73.

[10] M. F. Kabir, S. A. Ludwig, Enhancing the performance of classification using super learning, Data-Enabled Discovery and Applications 3 (1) (2019) 5.

[11] P. Harrington, Machine learning in action, Simon and Schuster, 2012.

[12] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.

[13] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[14] V. Chang, V. R. Bhavani, A. Q. Xu, M. Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, Healthcare Analytics 2 (2022) 100016.

[15] D. Shah, S. Patel, S. K. Bharti, Heart disease prediction using machine learning techniques, SN Computer Science 1 (6) (2020) 1–6.

[16] A. S. A. Rabby, R. Mamata, M. A. Laboni, S. Abujar, et al., Machine learning applied to kidney disease prediction: Comparison study, in: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), IEEE, 2019, pp. 1–7.

[17] J. I. Z. Chen, P. Hengjinda, Early prediction of coronary artery disease (cad) by machine learning method-a comparative study, Journal of Artificial Intelligence 3 (01) (2021) 17–33.

[18] C. Krittanawong, H. U. H. Virk, S. Bangalore, Z. Wang, K. W. Johnson, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai, et al., Machine learning prediction in cardiovascular diseases: a meta-analysis, Scientific Reports 10 (1) (2020) 1–11.

[19] S. M. Ganie, M. B. Malik, An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators, Healthcare Analytics 2 (2022) 100092.

[20] M. Li, X. Fu, D. Li, Diabetes prediction based on xgboost algorithm, in: IOP conference series: materials science and engineering, Vol. 768, IOP Publishing, 2020, p. 072093.

[21] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, et al., Convolutional neural networks-based mri image analysis for the alzheimers disease prediction from mild cognitive impairment, Frontiers in neuroscience 12 (2018) 777.

[22] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, M. Ingalhalikar, Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri, NeuroImage: Clinical 22 (2019) 101748.

[23] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, A. D. N. Initiative, et al., Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, NeuroImage: Clinical 21 (2019) 101645.

[24] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Ieee, 2015, pp. 1200–1205.

[25] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1) (2002) 389–422.

[26] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthcare Analytics 2 (2022) 100032.

[27] J. C. Ang, A. Mirzal, H. Haron, H. N. A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE/ACM transactions on computational biology and bioinformatics 13 (5) (2015) 971–989.

[28] H.-J. Chiu, T.-H. S. Li, P.-H. Kuo, Breast cancer–detection system using pca, multilayer perceptron, transfer learning, and support vector machine, IEEE Access 8 (2020) 204309–204324.

[29] W. U. Adiwijaya, E. Lisnawati, A. Aditsania, D. S. Kusumo, et al., Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification, Journal of Computer Science 14 (11) (2018) 1521–1530.

[30] V. J. Kadam, S. M. Jadhav, K. Vijayakumar, Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression, Journal of medical systems 43 (8) (2019) 1–11.

[31] W. Liu, H. Lin, L. Huang, L. Peng, T. Tang, Q. Zhao, L. Yang, Identification of mirna–disease associations via deep forest ensemble learning based on autoencoder, Briefings in Bioinformatics 23 (3) (2022) bbac104.

[32] P. Danaee, R. Ghaeini, D. A. Hendrix, A deep learning approach for cancer detection and relevant gene identification, in: Pacific symposium on biocomputing 2017, World Scientific, 2017, pp. 219–229.

[33] J. Huang, Y.-F. Li, M. Xie, An empirical analysis of data preprocessing for machine learning-based software cost estimation, Information and software Technology 67 (2015) 108–127.

[34] J. Han, J. Pei, H. Tong, Data mining: concepts and techniques, Morgan kaufmann, 2016.

[35] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural computation 10 (5) (1998) 1299–1319.

[36] J. A. Rosenthal, Statistics and data interpretation for social work, Springer publishing company, 2011.

[37] W. Pedrycz, S.-M. Chen, Deep learning: Concepts and architectures, Springer, 2020.

[38] M. F. Kabir, S. A. Ludwig, Classification of breast cancer risk factors using several resampling approaches, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 1243–1248.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[40] B. Davazdahemami, P. Peng, D. Delen, A deep learning approach for predicting early bounce-backs to the emergency departments, Healthcare Analytics 2 (2022) 100018.

[41] M. F. Kabir, S. A. Ludwig, Association rule mining based on ethnic groups and classification using super learning, Applied Smart Health Care Informatics: A Computational Intelligence Perspective (2022) 111–129.

[42] M. F. Kabir, S. A. Ludwig, Classification models and survival analysis for prostate cancer using rna sequencing and clinical data, in: 2019 IEEE international conference on big data (big data), IEEE, 2019, pp. 2736–2745.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (1) (2014) 1929–1958.

[44] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing 408 (2020) 189–215.

[45] MedlinePlus, Prostate cancer, last accessed 4 September 2022.
URL `https://medlineplus.gov/genetics/condition/prostate-cancer/#causes`

[46] M. Hossin, M. N. Sulaiman, A review on evaluation metrics for data classification evaluations, International journal of data mining & knowledge management process 5 (2) (2015) 1.

[47] D. J. Hand, R. J. Till, A simple generalisation of the area under the roc curve for multiple class classification problems, Machine learning 45 (2) (2001) 171–186.

[48] J. N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316.