UNIVERSITI TEKNOLOGI MARA

ENHANCING CLASSIFICATION PERFORMANCE ON HIGH-DIMENSIONAL DATA USING AN IMPROVED PROCEDURE OF SPARSE PRINCIPAL COMPONENT ANALYSIS (PCA) AND RANDOM FOREST

ZAHWA AHMED JASIM

MSc

June 2026

UNIVERSITI TEKNOLOGI MARA

ENHANCING CLASSIFICATION PERFORMANCE ON HIGH-DIMENSIONAL DATA USING AN IMPROVED PROCEDURE OF SPARSE PRINCIPAL COMPONENT ANALYSIS (PCA) AND RANDOM FOREST

ZAHWA AHMED JASIM

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Statistics)

Faculty of Computer and Mathematical Sciences

June 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 19th June 2025 to conduct the final examination of Zahwa Ahmed Jasim on her Master of Science thesis entitled "Enhancing Classification Performance on High Dimensional Data Using an Improved Procedure of PCA and Random Forest" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Marina Yusoff, PhD
Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Chairman)

Norshahida Shaadan, PhD
Associate Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

Hazrul Abdul Hamid, PhD
Senior Lecturer
School of Distance Education
Universiti Sains Malaysia
(External Examiner)

PROFESSOR DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA
Date : 30 June 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Zahwa Ahmed Jasim
Student I.D. No.	:	2021740989
Programme	:	Master of Science (Statistics) – CS753
Faculty	:	Computer and Mathematical Sciences
Thesis Title	:	Enhancing Classification Performance on High Dimensional Data Using an Improved Procedure of PCA and Random Forest
Signature of Student	:	…….
Date	:	June 2026

ABSTRACT

Principal Component Analysis (PCA) is a technique that reduces data dimensionality by converting correlated variables into uncorrelated components. It simplifies data while maintaining patterns. Combining PCA with Random Forest (RF) can increase efficiency and reduce overfitting. Nonetheless, its benefits depend on data standardization. PCA is a statistical strategy that calculates the covariance matrix of a data set, extracting eigenvectors and eigenvalues to minimize dimensionality while preserving significant variance. It involves covariance-based eigenvalue decomposition and Singular Value Decomposition (SVD), with the best PCs accounting for 95% of the variance. Modified PCA data can be used for training RF, but potential signal loss and interpretability issues may arise. High-dimensional data classification in genomics and medical diagnostics faces challenges due to noise accumulation, class imbalance, and dimensionality. Biased classification, noise buildup, and instability in feature selection hinder performance. Misclassification of training data affects both supervised and unsupervised techniques, leading to increased error rates and resource inefficiency. RF has garnered considerable attention from various researchers among numerous classification approaches. However, RF performs poorly with very high- dimensional data. Therefore, the objective of this study was to classify high- dimensional data using RF alone and a combined RF with PCA and Sparse PCA. The study used SVD and Random Orthogonal Matrix (ROM) to reduce a high-dimensional dataset. The Variance Maximization (VM), Reconstruction Error Minimization (REM), and SVD are effective methods for high-dimensional datasets, enhancing data variability assessment, model resilience, and dimensionality reduction, facilitating efficient summarization, noise reduction, and data analysis or predictive modeling. A PCA strategy that minimizes data reconstruction error, ensuring condensed data faithfully depicts original data, is instrumental in image processing and other critical data reconstruction and compression applications. SVD is a factorization that decomposes a matrix into three components: U, D, and V, which are used for tasks such as dimensionality reduction and solving systems of equations. The results show that RF alone (accuracy = 93.98%) performed best compared with VM combined with RF (accuracy = 73.23%), RF combined with PCA (accuracy = 92.48%), REM with RF (93.33%) and SVD with RF (93.33%). Sparse PCA is an alternative method for addressing dimensionality issues that hinder classification performance in RFs, particularly in genomic studies. The findings highlight that Sparse PCA techniques, particularly, effectively address dimensionality issues and significantly enhance classification performance in RF, especially for genomic applications. The study indicates that integrating dimensionality-reduction methods with RF classification substantially improves performance in high-dimensional domains, such as genomics. Techniques such as PCA, Sparse PCA, and VM-based approaches enhance classification performance on complex datasets.

ACKNOWLEDGEMENT

Praise be to Allah who enlightened my path and guided me in my endeavor with His grace, and peace and blessings be upon our master Muhammad and his family and companions. As I conclude my thesis, it is not enough for me to express my thanks, appreciation and pride to the honorable lecturer (Professor Dr. Ahmed Zia ul- Saufie Muhammad Japeri), the supervisor of the thesis, for the wise scientific guidance, follow-ups, and valuable efforts he provided that had a significant impact in producing it in the form it is in. In addition to his high morals. I also extend my sincere thanks to the honorable professors, the head of the discussion committee, and its honorable members for their kind acceptance of the thesis for discussion. It is also my duty to extend my sincere thanks to all my esteemed professors in the Department of Computer and Mathematical Sciences at the college, who shared their knowledge with me during my years of study and worked hard to support the academic progress of all students.

I also extend my thanks, appreciation, and pride to the honorable lecturer (Dr. Zalina Zahid) for the sound scientific guidance she provided.

I would also like to thank the staff at Perpustakaan Tun Abdul Razak UiTM for their guidance in formatting my thesis in accordance with the guidelines provided by the Institute of Postgraduate Studies, particularly Madam Syarifah Hanisah, who is part of the MZJ Formatting Method Teams.

I would also like to thank Ms. Norhidayah A. Kadir, Ms. Sharif, and Assoc. Prof. Dr. Shafaf for their assistance in following up on the study requirements throughout my stay at the college.

Finally, I would like to extend my sincere thanks to my family for their support and prayers, and to everyone who offered me assistance but whom I did not mention. I apologize to my colleagues and friends whom I failed to mention, and I appreciate your understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbols

A	Number of PLS or PCA Components in the Model and the Number of Selected Latent Variables in the Model
a	Number of the PLS or PCA Component
B	PLS Regression Coefficient
b	Coarse APM Block
C	Number of blocks (b=1,2,3,…….K)
C_p	Pooled Covariance Matrix for the Two Classes
C_g	Covariance Matrix for Class G

LIST OF ABBREVIATIONS

Abbreviations

3D	Three – Dimensional
AHP	Analytical Hierarchical Process
ANN	Artificial Neural Networks
CDSS	Clinical Decision Support Systems
CG	Conjugate Gradient
CS	Compressive Sensing
CV	Canonical Variate
DAE	Deep Auto -Encoder
DL	Deep Learning
DR	Diminishing Reduction
EBC	Exact Bayesian Computation
ESRD	End-Stage Renal Disease
FE	Feature Extraction
FFT	Ast Fourier Transform
FHI	Fast Hartley Transformation
FMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FS	Feature Selection

ICA	Independent Component Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linear Discriminant
LR	Logistic Regression
MCMC	Markov Chain Monte Carlo
MIM	Mendelian Inheritance in Man
ML	Machine Learning
MLR	Multiple Linear Regression
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PCA	Principal Component Analysis
PCs	Principal Components
PCT	Polar Coordinate Transforms
PM	Probabilistic Model
RF	Random Forest
RFE	Recursive Feature Elimination
SPCA	sparse principal component analysis
SVD	Singular Value Decomposition
SVM	Support Vector Machine
T1D	Type 1 diabetes
T2D	Type 2 Diabetes are Two Categories
TN	True Negative

US	Using Ultrasonography
VM	Variance Maximization

CHAPTER 1
INTRODUCTION

1.1 Research Background

The rapid development of computer technology has led to increased data generation and collection. Note that combining Principal Component Analysis (PCA) and Random Forest (RF) can improve data analysis, particularly for high-dimensional datasets. RF uses random feature selection to construct decision trees, assess feature relevance, and capture nonlinear interactions (Rahangdale et al., 2021). It is a modern machine learning technique that effectively extracts crucial information from large datasets.

Data classification involves selecting the optimal model to distinguish among data classes or concepts. Hence, PCA is a powerful tool for reducing data dimensionality, identifying biomarkers, and improving diagnostic accuracy in lung and breast cancer. Breast cancer, the most common cancer among women, affects 670,000 people in 2022. Risk factors include hormonal exposure, family history, genetic predispositions, age, lifestyle choices, and environmental exposure. Improved survival rates increase secondary malignancies. RF algorithms facilitate the prediction of breast cancer risk and outcomes in complex datasets (Fernández-Delgado et al., 2014). RF is an ensemble learning technique that generates unpruned decision trees from different training data subsets. Each tree is trained on a random subset of the data with replacement, and features are evaluated at each split. Consequently, this combination reduces overfitting and correlation, yielding more accurate predictions. RF is ideal for classification and regression tasks, as it minimizes variance and mitigates overfitting. It is simple, versatile, and reliable, even without complex parameter adjustments (Lin et al., 2021).

Massive data production has led to high-dimensional data, characterized by more features than observations. This type of data is often found in genetics and medical sciences, exemplified by extensive medical records of breast cancer patients. These records encompass numerous variables that researchers aim to correlate with patient outcomes, including hospital length of stay and treatment response.

Dealing with these vast, high-dimensional data sets is a key current issue for many conventional classification techniques. Typical examples include genomic data, which often contain many observations and thousands of features. The specific properties of extremely high-dimensional data can significantly impact the performance of classification systems. According to several studies, Lin et al. (2021), Kwon and Sim (2013), and Jia et al. (2022), three major problems can arise, including performance issues in the classification algorithm due to the high-dimensional characteristics of the data. One is that high-dimensional data may suffer from the curse of dimensionality, the tendency for data to have more features than observations. More specifically, when there are too many features, it will cause every observation to appear to be equally distant from every other observation. This will seem to demonstrate the tendency for every observation to be either equally similar or equally dissimilar, which makes it challenging for classification algorithms to forecast meaningful data groups.

The curse of dimensionality in high-dimensional data, characterized by more features than observations, poses significant challenges for machine learning and data analysis due to factors like uniformity, sparsity, computational complexity, visualization difficulties, and multicollinearity. High-dimensional data often has features that depend on multiple other features, leading to overfitting and poor classification performance. Correlation features can produce redundant information, which can lead to overfitting. Additionally, classification algorithms may take longer due to the large number of features, thereby increasing the minimum time required for processing.

High-dimensional data presents classification algorithms with challenges such as feature redundancy and overfitting, necessitating careful feature selection and dimensionality reduction strategies to improve efficiency and reduce overfitting. Therefore, PCA and RF are widely used techniques for analyzing high-dimensional data across various fields, including genetics, image processing, and finance. PCA reduces dimensionality by creating uncorrelated variables, simplifying the data while capturing variance. It offers feature extraction, visualization, and processing, providing efficiency and improved model performance in machine learning preprocessing. Meanwhile, RF is a tree-based ensemble learning technique used for feature selection, regression, and classification in high-dimensional data. It effectively manages missing data, ranks features, and balances bias and volatility. RF can be combined with PCA

toreduce dimensionality, as well as other methods such as t-SNE, autoencoders, and penalized regression.

Nevertheless, analysis of high-dimensional data has proven challenging. Unlike low-dimensional data, where the number of observations significantly outnumbers the number of features, in high-dimensional data analysis, the number of features necessitates consideration of potential difficulties arising from their abundance. When high-dimensional data are involved, traditional multivariate analysis methods are not readily applicable to classification. This circumstance has recently led to the establishment of a new research area, in which improvements in classification method accuracy have been achieved by integrating statistical methods and machine learning algorithms. Numerous studies on hybrid techniques have advocated methodological advances and improvements in classification-system accuracy. The growth in computer applications has also contributed to the development of the high-dimensional data classification challenge.

1.2 Problem Statement

It is well known that the primary goal of classification methods is to assign an observation under an unidentified class into a class that is known. Although classification is a standard statistical technique, determining an accurate method for assigning new observations to their correct classes remains challenging. RF has garnered considerable attention from various researchers across numerous categorization approaches. RF's approach achieves higher accuracy than state-of-the-art supervised classification methods.

RFs are regression or classification methods based on an ensemble of unpruned trees. Generally, the RF method is one of the most successful classification methods. However, according to previous research (Xu et al., 2012), RF faces challenges in dealing with high-dimensional data due to the inefficiency of random feature selection and the curse of dimensionality. Inefficient random selection often includes non-informative or noisy variables, leading to weak decision trees and reduced performance.

The sparsity of trees negatively impacts accuracy and reliability. RF's overfitting in high-dimensional data is due to noise and computational costs. It is less effective in sparse, irrelevant traits. Advanced adaptations, such as High-Dimensional

Random Forests (HDRF), employ techniques like Ridge Regression-based screening to enhance feature selection, manage multicollinearity, and improve predictive accuracy. These strategies are essential for mitigating overfitting, enhancing model robustness, and minimizing computational costs in high-dimensional settings. Note that (features) which are less informative or not informative to the explanation of the target variable. Uninformative variables, also referred to as noise variables, have little ability to predict the target variable because they are only weakly related to it. In general, data with a high concentration of uninformative features have fewer expected informative variables.

This makes it challenging to build a reliable RF model. Furthermore, in high-dimensional data, strong correlations often exist among numerous features. In high-dimensional datasets, strongly connected features typically reduce the predicted variable relevance scores of correlated features, making RFs less effective at identifying the strongest features (Fiagbe, 2021). Furthermore, when the number of characteristics is large, RF methods take longer to classify observations.

PCA is a method that enhances classification performance by addressing dimensionality, mitigating overfitting, and reducing computational costs. PCA focuses on significant features, reducing noise and extraneous data, thereby improving accuracy and precision. This approach supports RF classification by focusing on valuable features for decision trees. Empirical results support PCA's ability to produce good classification performance while reducing dimensionality and computing complexity.

Sparse Principal Component Analysis (Sparse PCA) is a dimensionality-reduction method that eliminates redundant variables by setting their loadings to zero, thereby improving interpretability and stability in high-dimensional datasets. Methods like the Variance Maximization Approach are suggested. The Reconstruction Error Minimization (REM) and Singular Value Decomposition (SVD) techniques are used to reconstruct data with minimal error. Sparse PCA, used in cancer research, reduces data dimensionality. Combining RF and Sparse PCA can improve classification accuracy in high-dimensional datasets by focusing on informative features. Therefore, this study will evaluate three Sparse PCA methods to determine which is most effective for feature selection when combined with RF for classification. It is hoped that combining RF with these three Sparse PCA methods will improve the accuracy of classifying high-dimensional data.

1.3 Research Questions

a) How does the existing RF method perform on high-dimensional data classification?
b) How to enhance the RF model for classification in high–dimensional data?
c) What is the performance of the proposed methods in classifying positive and non-positive cases?

1.4 Research Objectives

a) To investigate the classification performance of the existing RF method on high-dimensional data sets.
b) To enhance the existing RF by integrating PCA and Sparse PCA.
c) To investigate and verify the performance of the proposed methods.

1.5 Significance of the Study

An enhanced RF method integrated with Sparse PCA is proposed to improve classification performance on high-dimensional data. It is hoped that, through the proposed methodology improvement, the classification method will contribute to an accurate, efficient, and effective decision-making process. In gene-expression data classification, it is crucial to design effective models that are computationally efficient, rapid, and easy to execute. The integrated approach, utilizing Sparse PCA as a feature selection method, is expected to be notable for its potential to deliver outstanding performance and for its advantages in tackling dimensionality problems that impede classification, particularly in genomic investigations (Chuang et al., 2014). Research on lung and breast cancer benefits individuals by improving treatments, increasing survival rates, and improving quality of life. Moreover, it helps survivors cope with emotional difficulties and offers new therapeutic strategies. Cooperation between researchers, physicians, government agencies, businesses, and nonprofits ensures funding for impactful studies.

1.6 Scope and Limitation of Study

In this study, the classification method focuses on RF, which has been identified as the most powerful. Enhancing the RF as a seven-classification, three types of Sparse PCA will be considered as feature extraction methods, covering the Variance Maximization (VM) approach, REM approach, and SVD approach. For experimental purposes, simulation data will be generated using Markov Chain Monte Carlo (MCMC), accounting for high-dimensional data issues, such as when the number of features exceeds the sample size, and multicollinearity arises. Further model verification will utilize two benchmark data sets taken from Koç (2021).

RF is a complex machine learning technique with limitations, including high memory usage, interpretability challenges, slower prediction speeds, and overfitting in noisy data. Its conclusions may not be directly comparable to those of other techniques. Note that RF uses Sparse PCA for feature selection and artificial data generated by MCMC, which may not accurately capture the noise and complexity of real-world datasets. The research results may not generalize to real-world applications due to its focus on high-dimensional data and limited benchmarking against existing datasets. (Koç, 2021), necessitating more extensive validation for generalizability and robustness.

This research utilizes single-cell protein structure analysis (SPCA) techniques to study protein expression patterns and molecular alterations associated with breast and lung cancer. It aims to explore how SPCA can help identify early cellular changes, potential biomarkers, and mechanisms of cancer progression, as well as structural shifts between cancerous and non-cancerous cells. Data from cell samples will be collected and analyzed using computational and biochemical methods to determine correlations between SPCA findings and cancer type or stage Koç (2021).

Using Random Forest as the primary classifier and three Sparse PCA variants for feature extraction from high-dimensional genomic data, a framework is developed to categorize seven kinds of breast cancer. MCMC simulations use methods such as Singular Value Decomposition for sparse feature selection, Ridge-Elastic net Mixture, and Variance Maximization to handle multicollinearity and high-dimensional feature issues. Koç (2021) datasets are used to validate the reliability of the technique.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter discusses past studies on classification under machine learning for high-dimensional data. Section 2.2 focuses on Classification methods in machine learning. Section 2.3 covers Random Forest (RF) as a Classification Method. Next, Section 2.4 reviews the effects of high-dimensional data on the RF Method, followed by Section 2.5, which focuses on Feature Selection methods. Finally, the entire literature review is summarized in Section 2.6. High-dimensional data, characterized by large feature sets, sparsity, and complexity, is utilized across various sectors, including gene expression and image processing. Despite challenges such as overfitting, effective management can enable precise modeling of complex processes, including natural language processing and healthcare data.

2.2 Classification Method Under Machine Learning

Machine learning, a subfield of Artificial Intelligence (AI), employs algorithms to learn from data and make decisions. It underpins AI applications such as image identification and natural language processing, necessitating further research to explore specific methods for managing extensive data volumes and uncovering intricate patterns beyond conventional statistical techniques (Goldstein et al., 2017; Shickel et al., 2017). Other than that, it functions as an algorithmic model that learns to perform specific tasks by identifying patterns in the data it encounters, rather than being explicitly programmed by a human expert. This procedure comes in two types: supervised and unsupervised, with supervised methods being the most common (Erickson et al., 2017). Figure 2.1 on pages 24 illustrates the categorization of processes under machine learning techniques. Supervised learning occurs when an algorithm system is trained on a labeled dataset and then expected to classify a new, unknown data point. This is known as supervised learning. After the labeled dataset has completed its training, the algorithm is then tested on unlabeled data for classification or regression (Moore & Patton, 2019).

Meanwhile, unsupervised learning occurs when there is no training dataset for the corresponding output data. This type of learning aims to classify input data, much like supervised learning. Unsupervised learning techniques include clustering algorithms, such as K-Means clustering and K-Median clustering. The fundamental distinction in unsupervised learning is that categories of data are based on their integral feature because the input data are not labelled. This technique in machine learning enables researchers to learn more about the underlying distribution of data than would have been revealed otherwise (Chartrand et al., 2017).

2.3 Methods Used by The Previous Researchers to Classify High-Dimensional Data

Scholars have employed techniques such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) k-NN, RFs, and AdaBoost to categorize high-dimensional information. Hybrid classifiers combine conventional classifiers with feature selection to improve accuracy and interpretability. Techniques such as filtering, embedding, and regularization reduce superfluous characteristics, enhance model performance, and distinguish data distribution patterns. Machine learning, a subset of AI, is used to manage massive amounts of data and uncover intricate patterns that are inaccessible to traditional statistical methods (Goldstein et al., 2017). Instead of being explicitly coded by a human expert, it functions as an algorithmic model that learns to execute specific tasks by spotting and recognizing patterns in the data it sees. Several machine learning methods have been employed for classification and regression. Each of these machine learning techniques performs differently in terms of accuracy, efficacy, and other factors, depending on the type of data sets and methodology used. Only a few studies have compared high-dimensional classification approaches. However, those that did (Bolivar-Cime & Marron, 2013; Shao & Lunetta, 2012) indicated that the performance differences between the compared classification methods were not very noticeable. The lowest error rate was not always achieved by directly applying the classification algorithms. As a result, it was recommended to employ strategies such as feature selection before applying any categorization methods (Zekić-Sušac et al., 2014). As classification is a supervised learning task, its methodology includes identifying, comprehending, and classifying concepts and things into categories designated groupings, or “sub-populations” (Sun & Ongsomwang,

2023). In its simplest form, classification is a type of "pattern recognition" in which algorithms analyze training data to detect similar patterns in new datasets. According to Savargiv et al. (2021), predictive classification modeling is the process of predicting a mapping function (f) from discrete input variables (x) to output variables (y). Generally, in mathematical expressions, given the input feature vector is x`. Hence, the output score is:

$$y = (w\rightarrow, x) = f(\sum_j W_j \boldsymbol{X}_j) \quad (2.1)$$

The provided mathematical formula represents a standard linear model commonly used in machine learning and statistics, which involves an input feature vector (x) and a weight vector (w). Note that each feature is multiplied by its corresponding weight (w j). The output score is determined by a function f, which may be the identity function for linear regression or a sigmoid/SoftMax function for classification tasks, among other nonlinear functions, where $w\rightarrow\rightarrow$ is a real vector and f is a function that transforms the dot product of the two vectors into the targeted output. The weight vector w is estimated through learning from a set of labelled training samples. f can also be considered as a threshold function which maps w and $x\rightarrow$ above a certain threshold, say θ, to the first category and other values to the second category as indicated below:

$$f(X) = \begin{cases} 1 \text{ if } WT, & X > \theta, \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

Equation (2.2) describes the threshold activation function, where the output value is set to 1 if the weighted sum of the inputs exceeds a predefined threshold (θ); otherwise, the output value is 0. This function is commonly used in perceptron models for binary classification tasks. The provided function is a threshold (or step) function used in signal processing or machine learning. It is described as: f(x) = 1 if w^T x> 0, 1 otherwise. It represents the dot product of an input vector and a weight vector, multiplied by a scalar threshold θ\theta$\theta$. Perceptrons and binary classifiers often use this activation function. Many labels have been proposed to categorise classification tasks. In the simplest categories, classification can be divided into linear classifiers and nonlinear classifiers (Ouyang et al., 2019). These two categories of linear and

nonlinear classifications are separated based on whether strict assumptions regarding the shape of the mapping function are required. The premise of linear techniques is that features have only linear effects on the outcome. The performance of nonlinear approaches, in contrast, is less reliant on the model assumption.

Figure 2.1 illustrates the linear and nonlinear forms of classification algorithms, along with examples of these methods.

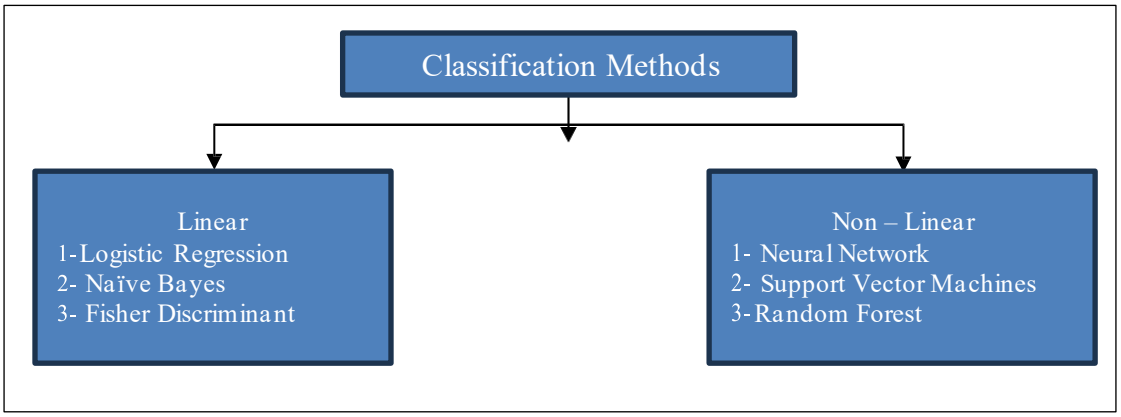

Figure 2.1 Types of Methods under Classification

Models known as "linear classifiers" categorise input vectors using linear (or hyperplane) decision boundaries. The classification target in linear classifiers used in machine learning is to categorize things based on comparable feature values. A linear classifier is often considered a rapid classifier when classification speed is a concern. Additionally, linear classifiers typically perform well when the number of dimensions is substantial. However, the margin has an impact on how quickly the variables in the data set converge. In general, the margin measures the ease with which a specific classification problem can be resolved by the dataset's capability to be linearly separable.Among the classifiers that divide categories using linear functions are the discriminant classifier, naive Bayes, and logistic regression, which can project features from a high-dimensional space into a lower-dimensional one. If the data are not linearly separable, a linear classification, however, cannot distinguish between the two classes with accuracy. Therefore, non-linear classification is a technique that can be used to categorise instances that cannot be separated linearly. Table 2.1 highlights the distinctions between linear classification and nonlinear classification. The first significant distinction focuses on whether the groups are split linearly or non-linearly. The distinction is whether classification is done with straight lines. Linear classification divides data into discrete classes using decision boundaries, while non-linear

classification models complex data patterns using curved or intricate boundaries, requiring a dataset and complexity for reliable predictions. Consequently, the hyperplane is utilized to separate data for linear classification, and kernels are incorporated to separate non-separable data.

Table 2.1
Differences Between Linear Classifiers and Non-Linear Classifiers

No.	Linear Classification	Non-Linear Classification
1	Linear Classification refers to the process of assigning a set of data points to a discrete class using a linear combination of its explanatory variables.	Non-linear Classification describes the process of dividing cases that cannot be divided linearly.
2	A straight line can be used to categorize data.	Data cannot be categorized along a straight line.
3	A hyperplane is used to categorize data and classify.	Data that cannot be separated is converted into separable data using kernels.

Table 2.1 presents the main differences between linear and nonlinear classifiers in terms of decision limits, complexity, computational requirements, and modeling capabilities. more theoretical or tied to a specific algorithm . Linear and non-linear classification are two key methods in machine learning. Each is tailored to specific data types and problems. Linear classifiers use a linear decision boundary, such as a line or hyperplane, to distinguish between classes. They are characterized by their interpretability, computational efficiency, and simplicity. They are best suited for linearly separable data, but have limitations in handling non-linear data. Non-linear classifiers employ complex decision boundaries to separate classes in datasets with inherent complexity, such as networks of neurons, KNN, or SVMs with non-linear kernels. Moreover, they offer flexibility, handling class issues that are not linearly separable. Nonetheless, they have limitations such as computational costs and risks of overfitting. An example of nonlinear models is the use of neural networks to analyze biological data, as they can reveal intricate biological linkages and retain predictive power even when linear signals are removed. Thus, linear models are preferred for linearly separable data, while complex datasets require careful adjustment to prevent overfitting. Numerous comparison studies have been conducted to evaluate the effectiveness of linear classifiers compared to non-linear classifiers (Kotsiantis et al., 2006). This analysis examined a few linear and non-linear classifiers to assess their performance with three different types of data, including those related to brain tumors,

diabetes, and brain cancer. The outcomes demonstrated that, across all datasets, nonlinear classifiers outperformed linear classifiers. Particularly, the Multi Layer Perceptron (MLP) and Learning Vector Quantization (LVQ) accuracy rates for the real brain tumour data set are 91% and 83%, respectively.

Next, based on the texture of canopy images, Min et al. (2017) investigated different non-linear machine learning approaches for biomass estimation in tropical forest areas. The collection included simulated photos of three different forest types, resulting in a variety of visual textures and canopy features (open vs. closed canopy, fine vs. coarse texture). Consequently, the findings demonstrate that nonlinear techniques applied to textural indices outperform linear techniques based on PCA and linear regression. Meanwhile, Dong et al. (2013) investigated the capability of an extreme learning machine. The Extreme Learning Machine (ELM) has been utilized in a study to enhance the accuracy of thyroid nodule categorization. The algorithm utilizes ultrasound images to identify sonographic features indicative of malignancy. It has demonstrated high specificity, sensitivity, and classification accuracy, aiding in risk stratification and clinical decision-making. However, its performance may vary depending on the dataset and features used. ELM to distinguish between benign and malignant thyroid nodules using sonographic characteristics in ultrasound images. Two critical factors were examined in relation to the effectiveness of ELM: the quantity of hidden neurons and the nature of the activation function. The results demonstrate that benign and malignant thyroid nodules exhibit significant differences (p-value 0.01), with echogenicity, calcification, margin, composition, and form being the most helpful criteria for distinguishing between them. The suggested non-linear method's use of the 10-fold Cross-Validation (CV) methodology has resulted in not only significantly lower computational costs compared to previous methods, but also in highly encouraging classification accuracy results. The recommended ELM-based strategy's ACC, AUC, sensitivity, and specificity are 87.72%, 87.72%, and 87.72%, respectively.

Overall, the results demonstrated that nonlinear classifiers outperformed linear classifiers across all data sets. Extreme machine learning (ELM), convolutional neural networks (CNN), and RFs are among the nonlinear techniques used to diagnose thyroid nodules, yielding encouraging results (Xia et al., 2017; Zhu et al., 2016). Nonetheless, that earlier work only trained models with a tiny sample size and did not rigorously test the effectiveness of different classifiers. Furthermore, Palatucci and Mitchell (2007) discovered that most non-linear classifiers perform poorly when the number of features

in the data is significantly higher than the number of training examples, or when the data is of a high-dimensional nature.

2.4 Random Forest as The Non-Linear Classification Method

A popular machine learning approach for classification is called RF. Breiman's suggested technique, RF, classifies data using several decision trees (Breiman, 2001). A variety of decision trees are built using various random sampling and replacement subsamples of the dataset. It creates a "forest" from a group of decision trees that are frequently trained using the "bagging" technique in order to improve the outcome (Breiman, 2001; Liu et al., 2017; Savargiv et al., 2021). Instead of depending only on one decision tree, the RF makes predictions based on the majority of votes. It accomplishes this by adding up all of the decision tree forecasts. This classifier increases the anticipated accuracy of the dataset by averaging the output of several decision trees applied to various subsets of the dataset. The RF creation flowchart is displayed in Figure 2.2.

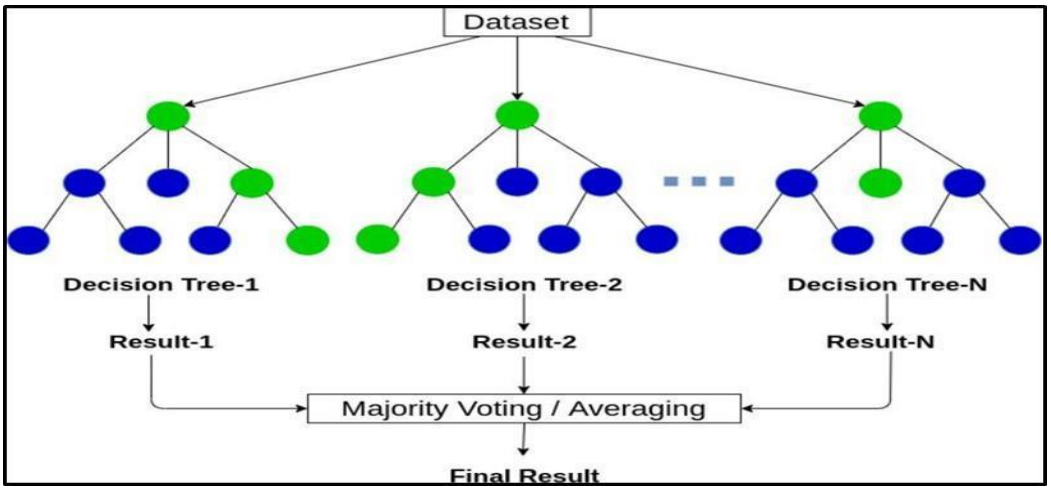

Figure 2.2 Construction of RF

Although no classifier is always the most accurate (Fernández-Delgado et al., 2014), RF is the top classifier among 182 classifiers (Dewi & Chen, 2019). The RF method is the best classifier and has the highest level of precision, according to a comparative examination of the Human Activity Recognition (HAR) dataset using machine learning techniques. Based on studies conducted by Breiman (2001), RF is a popular learner due to its versatility, ease of handling non-linearities, invariance to data

transformations, and minimal tuning required. Random Forest (RF) is a powerful model utilized for non-linear data processing in bioinformatics and genomics, using multiple decision trees to prevent overfitting. It excels with high-dimensional cancer genomics datasets, and hyperparameter fine-tuning can enhance results in complex cases.

2.5 Effects of High-Dimensional Data on the Random Forest Method

RF is a proven effective classification method. However, it faces challenges with high-dimensional data, as demonstrated in previous studies (Clarke et al., 2008; Xu et al., 2012). RF performs poorly with high-dimensional data or all data aspects, especially with data that has a low degree of dimensionality (Zebari et al., 2020). RF, like other classification methods, often faces challenges with high dimensions, particularly when dealing with high-dimensional data (Lin et al., 2017). First, the dimensionality curse causes models to overfit on training data, which reduces the model's capacity to generalize when there are several variables or features with high correlation in the classification model. Second, because much of the data is highly noisy, a model's robustness is essential. Third, a model's learning process may be incredibly slow if it is dimensionality sensitive. Fourth, the high-throughput data show significant variety and redundancy, making it challenging to extract relevant knowledge (Breiman, 2001). More closely related to the aforementioned issues are the RF issues with high-dimensional data. The first issue is that the curse of dimensionality will manifest itself in high-dimensional datasets, which have an excessive number of features. Note that strongly connected features frequently lower the projected variable relevance scores of correlated features in high-dimensional datasets, making RF less efficient at identifying the strongest features (Fiagbe, 2021). Highly connected trees have a significant negative impact on the efficacy of random forest models (Breiman, 2001).

Additionally, as the dimensions increase, the number of features also rises, frequently much more quickly, which implies that a significant amount of sparsity accompanies such high-dimensional features. Features may also be challenging to describe because there may be some association between many dimensions. RF will become overfitted as a result of this difficulty, which will impair its ability to generalize. The issue arises when randomly choosing features is one of the key steps

in creating an RF model, specifically when splitting data (bootstrap samples) to develop classifiers (decision trees). This stage also appears to be inappropriate for large-dimensional data, as these types of data frequently contain numerous variables that are irrelevant to the target variable. There is no criterion for choosing relevant variables in a broad random forest model. Thus, a uniform random sampling is used to sample all variables. Considering this, a split's randomly chosen splitting variables would often contain few or no relevant variables, especially when the proportion of actually useful variables is low. Due to their weak correlation with the target variable and low predictive power, uninformative variables, also known as noise variables, are challenging to forecast. Consequently, weak trees grow in the forest, which causes a severe drop in The RF model's classification ability (Amaratunga et al., 2008). The application of RF must be done carefully because RF parameters, specifically the number of features randomly selected at each node of a tree, must be considered. It must be carefully tuned to optimize its prediction performance, especially when the number of predictors p is significantly greater than the number of observations n. A classification model must have strong generalization capabilities. Such a model is expected to perform equally well on a different testing set, in addition to performing well on the training set. Table 2.2 illustrates the benefits and drawbacks of utilizing RF as a classification approach in general.

Table 2.2
The Temperature Information

No.	Advantages	Disadvantages
1	Capable of managing numerical and categorical data, and not affected by outliers and missing values.	When many trees are involved, it harms real-time forecasts. There may be a technical delay.
2	Typically, less prone to overfitting and have much better predictive ability than a single decision tree.	Since it builds numerous trees to combine the information, it is highly complex and necessitates significant processing resources.
3	A convenient algorithm, as it frequently utilizes the default hyperparameters to generate accurate prediction results.	When dealing with high-dimensional data, a model will be overfit and have low generalizability in terms of prediction performance.
4	Able to use a variable to indicate the contribution of each predictor.	

Source: Khadiev et al. 2019

2.6 Feature Extraction Method

A hybrid technique that combines feature selection and classification methods is introduced to address the issue of excessive features in high-dimensional data, thereby overcoming the curse of dimensionality (Hasan & Abdulazeez, 2021). The use of feature selection has significantly improved the efficiency of data mining and machine learning techniques by addressing the curse of dimensionality (Do et al., 2010). Correspondingly, machine learning techniques and their applications are enhanced by eliminating irrelevant and redundant features from data collection (Cai et al., 2018; Zhang, 2013). Feature selection enhances machine learning techniques by eliminating redundant features, improving classification performance, and reducing overfitting, High-dimensional datasets are particularly relevant.(Koç, 2021; Liu et al., 2017). The study highlights the benefits of feature selection in reducing data dimensionality, including improved data quality, faster calculations, improved algorithm performance, and simplified classification, all of which contribute to improved performance. There are three types of feature selection algorithms: supervised, non-supervised, and unsupervised (Janiesch et al., 2021), unsupervised (Farnaaz & Jabbar, 2016), and semi-supervised approaches (Cai et al., 2018). Supervised algorithms use labelled datasets to identify key features, while unsupervised methods use class label information. Semi-supervised strategies evaluate feature importance using both labeled and unlabeled data. Filter, wrapper, or embedding models are examples of supervised approaches (Venkatesh & Anuradha, 2019). Note that a filter model is designed to allow independent feature selection and model learning. Several filter-based strategies have been employed, including Information Gain, Relief Method, and Fisher Score Method.

They fail to consider how different characteristics interact when evaluating a feature. This highlights a failure in duplicate feature detection, leading to low stability and poor performance. The wrapper model trains the model using a constrained set of features. Through feature dependencies, it allows the model and features to communicate. The following wrapper models were investigated: Simulated Annealing, Genetic Algorithms, Greedy Forward Selection, and Recursive Feature Elimination (Chen et al., 2011). Hence, selecting features that perform well in terms of accuracy is a significant task for embedded models. The classification method incorporates the feature search process, making it impossible to isolate the learning process from the

feature selection process. Examples of embedded techniques include the Lasso and Ridge Regression algorithms. Therefore, unsupervised feature selection methods have been widely used.

Employed to eliminate irrelevant features through dimensionality reduction, they include Principal Component Analysis (PCA), Discriminant Analysis, and Feature Similarity (Solorio- Fernandez ´ et al., 2020). These methods consider data similarity without incorporating discriminative information into the data. PCA generates principal components from a dataset. To determine the most important principal components, the correlation between features is compared. It is a compression method that enables data sets with numerous interconnected features to be reduced in size, allowing the data to be represented with fewer variables (Nobre & Neves, 2019). By minimizing time and overfitting of the model, using variable correlation in this case produces characteristics that enhance algorithm performance. Since PCA does not limit how data can be used, it is not restricted to any specific dataset. Hybrid methods, where a feature selection method involving PCA is used before the classification process under RF, result in faster training and reduce the computational costs.

2.6.1 Principal Component Analysis

PCA was developed to explain variation in a set of multivariate data in terms of a set of uncorrelated variables. The primary goals of PCA are to analyze data and identify patterns that reduce the dataset's dimensions with minimal loss of information. It searches for K-n-dimensional orthogonal vectors, where K represents the number of relative characteristics. Combined into smaller sets with lower dimensions. Each set of attributes represents a direction that is perpendicular to the others as a unit vector. Principal Components (PC) are known to have the strongest significant vectors in terms of order of strength. The selection of an attribute does not take into account the least significant vectors. Instead, significant vectors may be taken into account. The data matrix of n observations on the p correlated variables x_1, x_2, x_p and the transformation of the xi into the p new uncorrelated variables yi. PCA is essentially based on the eigenvalue decomposition method. PCA, a popular technique for reducing dimensionality in other machine learning applications, is fundamentally based on eigenvectors. The PCA has achieved several notable successes in the field of genomic research. For instance, Rukhsar et al. (2022) attempted to use gene expression data to

divide patients with small, rounded blue cell tumors into different subgroups for various treatments in their analysis of gene microarray data. They started using PCA to lessen the dimensionality of the data. The Artificial Neural Network (ANN) analysis that followed properly categorized illness subtypes and identified the genes most related to the subgroups, utilizing the final 10 PCA components. Similar to this, Hsu et al. (2014) utilized PCA to divide gene expression data into three PCA components that could discriminate between different types of brain tumors and normal brain regions.

Since principal components are frequently linear combinations of all variables and loadings are typically nonzero, a disadvantage of PCA is the absence of sparsity of the main vectors. This makes it frequently challenging to apply in several situations where the fundamental components would be convenient if they were contained. Note that very few nonzero loadings. This results in the development of techniques for identifying sparse PCs, which explain the majority of the variance in the data.

2.6.2 Sparse Principal Component Analysis

A sparse representation exists when variables are related to just one or a few components (Guerra-Urzola et al., 2021). It has the benefits of simplifying interpretations and fixing the inconsistent estimated component loadings and weights in the high-dimensional situation (Lu et al., 2016). Unfortunately, the main vectors are not sparse because all of PCA's principal components are linear combinations of all the input attributes (Zebari et al., 2020). Due to the possibility of numerous zero loadings that are not significant, these components may be challenging to use. To address this issue, a new subfield of PCA, known as Sparse Principal Component Analysis (SPCA), has recently been developed. Sparse PCA identifies linear combinations that utilize the fewest available input qualities, resolving this issue (Greenacre et al., 2022). SPCA, which employs sparse loadings to modify PCs, overcomes these limitations. As with regularization methods in regression, SPCA approaches achieve this by inducing sparsity in the loading vectors. Alter PCA processes. For variables with multiple nonzero values, the analysis will yield only Sparse PCs. The benefit of this method is that it improves interpretability by lowering the number of loading coefficients. The emergence of Sparse PCA began with several strategies aimed at identifying only a few meaningful PCs for interpretation and understanding. PCA was utilized by Hsu et al. (2014) to identify risk variables for gastric and esophageal cancer. A primary

component was not considered significant if its loading coefficient was less than 0.2. Meanwhile, Zou et al. (2006) employed an algorithm to discretize all PC loading coefficients to values between -1 and 1. The algorithm is designed so that the first few components tend to be simpler than the later ones. Additionally, Hsu et al. (2014) examined 13 variable selection criteria to select a subset of the original variables, ensuring that the selected subset accurately reflects the PCs. Although these methods appear promising, they have several limitations, including difficulties in determining the threshold value and identifying the ideal discrete-valued loading coefficients. A significant amount of work has been done on sparse PCA, utilizing various PCA formulations and methods to achieve sparsity. They were created with the intent of using various types of penalty functions to reduce small, negligible estimations to zero. According to Min, Lee, & Yoon (n.d.), popular sparse PCA methods that exhibit high-dimensional statistical consistency fall into three categories: Variance Maximization (VM), Reconstruction Error Minimization (REM), and Singular Value Decomposition (SVD). All of them can be used to derive PCA.

Under the VM strategy, the L1-penalty function is incorporated into the PCA estimation process to quantify the degree of error. This function is known as the L1 penalty, defined as the sum of the absolute values of the error components. Models with L1 regularization can be sparse (i.e., have few nonzero coefficients). Some coefficients can be set to zero and removed. Meanwhile, the SPCA under the REM approach incorporates both L1 and L2 penalties during data training, where L2 denotes the sum of squares of the weights. The third SVD method identifies sparse PCs via regularized singular value decomposition, which addresses the low-rank matrix approximation issue. Only one study, conducted by Drikvandi and Lawal (2023) has applied a Sparse PCA variant within the VM framework as a feature selection method prior to RF and Gradient Boosting for classification of text data in the Natural Language Processing (NLP) field. Their experiment showed that Sparse PCA performed equally well as PCA in terms of accuracy and precision. On top of that, the advantages of Sparse PCA over PCA are such that the Sparse PC has a more straightforward interpretation and faster computational time to get the Sparse PC as compared to PCA.

2.7 Summary and Conclusion

Overall, most reviews of the literature found that nonlinear classifiers outperformed linear classifiers across all datasets. Among the nonlinear classifiers, RF is the best classifier among the 182 classifiers, according to Fernández-Delgado et al. (2014), even though no classifier has the highest accuracy in all circumstances. Due to its superior performance compared with other nonlinear classifiers, RF has been widely studied. RF is the best classifier among the 182 classifiers, according to Fernández-Delgado et al. (2014), even though no classifier has the highest accuracy in all circumstances. Among nonlinear classifiers, RF is widely used by researchers because of its strong performance relative to other classifiers. Although no single classifier consistently performs at the top level (Fernández-Delgado et al., 2014), research indicates that RF, among the 182 classifiers, is the most accurate. However, RF performs poorly on data with very high dimensionality. A hybrid strategy that combines feature selection with a classification method is introduced to address the problem of an excessive number of features that may be unimportant or suffer from the curse of dimensionality (Guerra-Urzola et al., 2021). PCA is a dimension-reduction method that can also serve as a feature-selection method. PCA has the disadvantage that its principal vectors are not sparse, as the PCs are typically linear combinations of all the variables, and the loadings are often non-zero. Hence, it is frequently challenging to apply the principle in numerous applications where it would be. It would be convenient if the principal components had very few nonzero loadings. By making the loading vectors sparser, Sparse PCA approaches alter PCA processes. Alternatively, the analysis will produce Sparse PCs only for variables with a limited number of nonzero values. Reduced loading coefficients, which improve interpretability, are a benefit of this strategy. Sparse PCA is easier to interpret and requires less computational time to produce than PCA, which is one of its benefits. Sparse PCA is also expected to enhance RF's performance when used in conjunction with it as a feature selection method.

2.7.1 Research Gap

High-dimensional data can hinder the performance of RF models when the number of features outnumbered the size of samples. It is known that RF leading to

overfitting and accuracy degradation from irrelevant or noisy features. Although RF has some feature selection capabilities, it is incapable to completely remove irrelevant or noisy variables. As a result, problems such as model overfitting may occur.

To overcome the dimensionality problem in RF. , utilizing PCA is often proposed to be used prior the implementation of RF. advanced techniques or preprocessing may be necessary to effectively identify key variables. PCA reduces dimensionality by creating orthogonal linear combinations of features, enhancing model performance and efficiency.

Sparse PCA, which is particularly effective in genomics and bioinformatics, selects a limited subset of variables for each component, thereby improving interpretability and computational efficiency in high-dimensional, low-sample-size settings. Combining Sparse PCA with Random Forest can enhance performance in high-dimensional genomic data by addressing issues like noise sensitivity and overfitting. Modified RF methods, which utilize SPCA and subsampling, may improve biomarker identification. Future research should aim to optimize genomic classification frameworks to reduce overfitting and enhance performance in biological contexts.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the methodology used in the study. The chapter begins with the explanation of the overall research framework in Section 3.2. Next, the research framework, which involves simulated data and real data, is presented in Section 3.3. The explanation of how the feature extraction method using Principal Component Analysis (PCA) and Sparse PCA is implemented before the Random Forest (RF) is explained in Section 3.4. The comparison of the performance of three Sparse PCA models involving simulated data is also included. Consequently, the classification process of using RF is described in Section 3.5. The next step is to Section 3.6, where the best Sparse PCA and RF are verified on simulated datasets. Finally, the performance criteria of RF are illustrated in Section 3.7.

3.2 Research Framework

Figure 3.1 displays the research framework of this study. First, the study begins with the data generation phase, which involves collecting and synthesizing relevant datasets. Synthetic data generation is a method used in research and machine learning to address challenges such as data scarcity, class imbalance, and privacy concerns. High-dimensional datasets require specialized analytical methods to manage their complexity and uncover underlying patterns. enabling comprehensive experimentation without relying solely on sensitive data. The advantages include ample data for model training, improved accuracy by addressing underrepresented groups, enhanced privacy because synthetic data lacks real-world identifiers, and cost and time efficiency. Techniques for generating synthetic data include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), rule-based systems, and statistical modeling. This practice is prevalent in fields such as computer vision, Natural Language Processing (NLP), and medical research, promoting the development of ethical Artificial Intelligence (AI) and facilitating data-driven decision-making using

Monte Carlo Markov Chain (MCMC) simulation. The next step involves the approach described in the first step: direct classification using RF without feature extraction.

Synthetic data refers to intentionally created datasets utilized in data science for various purposes, including research and machine learning. It helps to mitigate problems such as data scarcity, class imbalance, and privacy issues. The MCMC simulation is highlighted as a powerful approach for generating these datasets. At the same time, RF techniques support direct classification, reducing costs associated with data scarcity.

Simulated data provides a balanced dataset for developing robust models, ensuring an even class distribution, facilitating secure data exchange, and enabling controlled experimentation. It improves model behavior understanding and evaluation, especially in RF classification. Synthetic data enhances generalization while ensuring data privacy, making it a common term in the fields of data science and machine learning. The "curse of dimensionality" refers to the challenges of analyzing, organizing, and modeling high-dimensional data, which can lead to sparsity, poor performance, increased computational complexity, and a higher risk of overfitting. Meanwhile, synthetic datasets can indirectly exclude sensitive information. Feature extraction, a preprocessing procedure, reduces dimensionality, improves model performance, and reduces overfitting. In classification tasks, reducing dimensionality and minimizing feature redundancy are vital, as they enable more effective handling of high-dimensional issues without altering the problem's inherent dimensionality. This study employed a real genomic dataset for achieving objectives 1 and 2, meanwhile simulated dataset set for achieving objective 3.

The next step is to compare the performance of RF under the two aforementioned scenarios: a hybrid of standard PCA with RF, and hybrids of Sparse PCA (incorporating the Variance Maximization (VM), Reconstruction Error Minimization (REM), and Singular Value Decomposition (SVD) approaches) with Random Forest. Consequently, it is necessary to verify the performance of the best hybrid model using real high-dimensional data. Sparse Principal Component Analysis (SPCA) extends standard PCA by incorporating an L1-norm sparsity penalty, yielding components that are more interpretable in high-dimensional settings such as genomics. By producing loadings with fewer nonzero variables, SPCA facilitates biological insight while preserving key variance structure. When combined with Random Forest classifiers for cancer data, SPCA-enhanced features improve classification

performance and address the interpretability limitations of traditional PCA. These benefits are consistent with findings from SPCA variants such as Regression EM (REM), Sparse SVD, and Variable Matching (VM)

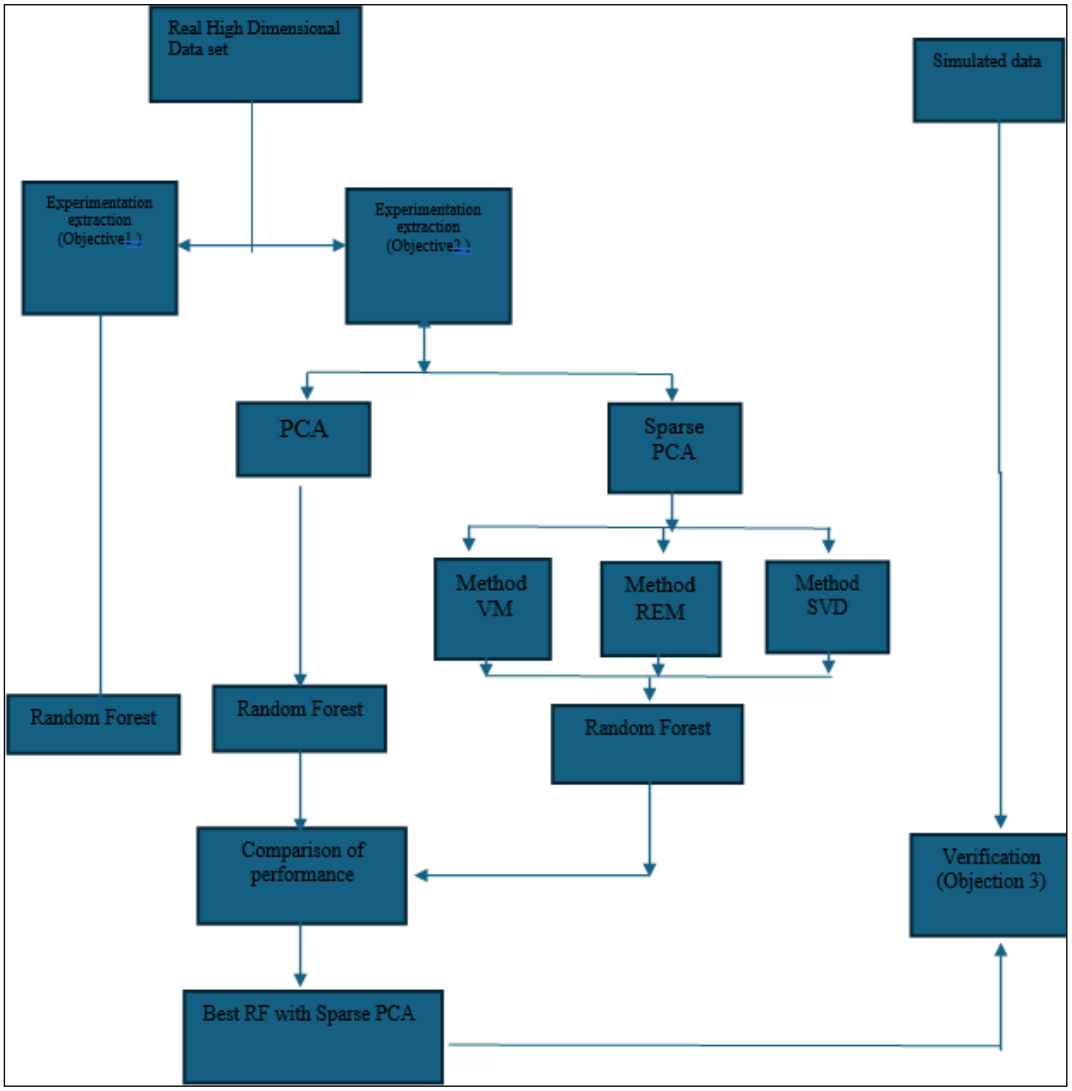

Figure 3.1 Research Framework

3.3 Research Data

Two types of data will be used in this study: simulated and real. Real data are used to derive empirical results on the performance of statistical approaches. Meanwhile, the simulated data are used to evaluate and compare hybrids of three types of Sparse PCA with RF across a range of performance criteria. The framework explains the relationship between input and target features, data types, target categorization groups, and study scenarios. It utilizes real data to validate statistical procedures and

compare methodologies. Note that simulated data are used to test hybrid models and to analyze them under various conditions. Class imbalance assessment evaluates model accuracy, interpretability, computational efficiency, and the stability of feature selection.

3.3.1 Data Simulation

The performance of RF alone, with PCA, and with Sparse PCA will be evaluated using simulated data generated by an MCMC Simulation for comparison purposes. The study evaluates hybrid methodologies for high-dimensional data, with a focus on feature selection, redundancy reduction, and enhancement of predictive accuracy. Findings indicate that these methods reduce errors and enhance model generalizability, particularly in bioinformatics and machine learning applications that use simulated or benchmark datasets. The simulation accommodates multicollinearity and varying sample sizes, utilizing the R packages Simstudy and Simdata for data modeling and generation. Simulation studies require multiple iterations to address sampling variability, which is constrained by computational resources and design, and yield variable outputs that depend on the simulation framework. The simulated data are used as training samples to train PCA, Sparse PCA under feature extraction methods, and RF as the classification model. The following data matrix $X = (X_1, X_2, \dots, X_p)$, where $X_i \in Rn$, will be generated as follows:

$$Xi=\mu+ ei,\ i=1,2,\dots,p \tag{3.1}$$

The document discusses the creation of an RF data matrix using simulated data, focusing on extraction methods such as PCA and Sparse PCA. It covers the analysis of the matrix via covariance matrix computation and the extraction of eigenvalues and eigenvectors, leading to classifier training using compressed features. The notation indicates that vector μ is multivariate normal with zero mean and a diagonal covariance matrix. In contrast, vector e has a covariance matrix of $\emptyset$I. This supports the understanding of multivariate normal distribution, determining expected values and correlations via the covariance matrix. Key components of the probability density function include the mean vector, vector space dimension, and covariance matrix, which are vital for simulating jointly standard random variables.

$$\mu \sim MVN(0,\sigma 2I\ ),\ e \sim MVN(0,\ \emptyset I\ ),\ v\ n\ i\ n. \quad (3.2)$$

μ is a random vector following a Multivariate Normal distribution (MVN) with mean vector 0 and covariance matrix I, signifying independent random variables that each have a univariate normal distribution with mean 0 and variance 1. The error vector e also follows an MVN distribution with zero mean and covariance matrix Σ, indicating that both μ and e are MVN distributed vectors. μ has uncorrelated components of unit variance while e is described by a general covariance matrix Σ.

A random vector v denotes the MVN distribution in Rn, where n is the number of variables. Its covariance matrix, represented as ∑ or Σ, is diagonal with variance σ^2 and a mean vector of zero. This structure simplifies analysis, with uncorrelated components indicated by equal variances. The MVN plays a crucial role in statistical methods, including hypothesis testing, PCA, Bayesian inference, and modeling linked data, serving as a framework for representing prior distributions and error structures in Bayesian statistics.

i) Datasets with three sets of n and p dimensions: 1) n = 200 and p = 1000 (small set), 2) n = 200 and p = 2000 (medium set), and 3) n = 500 and p = 2000 (big set).
ii) Level of multicollinearity ($\rho 2$) = 0.5 (low correlation), 0.7 (medium correlation), and 0.9 (high correlation).
iii) Distribution = Standard Normal distribution: N (0, 1).

The selection of datasets by dimensions-1) n = 200, p = 1000 (small), 2) n = 200, p = 2000 (medium), and 3) n = 500, p = 2000 (large)-provides a balance between data complexity and analysis feasibility. A sample size of n = 200 is considered moderate for small- to medium-sized datasets. In contrast, a sample size of n = 500 is more suitable for larger datasets. High-dimensional conditions with p = 1000 address complexity, and p = 2000 further increases it for larger datasets, enabling evaluation of machine learning techniques across multiple scales while balancing computational costs and statistical power. The study examined the statistical power, interpretability, and computational feasibility of small-to-medium-sized variable sets by benchmarking correlation performance across varying signal strengths. It concluded that larger samples are necessary to detect lower correlation coefficients (R). In comparison,

smaller samples can effectively identify stronger correlations. Note that increasing the sample size improves estimation accuracy and reduces the width of confidence intervals, thereby yielding more reliable results. The findings support the use of larger datasets for replication, with correlation thresholds informed by previous research.

Based on the above condition, the researcher will evaluate the performance of the proposed hybrid prediction models across varying levels of multicollinearity, using different degrees of correlation among the variables. Additionally, the effects of varying sample sizes should be examined for the proposed hybrid prediction models, which include small-, medium-, and large-sample models.

3.3.2 Benchmark Data

One set of genomic data, involving gene expression in lung cancer and Breast cancer cases, was used to evaluate the performance of hybrid Sparse PCA and RF models. The data comprise p = 998 genes and n = 442 observations. The response variable categorizes tumor tissues and normal tissues with 2000 predictors. The response variable categorizes patients with lung or breast cancer. The data sets presented are high-dimensional cancer gene expression data used for model validation. Hybrid approaches combining Sparse PCA and RF models are used to differentiate between lung and breast cancer.

Benchmark data in machine learning and bioinformatics is vital for evaluating algorithms, particularly in high-dimensional genomics such as cancer classification. Key datasets include TCGA datasets for lung and breast cancer, using methods like Sparse PCA and Random Forest classification. Incorporating real-world data to address biological sample noise, these benchmarks are important for platforms like Kaggle's cancer genomic challenges and the UCI ML Repository's Breast Cancer Wisconsin dataset.

3.4 Feature Extraction Methods

In order to improve the classification performance of RF, PCA, and Sparse PCA are considered to reduce the number of features while retaining as much of the information as possible. Feature extraction is crucial in machine learning for improving model performance and accuracy. RF, PCA, and Sparse PCA are methods for feature

reduction. RF ranks features based on their importance, handles high-dimensional data, and captures complex interactions without requiring explicit specification. PCA is a technique that reduces dimensionality by transforming features into PCs, retaining maximum variance. It is computationally efficient and effective for datasets with correlated features. Sparse PCA introduces sparsity constraints to maintain interpretability, thereby balancing dimensionality reduction and feature interpretation, and is particularly effective for high-dimensional datasets. Sparse PCA involves applying sparsity constraints during computation, selecting sparse PCs, and combining techniques such as RF, PCA, and Sparse PCA to enhance classification performance and accuracy.

3.4.1 Principal Component Analysis

PCA is a statistical method that employs orthonormal transformations to convert the observed correlated data into a set of linearly uncorrelated data, known as PCs. PCA's goals, according to Guerra-Urzola et al. (2021), are to (1) eliminate the most relevant information from the data table; (2) minimize the size of the data set by keeping just this significant information; (3) simplify the description of the data. Simplify data description by focusing on essential components, such as source, type, purpose, key features, and size, and (4) evaluate the observation structure and variables. To fulfill these objectives, PCA measures new variables called principal components, which are derived as a linear combination of the original variables.

3.4.1.1 Mathematical Formulation of PCA

The following mathematical formula:

Firstly, calculate the $i\ th$ principal component Y_i (given by:

$$\begin{aligned} Y1 &= a'_1X = a11\ X1 + a12\ X2 + \ldots + a1pXp, \\ Y2 &= a'_2\ X = a21\ X1 + a22\ X2 + \ldots + a2pXp, \\ Yp &= a'_pX = ap1X1 + ap2X2 + \ldots + appXp. \end{aligned} \tag{3.3}$$

Yi = function of linear combinations for every variable

Xp = Features or variables

The equations provide formulas for calculating each principal component. The first principal component (Y_1) explains the most variance in the data. The second principal component (Y_2), which explains the second most variance, etc. PCA standardizes data, calculates the covariance matrix, computes eigenvalues, and projects the original data onto new axes. Hence, we calculate:

$$Var\ (Yi) = a'i \sum ai = \lambda \text{ and } Cov\ (Yi, Yk) = a'i \sum ak = 0, \quad (3.4)$$

where i, k=1,2,… p and i≠ k.

λi = eigenvalue corresponding to each principal component

PCA uses formulas to calculate the variance and covariance of PCs. The eigenvalue λ_i\lambda_ λ_i represents the variance of the PC Y_i, which is a linear combination of original variables weighted by the vector a_i. When i≠ki T ki =k, the PCs Y_i and Y_k are uncorrelated, and their variance is equal to the corresponding eigenvalue. This result reduces dimensionality by projecting data onto orthogonal directions of maximum variance.

3.4.1.2 Assumptions of PCA

Note that PCA requires certain presumptions. These include data linearity, where variables combine linearly, and significant variability, where high-variance components are prioritized due to their importance. They are often ignored as noise. To ensure reliable PCA results, use the same measurement scale for each variable and normalize the data if different scales are used. A large sample size of at least 150 observations or a 5:1 ratio is required. Consequently, identify and address outliers to prevent bias and ensure dependable results. PCA effectively reduces dimensionality by converting correlated variables into uncorrelated components, ensuring the most relevant data remains from the original dataset. PCA assumes linearity, significant variance, importance of variance, and consistency of measurement scales. It identifies the essential underlying structure in components with higher variance, while low-variance components are often overlooked. To produce valid PCA findings, a sample size of 150 or 5 observations per variable is recommended. Outlier management is crucial to avoid bias. PCA assumes that the PCs are uncorrelated, reducing

dimensionality while maintaining the original dataset's key structure. The assumptions of PCA are as follows:

i) No significant outliers.

PCA assumes linearity and the absence of significant outliers, thereby reducing dimensionality by capturing variance in fewer components. It works best with a Gaussian distribution and assumes that significant differences reflect variance. For datasets with notable outliers, robust versions or techniques are recommended.

ii) There is a linear relationship between all variables.

The PCA assumes linearity in data representation, assuming that linear combinations of the original variables reflect the data's structure. It also assumes that there are no significant outliers, as outliers can cause inaccurate results. PCA minimizes dimensionality by capturing shared variance among variables with high correlation. Note that principal components are orthogonal and uncorrelated, capturing distinct variation. PCA assumes that maximum variance directions represent significant information, converting the data into a new coordinate system.

iii) The data correlate with variables.

PCA assumes that variables are correlated and seeks to reduce dimensionality by finding orthogonal components. It assumes linearity, identifying principal components as linear transformations of initial variables. It also assumes that there are no significant outliers, as outliers can skew the principal components. PCA reduces dimensionality by capturing shared variance among variables with high correlation. It captures distinct variation in orthogonal, uncorrelated primary components. The approach assumes that the maximum variance direction indicates the most significant information in the dataset. PCA requires correlation between variables to reduce dimensionality. It simplifies datasets by identifying principal components that cause variance in highly correlated variables. Without correlation, PCA may not offer significant dimensionality reduction gains.

3.4.1.3 Method of Determining the Number of Principal Components (PC)

The PCA is a method for reducing dimensionality, balancing variance preservation and noise reduction. It highlights the Scree Plot/Elbow Method, which involves plotting Principal Components (PC) eigenvalues against rank, to determine the ideal number of primary components. However, this method is subjective and relies on visual assessment. Programs like find PC use computational techniques to detect elbow points, while statistical and Bayesian techniques identify significant PCs. Resampling and cross-validation evaluate the generalizability and interpretability of component selection, with interpretability being key for biological applications. PCA offers benefits such as domain-specific value and categorized patterns. The PCA method, which maintains variation and flexibility, has drawbacks such as sensitivity to outliers and the linearity assumption. Variants such as Kernel and Robust PCA have been developed to address these issues. The study's objectives determine the method and component selection approach**.** This study employed the following method for determining the number of PCs.

i) Choose principal components with eigenvalues more than 1.

   The Kaiser Criterion, also known as the Kaiser-Guttman Rule, is a statistical technique used to select PCs with eigenvalues larger than one, ensuring significant components account for more variance than a single original variable in PCA. The Kaiser Criterion is a popular method for determining the number of PCs in exploratory data analysis. It can preserve components affected by random noise and is often combined with other strategies for reliable dimensionality reduction.

ii) The proportion of total variance must be greater than 70%.

   PCA calculates principal components by determining the percentage of total variance explained and selecting the fewest PCs with a cumulative explained variance exceeding a specified cutoff. The Proportion of Variance Approach is a method used in PCA to calculate the number of PCs. It involves determining the explained variance, calculating the cumulative explained variance, setting a threshold, and selecting the fewest PCs whose cumulative explained variance

exceeds a predetermined cutoff. Benefits reduce dimensionality while preserving dataset variability. PCA is a statistical technique designed to identify the number of PCs to retain by examining the percentage of total variance explained. It is essential for situations where significant data variability is necessary for analysis or modeling. The method assesses the overall variance represented by each component, expressed as a percentage from 0% to 100%, which helps evaluate the significance of various data components, including eigenvalues.

$$\text{Proportion} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \tag{3.5}$$

It utilizes PCA to calculate the proportion of variance explained by each PC. This helps assess the relevance of each component and determine the most significant contribution to overall variance. The formula for the proportion is λ_k, $\lambda_1+\lambda_2+\cdots+\lambda_p$, validating PCA. Results are highlighted by emphasizing the most significant components and their variance, facilitating the decision on which major components to retain.

3.4.2 Sparse Principal Component Analysis

Sparse PCA is a method for extracting features from datasets that is used in statistical analysis (usually multivariate). SPCA is a variation of the traditional PCA algorithm that reduces the dimensionality of the input feature vector by introducing sparsity in the structure of the input features. The difficulty with traditional PCA is that PCs are linear combinations of all input features. The SPCA solves this issue by locating linear combinations that use only a few input features. A parameter α determines the level of sparsity in SPCA, which can be categorized into three approaches: VM, REM, and SVD. Each SPCA variant utilizes α differently, typically in penalty terms that encourage sparse loading vectors.

$$\text{Maximize a } a' \sum xa - \alpha \|a\| 1. \tag{3.6}$$

The optimization process optimizes explained variance by focusing on sparse loadings, using the VM approach. A sparsity-promoting maximization problem in convex optimization seeks to maximize a specific function Maximize a a' ∑xa−α‖a‖1. The a method combining a fidelity term with an L1 penalty to minimize small coefficients, utilizing a soft-thresholding operator that zeros coefficients below a threshold (c_i ≤ α). This approach is applied in Sparse PCA for cancer genomics, tackling high-dimensional challenges through proximal gradient methods and Lasso duals for feature selection in lung and breast cancer datasets, with α optimized via cross-validation to improve accuracy and sparsity.

$$\text{Max } 'a' \sum xa - \alpha\|a\|1. \tag{3.7}$$

The REM-SPCA employs a sparsity penalty on the loading matrix B to minimize reconstruction error. Sparse regression models, such as Lasso, utilize L1 norm penalties to optimize coefficients and promote sparsity, key for convex optimization in high-dimensional datasets. This strategy reduces overfitting and boosts accuracy while enabling feature elimination, relevant in Sparse PCA. Methods like Random Forest and PCA extensions adopt these techniques for feature selection in cancer genomics, improving sensitivity and accuracy in lung and breast cancer data through proximal gradient optimization and parameter tuning.

$$min_{A,B} \ \|X - XBA'\|^2 F + \alpha\|B\|1. \tag{3.8}$$

The data matrix X, loading matrices A and B, and the Frobenius norm ‖.‖F, absolute values in B, and loading sparsity α are all crucial elements in this analysis. The Sparse SVD Approach is a method that constrains the sparsity of a single vector during decomposition. Regular matrix analysis aims to minimize reconstruction error using the quadratic Frobenius criterion and the L1 scatter penalty on matrix B. This approach, suitable for scattered low-rank approximations and bilinear models, aims to optimize B for feature selection and reduce classification noise in datasets such as breast and lung genomes, thereby enhancing scattered principal component analysis (Sparse PCA). Approximate gradient regression is used to manage the trade-offs between sensitivity and accuracy, alternating between B and A, with α being optimized through cross-validation.

$max_{U,V}$ ‖X−XBA′‖^2 F+ α‖B‖1 such that ‖U‖1 ≤ α1 , ‖U‖1 ≤ α2. (3.9)

The algorithm uses α1,α2 to control sparsity, enforce fewer nonzero loadings per component, and utilizes α in each equation. The algorithm improves low-rank approximations for outlier detection by optimizing the squared Frobenius norm with (\l_1\) penalties and matrix constraints to promote sparsity. It enhances adversarial robustness in genomics and cancer analysis, akin to Sparse PCA, by balancing sparsity and reconstruction errors. Implementation can be done using ADMM or proximal gradient methods in Python or R.

3.4.2.1 Variance Maximization (VM) Approach

The approach attempts to reduce the number of dimensions in which data points are projected (data reduction) while maintaining as much of the variety of the original sample points as possible. Mathematically, X is assumed to be the original data that is represented by a dimension of n samples with a p-variable matrix, that $V1$ is the vector of the first loading coefficient referring to X, and that XV_1 is the inferred PC. The following expression can be used to indicate how to maximize the variance of X, which represents the initial matrix of data. Usually, it is organized as an n×pn×p matrix, where pp is the number of variables (features), and n is the number of samples (observations). Each row represents a sample, while each column represents a variable or characteristic in the dataset. For instance, the input data points utilized for analysis in PCA or linear regression are contained in X. V: The covariance matrix obtained from X is frequently referred to as V in this work. The variance and correlations between the variables in X are captured by the covariance matrix V. Finding directions (principal components) that maximize the variance in X, for example, is known as variance maximization in PCA. This is theoretically expressed as optimizing a function involving V. Accordingly, the VM technique uses X as the input dataset and V, which is derived from X, to carry out optimization tasks linked to variance, such as regression or PCA.

Max $_{v1}$ (V_1,X,XV_1) subject to $V_1,V_1= 1.$ (3.10)

Equation (3.6) PCA seeks to identify the direction of maximum data variance while adhering to a unit-norm constraint on the loading vector. This optimization

problem involves determining the feature space direction characterized by the leading eigenvector of the covariance matrix of a data matrix X. The method applies to both traditional and Sparse PCA formulations for dimensionality reduction, particularly in high-dimensional datasets such as gene-expression data. A key aspect is the formulation in Equation 3.6, which defines the optimization for the first principal component. The objective function max (XV) aims to maximize the product of the variables X and V, subject to normalization or restriction requirements that guide feasible solutions. Indicate unique quantities, such as resistance or time in physics or engineering, or resources or production-influencing variables in economics or operations research, with more accurate interpretation provided by the study background.

$$\text{Max } v_1(V_1X, XV_1) + \boldsymbol{\lambda}_1 \|V_1\|_1 \quad \text{subject to } V_1,\ V_1=1 \tag{3.11}$$

The optimization formulation considers two variables: X and V_1, a vector often limited or normalized. The objective function aims to maximize terms involving V_V_1, X, and the product V_1X. The mathematical structure is explicit optimization, with the primary variable being a vector. The sparsity-inducing penalty promotes interpretable solutions, while the limit ensures V_V_1 is a unit vector or normalized. Contextual interpretation of X and V_1 is crucial in various fields. In machine learning, the V_11 refers to weights or loadings, while X represents the data matrix. The term $\lambda_1\|V_1\|1$ promotes sparsity for interpretability. The functional forms and interpretations of $f(V_1$, X) vary depending on the domain and application. Sparse PCA maximizes projected variance with an extra sparsity penalty, ensuring that trivial vectors are avoided and solutions scale correctly.

3.4.2.2 Reconstruction Error Minimization (REM)

Under the REM approach, two aspects of adjustments are introduced. First, the product of the loading coefficient matrix, V'V, is adjusted to become two matrices (A, B). Consequently, an L2 term, which represents the sum of squares of the norm penalty on B, is added. The two procedures transform the PCA into:

$$\text{Min}_{A,B} \sum_{i=1}^{n} \| x_i - AB'x_i \|^2 + \lambda \sum_{i=1}^{k} \| B_j \|^2 + \sum_{j=1}^{k} \lambda_j \| B_j \|, \quad (3.12)$$

where A and B represent a p×k matrix,

$$\| B_j \|^2 = \sum_{j=1}^{p} (\sqrt{B_{ij}})^2, \quad (3.13)$$

and $\boldsymbol{\lambda}$ is the parameter's penalty. It is probable that the phrase "λ" refers to a penalty parameter, which is frequently employed in Ridge Regression. To prevent the model from overfitting, this parameter regulates the regularization strength. The L2 norm of the coefficients is typically added as a regularization term to the loss function in Ridge. Regression, which helps shrink the coefficients toward, but not quite to, zero.

The paper discusses p×k matrices A and B, which are likely used in matrix factorization or decomposition procedures. A might represent user-related attributes in recommender systems, while B might represent another factorized matrix. The formula for determining B's norm is $\|B\|^2$, possibly connected to its Frobenius or spectral norm. These matrices are part of a factorization setup that resembles the original matrix. B is solved as the Ridge regression problem and obtains jth loading as $Vj = Bj \, .\|Bj\|$. Ridge regression is a type of regularized linear regression in which the L2-norm of the coefficients is used to determine the penalty that is added to the loss function. Usually, the Ridge regression loss function is written. Sparseness imposition requires the addition of an L1 function normality to B to obtain sparse loadings. Equation (3.6) then becomes

$$min \sum n \| x - AB, x \|_2 + \lambda \sum k \, 2 \sum k\lambda \, \| B \|. \quad (3.14)$$

Subject to $A,A = Ik$,

Where : A common $\boldsymbol{\lambda}$ is used for all PCs, but different $\boldsymbol{\lambda}$, s are allowed for penalizing the loading of different PCs. question (3.14) minimizes the reconstruction error while simultaneously incorporating ℓ_2 and ℓ_1 regularization to improve the stability and sparsity of the principal components, which is especially beneficial for high- dimensional genomic data.

3.4.2.3 Singular Value Decomposition (SVD)

Under the SVD approach, the PCs' extraction is done through finding the solution to a low-rank matrix approximation problem. According to Montanari and Venkataramanan (2021), low-rank approximation is a method for recovering the "original" (the "ideal" matrix before it was messed up by noise, etc.) low-rank matrix. It discovers the most consistent matrix (in terms of observable entries) with the current matrix and is low-rank enough to be utilised as a close approximation to the ideal matrix. Another perspective is that SVD determines the solution through decomposing a matrix into subcomponent matrices. Let the SVD of X be X=UDV'. Where U is an n by K orthogonal matrix, while D is a K–by–K diagonal matrix, assumed to be ordered so that d_1,>d,>…>d_k. The estimated values of U, D, and V can be formulated as follows:

$$\text{Min U,D,V} \|\boldsymbol{X}-\text{UDV}\|^2\boldsymbol{F} \text{ subject to U}'\text{U}= 1_k \tag{3.15}$$

A L1–norm penalty is incorporated into 3.10 to obtain sparse loadings. Therefore, the penalized function turns out to be:

$$MinU,D,V \|X - UDV'\|2 + \sum k \text{ j=1 } \lambda \|V\| \text{ Subject to } U'U = Ik \text{ and } V'V = Ik, \tag{3.16}$$

where λ the penalty parameter for each component.

3.5 Random Forest Classification Model

RF is an ensemble learning method that improves classification accuracy by merging multiple decision trees. Each tree is trained on a different bootstrap sample and employs a random subset of features, adding randomness to the model. This technique relies on majority voting for predictions, which enhances resilience to overfitting and is particularly effective for complex, high-dimensional datasets, such as those in genomic research. RF can manage missing data, assess variable importance for feature selection, and work with both continuous and categorical data. Its use of bagging and random feature subspaces leads to strong generalization and minimized overfitting risks.

3.5.1 Random Forest Algorithm

The RF construction process involves various procedures, including developing an algorithm for reliability analysis. The process involves using different bootstrap subsets from the training dataset to generate decision trees using the sampling with replacement method. Note that each tree is formed in an RF model, and the optimal split point is determined by entropy and information gain calculations. The algorithm iteratively produces a tree until it reaches leaf nodes, obtaining the prediction value by averaging values. The hyper-parameter tuning process adjusts parameters like tree number, splitting criteria, accuracy, and least squares, considering factors like depth, pruning, confidence, leaf size, split size, and voting processes. The RF model will be tested on the remaining 20% of the dataset, repeating the training phase, and the projected ozone concentration will be created by averaging prediction values over decision trees. The RF algorithm is an ensemble learning technique used for classification, regression, and other applications. It generates predictions by building multiple decision trees during training and combining their results, reducing overfitting and increasing accuracy through the use of randomness in feature selection and data sampling. The process involves selecting features at random for each tree, constructing each decision tree using optimal split parameters, and building 64-128 trees. After construction, individual trees receive test data to produce predictions, resulting in a more diverse and less correlation-prone tree. RF accuracy enhances precision and dependability by combining forecasts from multiple trees through majority vote and average calculation for classification tasks. RF is a robust, scalable machine learning algorithm used for classification and regression problems, combining durability, flexibility, and simplicity. It effectively manages missing data and scales well with complex datasets. On the other hand, PCA is a fundamental mathematical technique that transforms data into a new coordinate system with maximum variance directions. It involves computing the covariance matrix, extracting eigenvectors and eigenvalues, and utilizing these eigenvectors to identify orthogonal directions that reflect significant data variability patterns. PCA applies mathematical ideas algorithmically to computing processes. The PCA approach simplifies data analysis by uniformly calculating the covariance matrix, applying SVD, and selecting the top eigenvectors. This computation technique reduces dimensionality, extracts features, and visualizes data,

illustrating how mathematical ideas from statistics and linear algebra are applied to real-world problems.

3.6 An Enhanced Random Forest Approach

This hybrid method combines RF classification with dimensionality-reduction techniques, specifically PCA and SPCA, to improve performance on high-dimensional genomic data. PCA reduces redundancy and noise, improving RF efficiency, while SPCA enhances interpretability and feature selection. The hybrid method employs Sparse PCA with a VM to optimize features, improving the signal quality by reducing noise. Furthermore, integrating SPCA with REM enhances feature selection, yielding a robust feature set that improves the performance of the RF classifier. The combination of SVD and SPCA is utilized for improved low-rank data representation, capturing significant patterns and complex relationships, ultimately enhancing accuracy, interpretability, and model stability in high-dimensional datasets.

3.7 Model Verification

Figure 3.2 illustrates the procedure for selecting and validating the optimal hybrid model, which combines SPCA and RF, based on performance metrics. The process begins by identifying the optimal SPCA type and RF combination for high-dimensional gene-expression data. The model is chosen based on accuracy, precision, sensitivity, recall, and the F1-score. Correspondingly, the model is compared based on its robustness, correct case identification, and overall performance. Discrimination using metrics like recall+precision2·Recall·Precision. It is validated using benchmark gene-expression datasets for high-dimensionality and relevance. This method improves the model's credibility and reproducibility, especially for high-dimensional pattern recognition tasks in biological data analysis. Evaluation of real-world accuracy includes debugging, data flow analysis, and testing methods. It highlights comparison testing and consistency checks in high-dimensional models, especially in bioinformatics, where Sparse PCA must meet theoretical standards before use on cancer datasets. Evaluation entails data flow diagrams, accuracy assertions, and input sensitivity analysis for models like Random Forest.

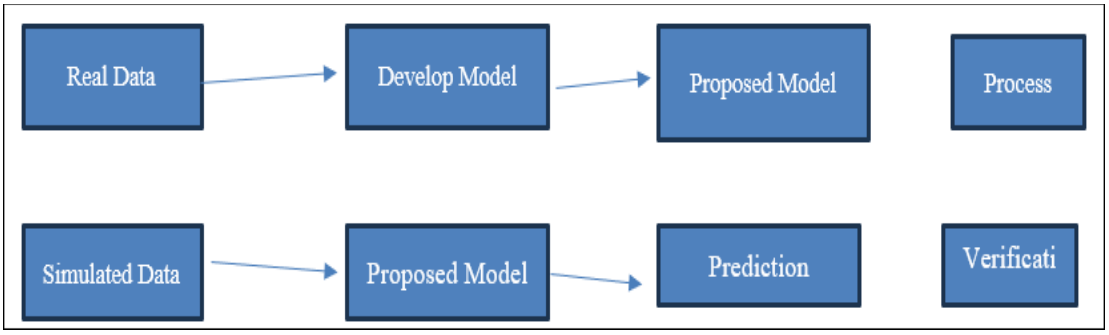

Figure 3.2 Model Verification

3.8 Model Performance

Three performance indicators are listed below to validate the RF method by assessing classification accuracy. The accuracy instruments in the performance models include the Coefficient of Determination (R^2), and the error instruments include the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). Three essential criteria are frequently employed to assess the RF method's effectiveness for classification or regression tasks:

3.8.1 Determination Coefficient (R^2)

The model's ability to explain the variance in the target variable is measured by the coefficient of determination, also known as R^2. This statistical measure measures the percentage of the dependent variable's variance that can be predicted from the independent variables. Regarding RF models, a higher R2 score indicates better predictive accuracy. R^2 values near 0 denote subpar performance, while values closer to 1 show that the model is successfully capturing the variability in the data. The coefficient of determination (R^2) indicates the percentage of variance in a dependent variable that is explained by independent variables in a regression model. defined as $R^2=1-\frac{SSres}{SSt0t}$ R2, ranging from 0 to 1, The coefficient of determination (R^2) indicates $SSt0t$ the percentage of variance in a dependent variable that is explained by independent variables in a regression model. defined as $R^2=1-\frac{SSres}{SSt0t}$. R^2, ranging from 0 to 1, R^2, the square of the Pearson correlation coefficient, indicates explanatory power and varies in effect sizes per Cohen's standards: small (0.01), medium (0.06), and large (0.14). While a high R^2 suggests a good model fit, it may also indicate

overfitting and does not imply causation. In genomics, R^2 is crucial for evaluating model fit and selecting features in cancer outcome predictions using high-dimensional data, with adjusted R^2 correcting biases in sparse PCA integrations.

3.8.2 RMSE, or Root Mean Square Error

The square root of the average squared discrepancies between expected and actual values is measured by the error metric known as RMSE. It provides information about the degree of prediction error. Given the squaring, RMSE penalizes greater errors more severely than smaller ones. When evaluating models that involve significant departures from actual values, it is constructive. Root Mean Square Error (RMSE) measures prediction errors in regression by evaluating residuals' distribution against the regression line. It is the square root of the mean of squared differences between observed and predicted values, with a lower RMSE indicating better model performance. RMSE is particularly important in machine learning contexts with high-dimensional data, such as cancer genomics, as it focuses on reducing larger errors and enhancing classification accuracy.

3.8.3 MAE, or Mean Absolute Error

The MAE between expected and actual values is computed. Since MAE does not square deviations, it is resilient to outliers, in contrast to RMSE, which considers all mistakes equally. It helps comprehend overall forecast accuracy, as it offers a clear explanation of the average number of mistakes.

3.9 Summary

Overall, based on the specified methodology, the following objectives are targeted for achievement using the following methods.

Table 3.1
Summary of Objectives of Study

No.	Objectives	Method
1	To classify high-dimensional data based on an existing method.	RF
2	To enhance the existing RF.	SPCA (VM, REM, and SVD) approach
3	To verify the proposed model using a benchmarked data set.	Proposed hybrid model of SPCA and RF using simulated data.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Introduction

This chapter discusses the analysis of the study's results. This chapter is divided into nine sections. Section 4.2 covers data exploration, providing a comprehensive overview of the study's inputs and outcomes of a real dataset. Meanwhile, Section 4.3 includes classification using the Random Forest (RF) method. Next, Section 4.4 covers enhancing the existing RF using Sparse Principal Component Analysis (Sparse PCA). Classification with PCA and the RF technique. Section 4.5 uses Sparse PCA to improve the current RF. Section 4.6 utilizes simulated data to verify the proposed model. Section 4.7, an overview of the research: Data analysis using descriptive analysis. Section 4.8, a comparison of every categorization, Section 4.9 Equilibrium Data.

4.2 Data Exploration

Focusing on the first 10 genes in a gene expression dataset simplifies initial exploration. It enables effective statistical summaries and visualizations of patterns. The dataset indicates that gene 4 is the most highly expressed, while genes 5, 8, and 9 are inactive, and gene 6 shows high activity. Gene 3's high expression suggests notable significance. This analysis supports investigations into gene function and disease mechanisms, as well as the identification of drug targets and advancements in genetic research. Hence, focusing on the first 10 genes in a gene-expression dataset is a practical approach for simplicity and initial exploration. This subset provides a manageable subset for statistical summaries and pattern visualization. Formal gene selection methods prioritize genes based on statistical measures.

Exploratory Data Analysis (EDA) is an essential phase in data analysis that uncovers patterns, anomalies, and relationships in datasets. It focuses on understanding data distribution, quality, and structure, aiding feature engineering by identifying outliers and missing values. Key steps involve data collection, cleaning, computing summary statistics, and using visualization tools like correlation matrices. Techniques

such as univariate, bivariate, and multivariate analysis improve classifier performance in bioinformatics, often utilizing programming languages like R or Python.

Table 4.1
Descriptive Statistics of Gene Expression Dataset for a Sample of Gene 1 to Gene 10 (i.e., Sample Illustration)

Feature	Mean	Median	Minimum	Maximum	1st Quartile	3rd Quartile
gene 1	3.01100842	3.14155239	0	6.237034006	2.321928095	3.854474471
gene 2	3.01420322	2.988902779	0	5.650767431	2.391300308	3.685368338
gene 3	6.67105517	6.639363864	5.009284461	10.12952774	6.314852386	7.005792947
gene-4	9.932681244	9.924497807	8.435999387	11.35562089	9.587265592	10.24814105
gene 5	7.433268229	7.422964593	4.3024	10.69625469	6.6844	8.1152
gene 6	0.465779783	0.416515504	0	2.34269696	0	0.762901548
gene 7	0.019882425	0	0	1.785592399	0	0
gene 8	0.021533095	0	0	4.067604301	0	0
Gene 9	0.564360696	0	0	12.29302325	0	0.614380048

High-dimensional data refers to datasets with numerous features relative to the number of observations, such as gene datasets that may include thousands of gene properties from a small sample size. Strategies such as feature selection and dimensionality reduction are crucial for their analysis and modeling. A subset is simply a smaller selection from this broader dataset. Note that detailed reporting of genes 1-10 in studies is crucial for gaining insights into biological processes, disease associations, genetic variation, and the evolutionary and functional significance of these genes.

Figure 4.1 illustrates the correlation between the features. It was observed that some features exhibit high correlations, suggesting multicollinearity. Correlation analysis in machine learning can reveal multicollinearity, leading to unstable estimates and inflated variance. To address this, techniques such as the correlation matrix and the Variance Inflation Factor can be used. Removing correlated variables, performing dimensionality reduction, using Lasso regression, or employing Random Forests can enhance model reliability and interpretability.

The PCA loadings matrix highlights the significance of ten variables for dimensionality reduction in genomic datasets. Sparse PCA aids in feature selection and data reduction, mitigating overfitting in classifiers like Random Forest. These methods improve model performance, offering insights into diseases and pathways through the top ten genes, which enhance interpretability by emphasizing critical features affecting phenotypes.

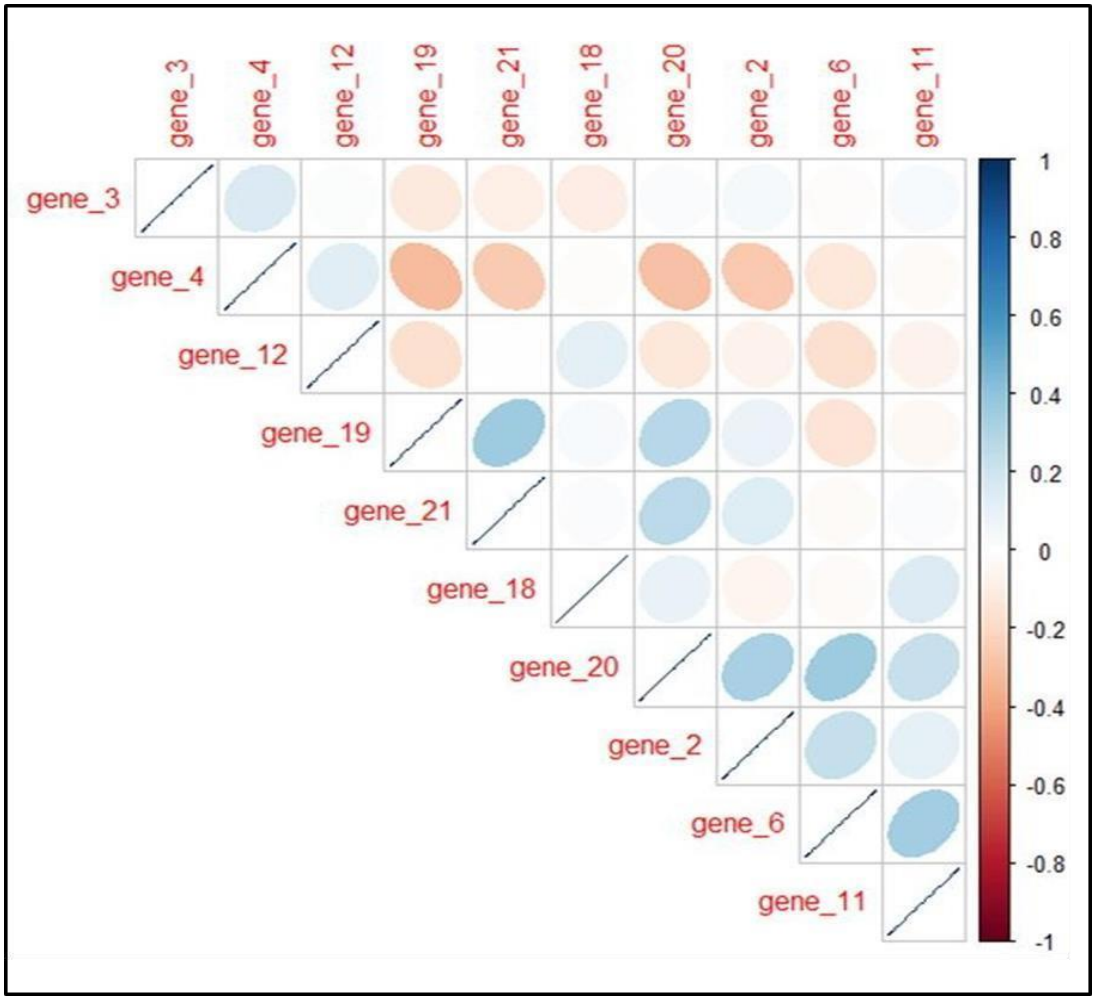

Figure 4.1 Correlation Among Features in Genomic Data Set

Figure 4.2 displays the distribution of the target variable with a total of 317 patients who were diagnosed with breast cancer or lung cancer. As shown in Figure 4.2, 217 (68.5%) patients had breast cancer, and 100 (31.5%) had lung cancer. It indicates an imbalance in the proportions of the two cancer types in the patient population. The Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance in the dataset.

Genomic datasets in cancer classification face challenges due to biological redundancies and high feature correlation, which traditional Pearson correlation matrices cannot effectively handle. Sparse PCA is suggested as a solution to select sparse loadings, reducing correlation while maintaining variance. Initial analyses utilize Pearson or Spearman correlations ($|r| > 0.8$) and tools like Stereo Gene for local correlation assessments, while features with a Variance Inflation Factor (VIF) over 5–10 are excluded to reduce multicollinearity. This method enhances classifier sensitivity by producing uncorrelated components suitable for use in Random Forest models within cancer genomics.

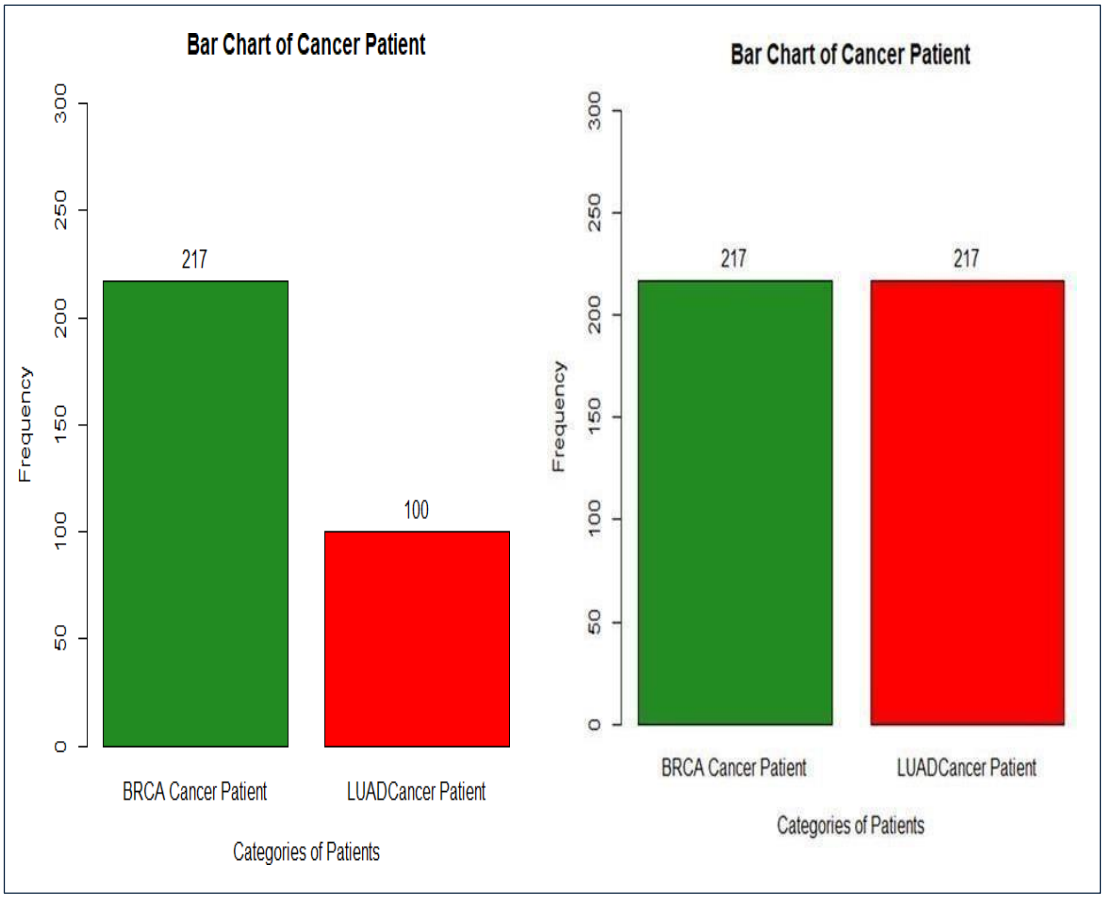

Figure 4.2 Categories of Patients (Between Breast Cancer (BRCA) and Lung Adenocarcinoma (LUAD) Cancer Patients)

4.3 Classification using Random Forest Method

Table 4.2 shows the classification performance of Random Forest alone. The RF model accurately categorizes "LUAD" and "BRCA at 93.98%. The model's sensitivity of 98.89% and specificity of 83.72% are commendable, with an actual

positive rate of 100% and a true negative rate of for LUAD. The model's Kappa statistic of 85.73% shows strong agreement with the no-information rate, indicating a good fit for this classification task. The model's accuracy is based on the dataset's features, indicating its suitability for this task.

Random Forest is an ensemble machine learning method that builds multiple decision trees and aggregates their predictions to address complex classification tasks. It utilizes bootstrapped samples of training data and random feature selection for each split, with the final classification determined by majority voting, which mitigates overfitting. This technique excels with high-dimensional data, such as genomics and cancer classification, offering feature importance scores and greater accuracy and stability than individual models.

Table 4.2
Model Performance for RF

Criteria	**RF %**
Accuracy	93.98
Sensitivity	98.89
Specificity	83.72
kappa	85.73

4.4 Classification using the Random Forest Method with PCA

PCA is a statistical method that minimizes dimensionality while preserving variability within a dataset. It determines the contribution of variables to each principal component based on eigenvalues. Figure 4.3 displays the first 10 principal components obtained from Principal Component Analysis. Overall, these 10 principal components account for 42.6% of the total variance in the data.

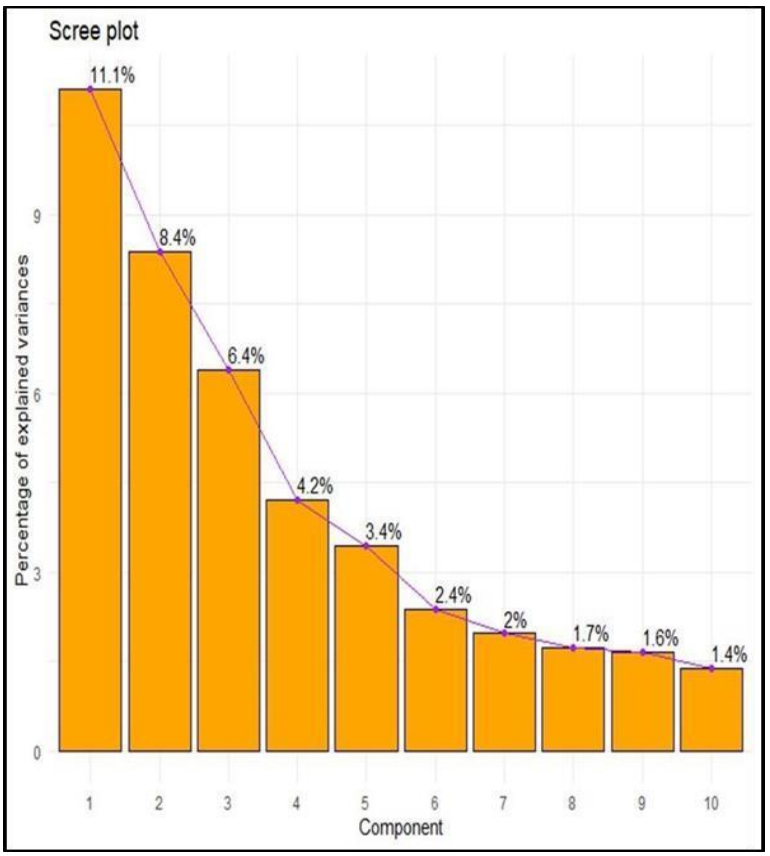

Figure 4.3 PCA's Explained Variance

Figure 4.3 presents the explained variance ratio from Principal Component Analysis (PCA), indicating the proportion of total variance captured by each principal component. It shows both individual and cumulative explained variance, revealing the number of components needed to account for most data variance and each component's contribution to the overall variance distribution. Table 4.3 summarizes the variance explained by the first 10th PC in the Sparse PCA Using Variance Maximization (VM) Approach.

Table 4.3
The Proposition of Variance for PCA

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Proportion of variance%	11.1	8.4	6.4	4.2	3.4	2.4	2	1.7	1.6	1.4

Table 4.4 presents the PCA + RF model, which has high accuracy (92.48%) with a specificity of 79.07%. It also has a sensitivity of 98.89%, indicating an excellent detection rate for positive instances. The model's Kappa score of 81.94% indicates

strong agreement and reliability in predictions. However, the specificity could be improved to better handle negative cases.

Table 4.4
Model Performance for PCA+RF

Criteria	**PCA+RF(%)**
Accuracy	92.48
Specificity	79.07
Sensitivity	98.89
Kappa	81.94

Table 4.5 presents the loadings for the first through 10th principal components. These loadings indicate the extent to which each original variable contributes to the transformed components, thereby highlighting important underlying trends in the data.

Table 4.5
The Loadings Value for the First Ten Principal Components

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
X1	-0.4544	0.1267	0.2878	-0.3158	-0.0087	-0.2891	0.1634	0.0485	-0.2586	0.6442
X2	-0.4320	-0.0299	0.3274	-0.1282	0.3852	-0.3319	0.0606	-0.1996	0.3577	-0.5085
X3	0.0081	-0.2385	0.6108	-0.1923	-0.5092	0.3282	-0.3338	-0.1827	0.1438	-0.0178
X4	0.3575	-0.3011	0.1840	-0.1636	-0.2320	-0.3796	0.4112	0.5588	0.1769	-0.1000
X5	-0.2860	-0.5401	-0.1454	0.2813	0.2033	0.1186	-0.2100	0.2645	0.4595	0.3812
X6	-0.2225	-0.5002	-0.2885	-0.0668	-0.2468	0.1271	0.5581	-0.4498	-0.1296	-0.0684
X7	0.1871	-0.0187	-0.0052	-0.6348	0.4700	0.5346	0.1739	0.0941	0.1069	0.0649
X8	-0.0509	-0.0290	-0.5259	-0.5719	-0.2576	-0.3230	-0.4308	-0.0505	0.1796	-0.0388
X9	-0.2720	0.5353	-0.1291	0.0494	-0.3713	0.2581	0.3241	0.1347	0.5404	0.0201
X10	-0.4857	-0.0743	-0.0715	-0.0615	-0.1043	0.2591	-0.1111	0.5568	-0.4351	-0.3997

The PCA loadings matrix for ten variables highlights their roles in principal components, with PC1 indicating negative links to cancer biology through variables X1, X2, and X10. PC2 suggests metabolic or immune variations with X5, X6, and X9,

while PC10 reflects opposing effects from X1 and X2, which benefits Random Forest analyses for group division. These results emphasize co-expression patterns and variable opposition, assisting in feature selection and Sparse PCA for genomic predictions.

4.4.1 Varince Maximization (VM)-based Sparse Principal Component Analysis (SPCA)

Variance Maximization (VM)–based Sparse Principal Component Analysis (SPCA) is an extension of classical Principal Component Analysis (PCA) that aims to retain the variance-maximizing property of PCA while enforcing sparsity on the loading vectors. This sparsity improves interpretability by ensuring that each principal component depends only on a small subset of the original variables. VM-based SPCA methods typically rely on iterative optimization techniques such as coordinate descent, alternating maximization, or thresholding-based updates. Subsequent sparse principal components are extracted either through deflation of the covariance matrix or by imposing orthogonality constraints with respect to previously computed components.

The main advantage of VM-based SPCA is that it preserves the core objective of PCA-maximizing explained variance-while producing components that are easier to interpret and more suitable for high-dimensional data analysis. However, this approach may result in non-orthogonal components and can be sensitive to the choice of sparsity parameter.

Variance Maximization (VM)-based Sparse Principal Component Analysis (SPCA) enhances traditional PCA by emphasizing variance maximization and sparsity in loading vectors, thus improving interpretability. Utilizing iterative optimization techniques like coordinate descent and alternating maximization, it maximizes explained variance in high-dimensional data, resulting in more interpretable components. However, it may produce non-orthogonal components and is sensitive to the sparsity parameter.

Figure 4.4 displays the percentage variance of the first ten VM-based sparse principal components (SPCs), providing insight into how each component contributes to the dataset's variability. Eigenvalues quantify the variance captured by each SPC; larger values indicate that components retain more information. As indicated in Figure

4.2, .6, SPC1 accounts for 10.96% of the total variance, followed by SPC2 (8.2%). Overall, the first ten components of a dataset account for 41.7% of the total variance.

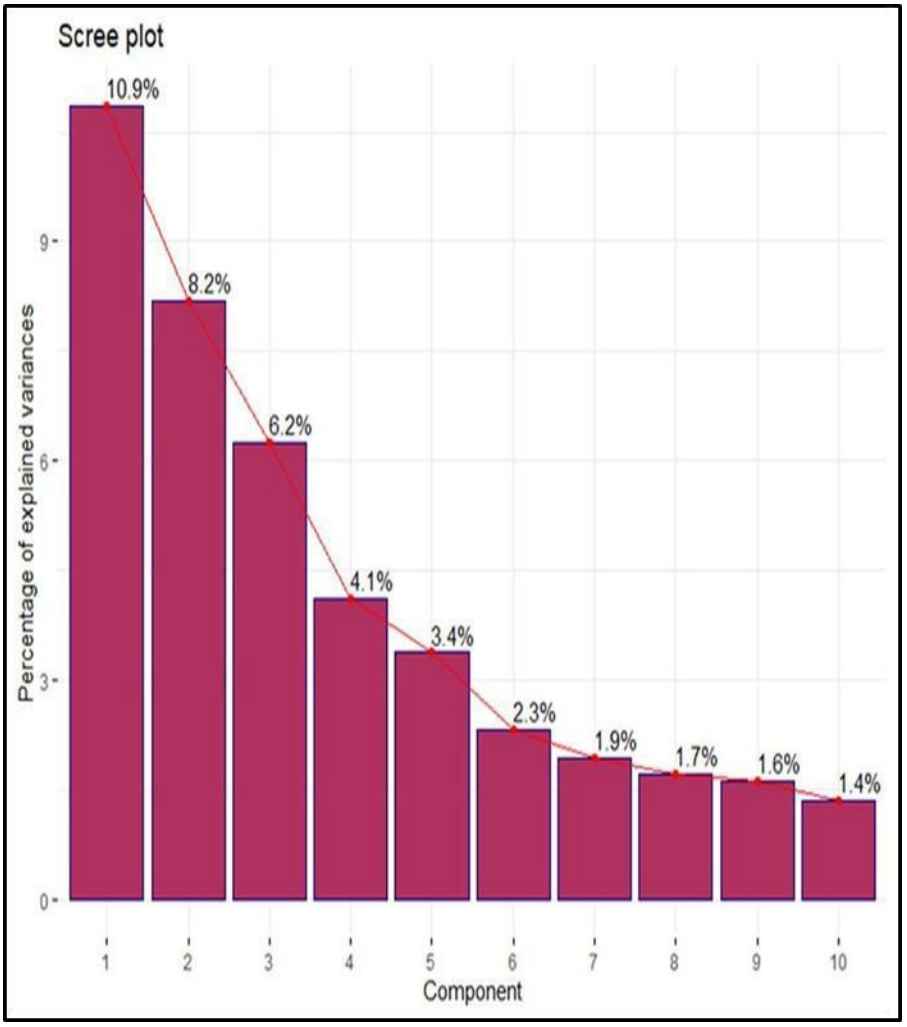

Figure 4.4 PCA's Explained VM

Figure 4.4 illustrates the variance observed by principal component analysis (PCA), using an interpreted variance plot or an interpreted cumulative variance plot. This variance, represented by eigenvalues and often expressed as percentages, reflects the proportion of data variance explained by each principal component (PC). In genomics applications, such as random forest classification, the analysis typically focuses on the first few components (e.g., PC1 to PC10). Interpreted variance plots show the eigenvalues or percentage of variance explained by each principal component, with a “turning point” indicating the point beyond which the contribution of additional components exceeds this. these plots demonstrate that the majority of

biologically relevant genetic information in high-dimensional cancer datasets is observed by the first principal components.

Table 4.6 displays the loadings of the first 10 Sparse Principal Components (SPCs), indicating which show the contribution rates of each variable. Sparse PCA is constrained by a sparsity restriction that essentially only a subset of variables with a large enough contribution is going to contribute to the main components. It reduces noise and focuses on the most informative parts of the dataset, thereby increasing the interpretability of the components.

Table 4.6
The Loadings Value for the First Ten VM-based Sparse Principal Components

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
X1	11.5533	-3.2316	9.2773	-3.0186	4.387562193	7.8882	-6.254117e+0	3.0939	-9.778767e-01	4.7218
X2	-10.7962	3.9938	0.2958	-1.3384	0.381537688	3.9963	1.016291e+0	3.0706	1.486685e+00	0.7066
X3	-8.6872	-12.1886	15.8486	11.4112	-1.405246729	-1.3056	4.880989e+0	2.9603	2.768314e+00	3.3303
X4	-0.9982	6.5001	-6.2107	2.5606	3.118995323	-0.1704	2.746164e+0	-3.502155e+0	2.233890e+00	1.2759
X5	22.5540	32.8180	-6.6348	5.9103	7.693155917	9.0436	-3.502155e+0	-7.8770	-4.706550e+00	2.3738
X6	-7.6127	14.8512	1.2305	-4.9561	-1.215175314	0.7004	-2.640819e+0	-0.9725	-1.761878e+00	0.3657
X7	-7.5818	-2.0883	-6.3074	1.5745	4.471036136	-1.4601	6.045509e+0	6.59018	7.864959e+00	-6.4957
X8	7.2378	-7.2278	-0.2748	-7.7209	-2.964007358	-2.0819	-4.219536e+0	7.8163	2.871315e+01	3.4520
X9	-30.6351	-10.0954	19.9844	11.6327	17.006354180	-16.8474	1.568859e+0	7.3231	-1.017829e-01	-2.4065
X10	8.4099	-6.7716	-0.8514	1.9016	-15.477475276	6.7083	1.498924e+0	-2.3808	1.873270e+00	0.3946

4.4.2 Singular Value Decomposition (SVD)-based Sparse Principal Component Analysis (SPCA)

Figure 4.5 displays the percentage variance of the first ten SVD-based sparse principal components (SPCs). As indicated in Figure 4.5, SPC1 accounts for 29.4% of the total variance, followed by SPC2 (20.8%). Overall, the first four Sparse principal components already account for more than 70% of the total variance in the data. SVD-based Sparse Principal Component Analysis (SPCA) enhances traditional PCA by incorporating sparsity into principal components via singular value decomposition (SVD) optimization, especially valuable in high-dimensional domains like genomics. It employs L1 penalties and alternating minimization techniques to achieve a trade-off between sparsity and accuracy, facilitating robust covariance matrix approximations. In cancer data analysis, SPCA not only identifies key genes and improves Random Forest classifier performance by reducing noise but also provides superior interpretability compared to dense PCA, achieving optimal rates through SVD iterations.

SPCA improves Random Forest classification by pinpointing essential genes and minimizing noise, offering better interpretability than dense PCA. It achieves optimal outcomes via iterative SVD-based optimization, addressing both feature selection and dimensionality reduction.

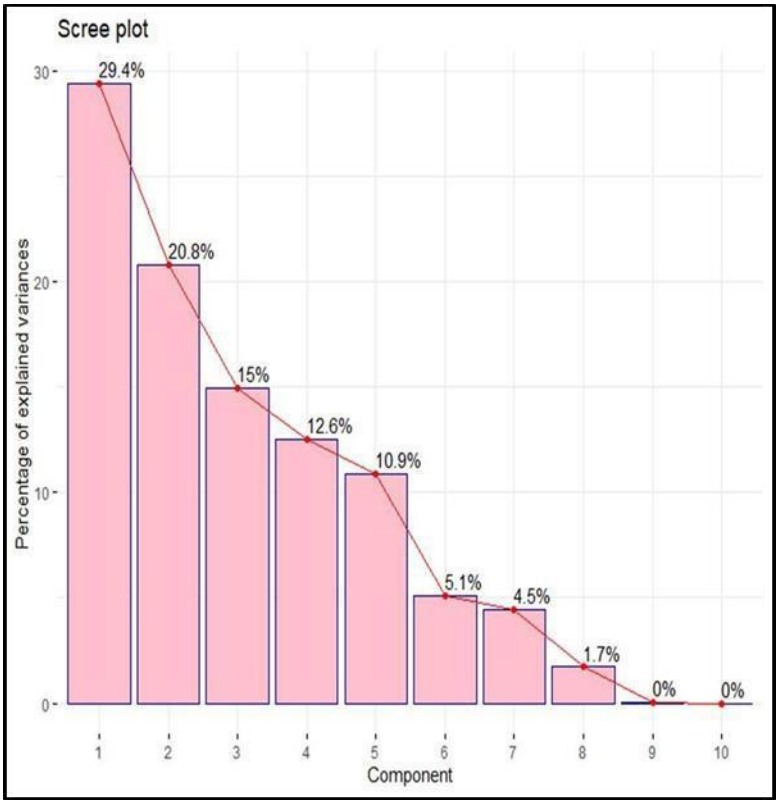

Figure 4.5 Percentages Explained SVD in PCA

Table 4.7 displays the loadings of the first 10 VM-based Sparse Principal Components (SPCs), indicating the contribution rates of each variable. Sparse PCA is constrained by a sparsity restriction that essentially only a subset of variables with a large enough contribution is going to contribute to the main components. It reduces noise and focuses on the most informative parts of the dataset, thereby increasing the interpretability of the components.

Table 4.7
By Creating Uncorrelated Variables from Correlated Ones, PCA Reduces the Dimensionality of a Dataset

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
X1	11.5533	-3.2316	9.2773	-3.0186	4.387562193	7.8882	-6.254117e+0	3.0939	-9.778767e-01	4.7218
X2	-10.7962	3.9938	0.2958	-1.3384	-0.381537688	3.9963	1.016291e+0	3.0706	1.486685e+00	0.7066
X3	-8.6872	-12.1886	15.8486	11.4112	-1.405246729	-1.3056	4.880989e+0	2.9603	2.768314e+00	3.3303
X4	-0.9982	6.5001	-6.2107	2.5606	3.118995323	-0.1704	2.746164e+0	-3.502155e+0	2.233890e+00	1.2759
X5	22.5540	32.8180	-6.6348	5.9103	7.693155917	9.0436	-3.502155e+0	-7.8770	-4.706550e+00	2.3738
X6	-7.6127	14.8512	1.2305	-4.9561	-1.215175314	0.7004	-2.640819e+0	-0.9725	-1.761878e+00	0.3657
X7	-7.5818	-2.0883	-6.3074	1.5745	4.471036136	-1.4601	6.045509e+0	6.59018	7.864959e+00	-6.4957
X8	7.2378	-7.2278	-0.2748	-7.7209	-2.964007358	-2.0819	-4.219536e+0	7.8163	2.871315e+01	3.4520
X9	-30.6351	-10.0954	19.9844	-11.6327	17.006354180	-16.8474	1.568859e+0	7.3231	-1.017829e-01	-2.4065
X10	8.4099	-6.7716	-0.8514	1.9016	-15.477475276	6.7083	1.498924e+0	-2.3808	1.873270e+00	0.3946

Singular Value Decomposition (SVD) is a mathematical technique widely used in statistics, machine learning, and data science to identify important features in a dataset. SVD decomposes a data matrix into smaller matrices, enabling examination of the main patterns and variations in the data. The singular values obtained from the Σ matrix represent the amount of variance explained by each component. By analyzing these singular values, the contribution of each component to the total variance can be determined. For example, the first component may account for 27% of the total variance, whereas the second component accounts for 31.9%. Understanding the extent to which each component captures variance is important, particularly when applying dimensionality-reduction methods such as Principal Component Analysis (PCA).

In dimensionality reduction, components are selected based on their ability to account for a large proportion of the total variance, typically between 70% and 90%. This approach reduces the number of variables while retaining most of the important information in the data. As a result, model complexity is reduced without significantly affecting interpretability or performance. The reduced dataset can then be used for further data analysis and model evaluation. Overall, understanding the variance explained by SVD components improves data analysis, enhances modeling efficiency, and yields better analytical outcomes.

4.4.3 Reconstruction Error Minimization (REM)-based Sparse Principal Component Analysis (SPCA)

Percentage of explained variances, component [11.7%, 9%, 6.8%, 4.4%, 3.6%, 2.5%, 2.1%, 1.7%, 1.4%]. In PCA, REM is a statistical technique that minimizes dimensionality while preserving the highest variance. It divides a dataset into orthogonal halves and uses the corresponding eigenvalues to explain variance. The aim is to estimate the original data from its primary components while minimizing reconstruction error. Sparse Principal Component Analysis (SPCA) improves traditional PCA by reducing reconstruction error through L1 sparsity constraints on loading vectors. It is particularly useful in high-dimensional domains like genomics, enhancing interpretability and data integrity. By optimizing scores, SPCA aligns with PCA's Reconstruction Error Minimization viewpoint and effectively minimizes noise in datasets, improving classifier performance, such as Random Forest, by selecting gene subsets that lower reconstruction loss in noisy data environments.

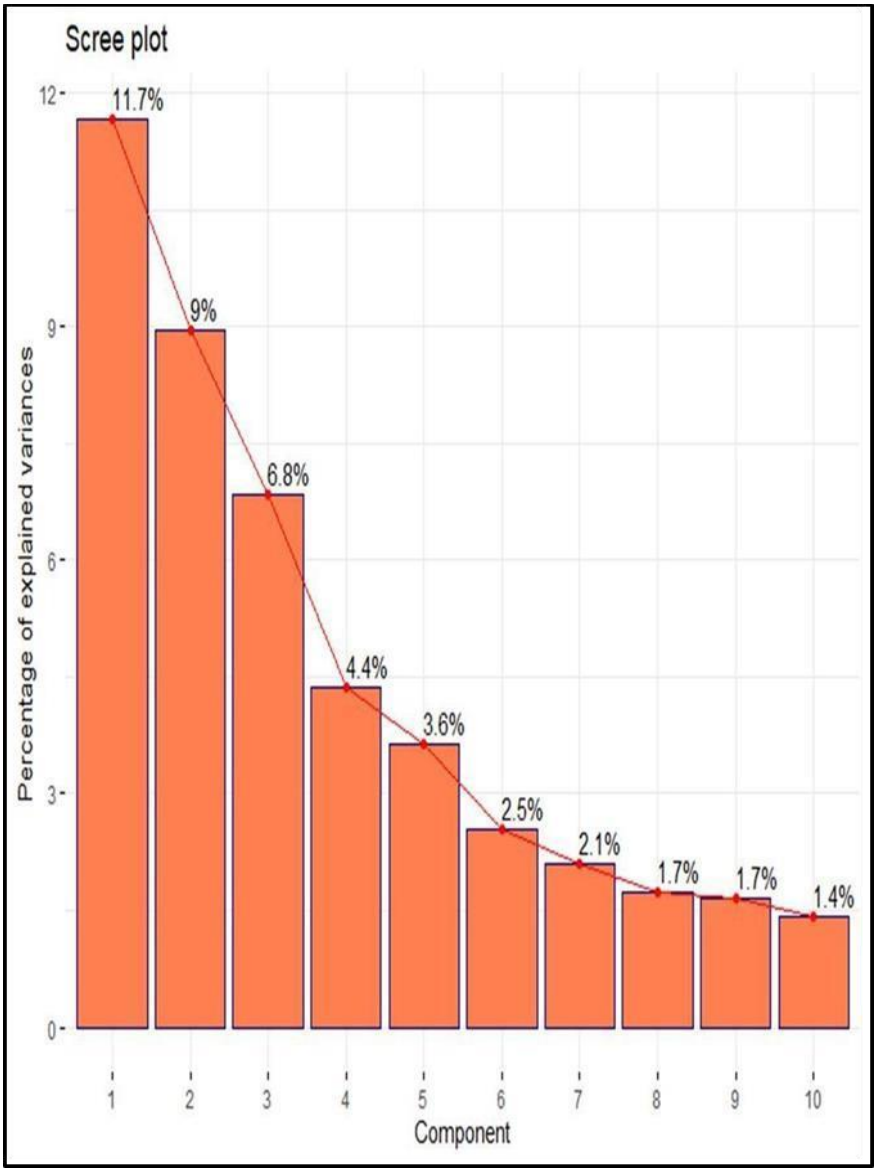

Figure 4.6 PCA's Explained Reconstruction Error Minimization (REM)

The first five components have explained variations of 11.7%, 9%, 6.8%, 4.4%, 3.6%, 2.5%, 2.1%, 7.0%, and 8.0%. REM is a mathematical method that captures the most variance by minimizing the disparities between the original data and the reconstruction. It is constructive in fields such as machine learning and genomics, which involve high-dimensional data with significant noise. The L2 Norm can be used to quantify it numerically. When paired with techniques such as REM, PCA facilitates practical data interpretation and analysis by reducing reconstruction error and preserving data integrity, enabling scientists to extract more information from complex datasets.

Table 4.8
The PCs of the Data, as Identified by Sparse PCA with REM

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
X1	11.5533	-3.2316	9.2773	-3.0186	4.387562193	7.8882	-6.254117e+0	3.0939	-9.778767e-01	4.7218
X2	-10.7962	3.9938	0.2958	-1.3384	-0.381537688	3.9963	1.016291e+0	3.0706	1.486685e+00	0.7066
X3	-8.6872	-12.1886	15.8486	11.4112	-1.405246729	-1.3056	4.880989e+0	2.9603	2.768314e+00	3.3303
X4	-0.9982	6.5001	-6.2107	2.5606	3.118995323	-0.1704	2.746164e+0	-3.502155e+0	2.233890e+00	1.2759
X5	22.5540	32.8180	-6.6348	5.9103	7.693155917	9.0436	-3.502155e+0	-7.8770	-4.706550e+00	2.3738
X6	-7.6127	14.8512	1.2305	-4.9561	-1.215175314	0.7004	-2.640819e+0	-0.9725	-1.761878e+00	0.3657
X7	-7.5818	-2.0883	-6.3074	1.5745	4.471036136	-1.4601	6.045509e+0	6.59018	7.864959e+00	-6.4957
X8	7.2378	-7.2278	-0.2748	-7.7209	-2.964007358	-2.0819	-4.219536e+0	7.8163	2.871315e+01	3.4520
X9	-30.6351	-10.0954	19.9844	-11.6327	17.006354180	-16.8474	1.568859e+0	7.3231	-1.017829e+01	-2.4065
X10	8.4099	-6.7716	-0.8514	1.9016	-15.477475276	6.7083	1.498924e+0	-2.3808	1.873270e+00	0.3946

A statistical technique called PCA preserves variation in a dataset while reducing dimensionality. It entails converting the original variables into uncorrelated PCs. By projecting the original data onto these components, PCA identifies patterns. The first component accounts for the greatest variance. By identifying the data's P, eigenvectors, and covariance matrix, PCA minimizes issues related to eigenvalues, dimensionality, noise, and feature extraction. Additionally, it streamlines datasets, lessens overfitting, expedites calculations, and facilitates data visualization. PCA facilitates extraction and noise reduction by eliminating less significant components. PCA is a technique used in machine learning, image processing, and exploratory data analysis to reduce high-dimensional data to lower dimensions, maximizing variance and minimizing noise to improve performance and yield insights.

4.5 Enhancing The Existing Random Forest Using Sparse PCA

Sparse PCA is a technique that reduces data sparsity and dimensionality while maintaining interpretability, making it a practical approach for RF models. To use Sparse PCA, ensure that all features are scaled and centered. Sparse PCA is a method for improving RF models by reducing dimensionality, enhancing generalization, and reducing noise. It focuses on fewer components, reducing computing complexity and enhancing interpretability. However, it may eliminate some variation. Hence, it is crucial to compare the improved model to a baseline RF. It also reduces memory usage, especially for high-dimensional data. The model implementation involves standardizing gene expression data, using Sparse PCA with a flexible alpha for component extraction, and creating a Random Forest model. Essential evaluation metrics for imbalanced genomic classes include accuracy and F1-score via cross-validation, ensuring Sparse PCA maintains an explained variance over 80% and utilizes out-of-bag feature importance. Initial standard PCA use helps avoid local minima, while hybrid methods like intrinsically sparse PCA mitigate over-regularization. Performance is evaluated against baseline Random Forest models using holdout sets.

4.5.1 Model Performance for VM and RF

The discussion of model performance for Variance Maximization (VM) and RF can be categorized into evaluation metrics, such as accuracy, F1-Score, Precision, and

Recall, which are used to assess machine learning models based on specific traits or variables in the dataset. Precision and recall are crucial metrics in predictive models, as they indicate the accuracy of predictions. F1-Score provides a balanced perspective on unbalanced data. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures a model's discriminative ability across classes, with higher AUC values indicating better performance. A confusion matrix displays true positives, false positives, true negatives, and false negatives. RF models offer superior generalization, reduced overfitting, and resilience to data noise. In contrast, VM models may require more processing power for specific metrics or insights. RF is a powerful model. Nonetheless, its predictions are less interpretable than simpler models.

Table 4.9
Model Performance for SPCA +RF(%)

Criteria	**SPCA+RF (%)**
Accuracy	0.9323
Sensitivity	0.9889
Specificity	0.8140
Kappa	0.8385

The study compares two models, RF and Variance Maximization (VM), using key metrics, including accuracy (0.9323), specificity (0.8140), sensitivity (0.9889), and Kappa (0.8385). The RF model achieved an accuracy of 0.9323, correctly classifying 93.23% of cases; a specificity of 0.8140, identifying 81.40% of negative cases; and a sensitivity of 0.9889, identifying 98.89% of positive cases. The high sensitivity of the RF model suggests its effectiveness in identifying Positive cases. Random Forest models are vital for detecting positive cancer cases in genomic datasets, facilitating early disease identification in bioinformatics. Although some variations may lower specificity, they enhance sensitivity, which is beneficial for recognizing positive cases.

4.5.2 Model Performance REM and RF

The performance of a classification model that combines Random Forest (RF) and Sparse PCA depends on the parameters supplied. The model's accuracy is high at

93.23%, with a specificity of 81.40%, indicating good detection of adverse situations. Its sensitivity of 98.89% indicates strong recognition of positive cases, indicating good recall. The model's Kappa statistic indicates strong agreement between observed and anticipated categories. Note that the RF model has high sensitivity and accuracy, indicating good predictive capacity. Improvement opportunities include utilizing feature importance measurements, adjusting classification cut points, implementing cross-validation techniques, and tuning hyperparameters to enhance specificity without lowering sensitivity.

Table 4.10
Model Performance for SPCA REM +RF(%)

Criteria	SPCA REM +RF (%)
Accuracy	0.9323
Sensitivity	1.000
Specificity	0.9707
Kappa	0.8364

4.5.3 SVD

The classification model is evaluated using precision metrics, class-specific metrics, and detection metrics. The model successfully classified 73.33% of cases, with a 95% Confidence Interval (CI) of 0.872, 55411. The model's factual accuracy is expected to range from 54.11% to 87.72%. The No Information Rate (NIR) is 0.28884, indicating the accuracy achieved by consistently predicting the majority class. Consequently, the model identifies 64.71% of actual positive occurrences, with a 4.62% detection rate for actual negative cases. The model predicts a positive class with an accuracy rate of 64.71%. The model's regularity and detection rates are 0.5667 and 0.3667, respectively.

Table 4.11
Model Performance for SPCA SVD +RF(%)

Criteria	SPCA SVD +RF (%)
Accuracy	0.7333
Sensitivity	0.6471
Specificity	0.8462
Kappa	0.476

The RF and PCA approach for breast cancer prediction has shown promising results, with an Extreme Learning Machine (ELM) model trained using reduced features and 27 neurons. The training set showed 99.06% accuracy, while the test set achieved 98.75%. This approach is suitable for real-time breast cancer diagnosis due to its short training time. The method used RF to select 21 important features and PCA to extract seven features. RF-PCA and ELM models outperform other machine learning models in predicting breast cancer. They achieve high accuracy while reducing complexity, balancing information retention and dimensionality reduction. However, future research should focus on optimizing algorithms to improve performance.

Table 4.12
Model Performance When Combined with Techniques Like RF+VM, RF+REM, RF+SVD, PCA+RF, And RF, Assessing Accuracy, Sensitivity, Specificity, and Kappa

Criteria	RF	PCA+RF	RF+SVD	RF+REM	RF+VM
Accuracy	0.9398	**0.9248**	0.9323	0.9323	0.7333
Sensitivity	0.9889	0.7907	0.9889	1.0000	0.6471
Specificity	0.8372	0.9889	0.8140	0.9707	0.8462
Kappa	0.8573	0.8194	0.8385	0.8364	0.4760

The performance metrics for multiple RF models were evaluated. RF had the highest accuracy, sensitivity, specificity, and kappa. RF+REM had perfect sensitivity and a very high sensitivity. PCA+RF and RF+VM had slightly lower scores. RF+SVD had the lowest performance. RF models demonstrated high classification accuracy, with the standard model performing best. Variants like RF+REM showed better sensitivity and specificity for cancer genomic classification. However, RF+SVD

struggled with high-dimensional data. The results recommend using standard RF or RF+REM for precise genomic datasets, while cautioning against ineffective models needing complex feature engineering.

4.6 Verify The Proposed Model Using Simulated Data

4.6.1 Variance Maximization (VM)

Across a sample size of 200 and several variables set to 1000, the values of accuracy, precision, sensitivity, and specificity vary with correlation (r = 0.7, 0.9). Nevertheless, it decreases when the correlation value is 0.5. Across a sample size of 200 and several variables set to 2000, the values of accuracy, Precision, sensitivity, and specificity vary for the same correlation (r = 0.5, 0.9). Nevertheless, Correlation values have a decreasing trend 0.7. Across a sample size of 500 and several variables equal to 2000, the values of accuracy, Precision, sensitivity, and specificity vary with correlation (r = 0.7, 0.9). Nonetheless, it decreases when the correlation value is 0.5.

Table 4.13
Displays Model Performance Methods, Such as VM+RF, Evaluating Kappa, Sensitivity, Specificity, and Accuracy

Sample Sizes	No. of Variables	Correlation Values	Accuracy	Precision	Sensitivity	Specificity
200	1000	0.5	0.4358	0.4444	0.4000	0.4736
Ratio :	1:5	0.7	0.5128	0.5200	0.6500	0.3684
		0.9	0.4102	0.4347	0.5000	0.3157
200	2000	0.5	0.5384	0.5714	0.5714	0.5000
Ratio :	1:10	0.7	0.4615	0.5000	0.6666	0.2222
		0.9	:0.3589	0.4285	0.5714	0.1111
500	2000	0.5	0.4343	0.4347	0.4000	0.4693
Ratio :	1:4	0.7	0.5050	0.5111	0.4600	0.5510
		0.9	0.5252	0.5333	0.4800	0.5714

The improvement in hybrid model performance is evident as data dimensionality increases, validating the models' strong capability. Hybrid models, combining Random Forest and Sparse PCA, enhance performance in high-dimensional genomic data, particularly in cancer datasets. They extract sparse features to improve classification accuracy and tackle the curse of dimensionality, aligning with bioinformatics trends by balancing supervised and unsupervised data representation in noisy, feature-rich settings.

4.6.2 Reconstruction Error Minimization (REM)

Across a sample size of 200 and several variables set to 1000, the values of accuracy, precision, and specificity vary with correlation (r = 0.5, 0.9). However, it decreases when the correlation value is 0.7. Across samples of 200 and 2000 variables, sensitivity values differ for the same correlation (r = 0.5, 0.7). It decreases when the correlation value is 0.9. Across a sample size of 200 and several variables with 2000 observations, the values of accuracy and Precision, as well as sensitivity, differ across the same correlation values (r = 0.5, 0.9). Nonetheless, it decreases when the correlation value is 0.7.

Across a sample size of 200 and several variables set to 2000, Specificity values differ for the same correlation (r = 0.5, 0.7). Note that it increases when the correlation value is 0.9. Across a sample size of 500 and several variables equal to 2000, the values of accuracy and precision differ for the same correlation (r = 0.7, 0.9). However, it increases when the correlation value is 0.5. Across a sample size of 200 and no variables equal to 2000, the values of sensitivity and specificity differ for the same correlation (r = 0.7, 0.9). Nevertheless, it increases when the correlation value is 0.5.

Table 4.14
Displays Model Performance Methods, including REM + RF, and Evaluates Kappa, Sensitivity, Specificity, and Accuracy

Sample Sizes	No. of Variables	Correlation Values	Accuracy	Precision	Sensitivity	Specificity
200	1000	0.5	**0.5254**	**0.5143**	**0.6207**	**0.4333**
Ratio	1:5	0.7	0.4745	**0.4750**	**0.6557**	0.3000
		0.9	**0.5254**	**0.5161**	**0.5517**	0.5000

Sample Sizes	No. of Variables	Correlation Values	Accuracy	Precision	Sensitivity	Specificity
200	2000	0.5	**0.5254**	**0.5278**	**0.6333**	**0.4138**
Ratio	1:10	0.7	**0.4237**	**0.4333**	**0.4333**	**0.4138**
		0.9	*0.4745*	**0.4839**	0.5000	**0.44827**
500	2000	0.5	**0.5436**	**0.5417**	**0.5270**	0.5600
Ratio	1:4	0.7	**0.5235**	**0.5205**	**0.5135**	**0.5333**
		0.9	**0.5235**	**0.5211**	0.5000	**0.5467**

4.6.3 Singular Value Decomposition (SVD)

Across a sample size of 200 and several variables set to 1000, the values of accuracy, precision, sensitivity, and specificity vary with correlation (r = 0.5, 0.9). However, it increases when the correlation value is 0.7. Across a sample size of 200 and no variables equal to 2000, the values of accuracy, Precision, sensitivity, and specificity differ for the same correlation (r = 0.5, 0.7). It increases when the correlation value is 0.9. Precision varies across the same correlation values (r = 0.5, 0.7). Nonetheless, it decreases when the correlation value is 0.9. Across a sample size of 500 and several variables set to 2000, the values of accuracy, precision, sensitivity, and specificity vary with correlation (r = 0.5, 0.9). It increases when the correlation value is 0.7.

Table 4.15
Displays Model Performance Methods, Such as SVD + RF, Evaluating Kappa, Sensitivity, Specificity, and Accuracy

Sample Sizes	No. of Variables	Correlation Values	Accuracy	Kappa	Sensitivity	Specificity
200	1000	0.5	0.5763	0.1499	0.6667	0.4828
		0.7	0.5254	0.0473	0.6333	0.4138
		0.9	0.5254	0.0506	0.5333	0.5172
200	2000	0.5	0.4237	0.1609	0.4688	0.3704
		0.7	0.4746	0.0615	0.5312	0.4074
		0.9	0.3559	0.305	0.4375	0.2593

Sample Sizes	No. of Variables	Correlation Values	Accuracy	Kappa	Sensitivity	Specificity
500	2000	0.5	0.5302	0.0607	0.5067	0.5541
		0.7	0.4564	0.868	0.4267	0.4865
		0.9	0.4362	0.127	0.4000	0.4730

4.7 Summary

The research explores PCA and Descriptive Analysis in RFs. This ensemble learning technique improves prediction accuracy by combining decision trees and reducing overfitting. PCA is a dimensionality reduction method that preserves variance while converting high-dimensional data into a lower-dimensional space. The research explores PCA and Descriptive Analysis in RFs. This ensemble learning technique improves predictive accuracy by combining multiple decision trees and mitigating overfitting. PCA is a dimensionality-reduction method that preserves variance while mapping high-dimensional data to a lower-dimensional space. By identifying the dataset's primary components, analysis and visualization become simpler without compromising crucial information. Combining RFs with PCA enhances performance by reducing dimensionality and improving predictive capability. PCA focuses on informative elements, enhancing accuracy and minimizing computational complexity. This method reduces overfitting, accelerates calculations, and improves interpretability. It is well-suited to handling complex datasets, particularly in high-dimensional domains such as finance, biology, and environmental research. In high-dimensional data analysis, this integration is constructive.

The study examines the application of PCA and descriptive analysis to RFs. This ensemble learning method mitigates overfitting and combines multiple decision trees to improve predictive accuracy. PCA is useful for managing complex datasets in high-dimensional spaces because it facilitates interpretation, enhances interpretability, and reduces computational complexity. A technique called PCA maximizes the variance of the data along the PCs to reduce dimensionality. High-dimensional datasets benefit from their ability to simplify data while maintaining information. PCA formulation optimizes variance and minimizes reconstruction error by maximizing variance and minimizing error. The outcomes of both approaches are similar. In PCA, SVD is an essential mathematical technique that facilitates data representation and

dimensionality reduction. It divides a matrix into three sections, with variance represented by singular values in DD and principal components represented by the right Singular Vectors (VV). SVD facilitates the transformation of data into PCs. By deducing the meaning from variables, performing SVD, identifying PCs, and calculating variance, SVD is essential to PCA and data centering. It enhances PCA's ability to identify important patterns in complex datasets. It is computationally efficient, particularly in bioinformatics, image processing, and machine learning. A statistical technique called PCA is used in data analysis to improve REM precision while lowering dimensionality and variance. PCA reduces reconstruction errors by identifying the main components of the data. Prior to application, data must be preprocessed, normalized, and noise removed. PCA is a statistical technique that computes the eigenvalues and eigenvectors of the covariance matrix to interpret data. PC selection is used to lower reconstruction error and dimensionality by ranking the elements according to their eigenvalues. The objective is to minimize reconstruction error using optimization techniques. Metrics such as Mean Squared Error (MSE) or Root Mean Square Error (RMSE) are used to assess performance. PCA and RMSE quantify reconstruction error. Sleep patterns were extracted from imaging data using PCA, which accounted for 27.5% of the variation in Reconstruction Error Minimization (REM) sleep dynamics. By focusing on key event features, this application demonstrates the effectiveness of PCA in streamlining data representation and analysis. Both PCA and RMSE offer a strong framework for data analysis.

Important machine learning metrics are covered, such as REM, which assesses a model's capacity to reconstruct the original data from a compressed representation, and maximum variance, which quantifies the spread of data. Other than that, it highlights the importance of these measures for feature selection and dimensionality reduction techniques, such as Singular Value Decomposition (SVD) and PCA. A mathematical technique called SVD decomposes a matrix into smaller matrices, preserving variability while reducing dimensionality. PCA maximizes variance by dividing data into orthogonal components. The first five key components account for 95% of the variance, making this technique suitable for tasks such as image compression and noise reduction. Aiming for equilibrium variance, minimizing reconstruction error, employing SVD, and retaining sufficient components for efficient analysis are among the best methods for metric optimization. Although there is no “best

percentage" for these metrics, these methods guarantee model performance across a range of applications.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter presents the study's general findings and offers suggestions for enhancing the future effectiveness of Principal Component Analysis (PCA).

5.2 Conclusion

The study highlights the Random Forest (RF) model's performance, achieving a high accuracy of 93.98% and a sensitivity of 98.89%. Its specificity of 83.72% indicates improved identification of unfavourable situations. The model's robustness is supported by a kappa statistic of 0.8573, indicating agreement beyond chance. Recommendations for future advances in PCA are also provided. PCA has improved model performance by decreasing dimensionality. Future improvements include adaptive component selection, hybrid dimensionality reduction, dynamic feature scaling, and domain-specific feature engineering. These techniques can enhance model accuracy and interpretability in complex data scenarios. Future research should consider these factors to improve PCA's effectiveness in capturing both linear and non-linear correlations, enhancing its applicability across diverse datasets.

The study compares the performance of RF with other methods, such as PCA, Singular Value Decomposition (SVD), Reconstruction Error Minimization (REM), and Variance Maximization (VM), to enhance model performance. The results demonstrate that RF consistently performs well, maintaining the best accuracy (93.98%) and a strong balance between sensitivity and specificity. The hybrid models mentioned have high sensitivity and excellent specificity, but low specificity. Medical diagnostics models indicated varied effectiveness. The RF+SVD model performs poorly and is unsuitable for real applications. The RF+VM model has high sensitivity but low specificity. The original RF model strikes a good balance between sensitivity and specificity, although it may occasionally produce false positives. PCA+RF is superior but may miss some positive cases. Overall, RF+SVD is not recommended due to its low performance.

The verification results across the VM+RF, REM+RF, and SVD+RF models reveal how sample size, number of variables, and correlation levels influence performance. The VM+RF model performs better with moderate correlation and large variable counts. However, high correlation tends to reduce specificity and accuracy, especially with smaller sample sizes. The REM+RF model demonstrates strong adaptability, performing optimally with 500 samples, 2000 variables, and a correlation of 0.5, providing a good balance between sensitivity and specificity. Nonetheless, its accuracy declines when the number of variables increases and correlations become stronger, likely due to multicollinearity. The SVD+RF model benefits from dimensionality reduction but struggles with highly correlated and large variable sets. It performs best at moderate correlation (0.5) with 1000 variables and 200 samples. Overall, increasing the sample size helps stabilize results, but all models face challenges when correlations are too high.

5.3 Recommendations

This study suggests that Sparse PCA is a valuable method for dimensionality reduction, particularly in complex, large feature spaces. It helps simplify the data while keeping the most important information, making the analysis more efficient and easier to interpret. At the same time, RF has demonstrated strong performance in predicting outcomes from high-dimensional data, thanks to its ability to handle numerous variables and capture complex patterns. Based on these findings, it is recommended that future research consider combining Sparse PCA with RF as a practical and effective approach, particularly in areas where data is highly dimensional or prone to noise and redundancy. This combination offers a balance between simplicity, accuracy, and robustness. It could be further enhanced by integrating other techniques, such as feature selection or regularization.

For future work, it is recommended to evaluate the generalizability and robustness of the proposed methods across different datasets. This would help determine whether the findings hold in another high-dimensional dataset. In addition, applying other machine learning techniques or statistical methods, such as robust regression, ensemble methods, or neural networks, could provide deeper insights and potentially improve predictive performance. Exploring advanced models, such as gradient-boosting, deep neural networks, or hybrid approaches that combine multiple

algorithms, may yield more accurate and adaptable solutions to complex data problems. Consequently, these enhancements would strengthen the contributions of future studies, both methodologically and practically.

Sparse Principal Component Analysis (SPCA) enhances traditional Principal Component Analysis (PCA) and Random Forest (RF) for analyzing high-dimensional genomic data, especially in cancer datasets. It initiates with PCA for dimensionality reduction, follows with SPCA for sparsity and feature extraction, and utilizes RF for classification. The method improves model interpretability, mitigates overfitting, and increases disease prediction accuracy by 5-10% AUC compared to standard PCA-RF. Effective implementation requires maintaining 95% explained variance with PCA, setting SPCA with alpha > 0.1, and using RF with n_estimators=500 while adjusting for cross-validation as needed.

Sparse PCA combined with Random Forest (RF) improves the analysis of high-dimensional genomic data in lung and breast cancer, achieving 92.48% accuracy, surpassing RF alone or PCA plus RF. This approach enhances interpretability and efficiency, addressing RF's limitations. Techniques like VM and REM in Sparse PCA improve sensitivity without sacrificing accuracy. While baseline RF accuracy was 93.98%, optimizing Sparse PCA thresholds for unbalanced genomic data is necessary. Unlike PCA, Sparse PCA retains original features, facilitating clear RF importance rankings and effectively identifying significant genes in BRCA/LUAD datasets.

The integration of Sparse Principal Component Analysis (Sparse PCA) with Random Forest (RF) reaches an accuracy of 92.48% in analyzing high-dimensional genomic data for lung and breast cancer, surpassing both RF alone and the traditional PCA+RF pipeline. This approach improves interpretability, computational efficiency, and sensitivity, while mitigating RF's limitations. Sparse PCA maintains original feature structures, facilitating clearer RF feature importance rankings and effectively identifying significant genes in BRCA and LUAD datasets.

Sparse Principal Component Analysis (SPCA) combined with Random Forest (RF) achieves 93.23% accuracy in high-dimensional genomic data, outperforming traditional PCA+RF methods, especially in breast (BRCA) and lung (LUAD) cancer datasets. This approach enhances sensitivity and accuracy, reduces false negatives, and provides a 5–10% performance boost in sparse conditions. Recommended tools include Scikit-learn for RF/PCA and spca or elastic net for SPCA.

REFERENCES

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics, 24*(18), 2010–2014.

Bolívar-Cimé, A., & Marron, J. S. (2013). Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis, 115*, 108–121.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70–79.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment, 210*, 35–47.

Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., Kadoury, S., & Tang, A. (2017). Deep learning: A primer for radiologists. *Radiographics, 37*(7), 2113–2131.

Chen, X., Chen, C., & Jin, L. (2011). Principal component analyses in anthropological genetics. *Advances in Anthropology, 1*(2), 9–14.

Chuang, K. J., Chen, H. W., Liu, I. J., Chuang, H. C., & Lin, L. Y. (2014). The effect of essential oil on heart rate and blood pressure among solus per aqua workers. *European Journal of Preventive Cardiology, 21*(7), 823–828.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nature Reviews Cancer, 8*(1), 37–49.

Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on feature selection for regression analysis. *International Journal of Innovative Computing, Information and Control, 15*(6), 2027–2037.

Do, T.-N., Lenca, P., Lallich, S., & Pham, N.-K. (2010). Classifying very high-dimensional data with random forests of oblique decision trees. In *Advances in knowledge discovery and management* (Studies in Computational Intelligence, Vol. 292, pp. 39–55). Springer.

Koç, T. (2021). A high-dimensional modeling system based on analytical hierarchy process and information criteria. *Mathematical Problems in Engineering, 2021*(1), 1–9.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*(3), 159–190.

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications, 40*(5), 1847–1857.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access, 5*, 16568–16575.

Lin, Z., Feng, W., Liu, Y., Ma, C., Arefan, D., Zhou, D., Cheng, X., Yu, J., Gao, L., & Du, L. (2021). Machine learning to identify metabolic subtypes of obesity: A multicenter study. *Frontiers in Endocrinology, 12*, 1–12.

Liu, J., Wang, X., Cheng, Y., & Zhang, L. (2017). Tumor gene expression data classification via sample expansion-based deep learning. *Oncotarget, 8*(65), 1–15.

Lu, M., Huang, J. Z., & Qian, X. (2016). Sparse exponential family principal component analysis. *Pattern Recognition, 60*, 681–691.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics, 18*(5), 851–869.

Montanari, A., & Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *Annals of Statistics, 49*(1), 321–345.

Moore, Z. E. H., & Patton, D. (2019). Risk assessment tools for the prevention of pressure ulcers. *Cochrane Database of Systematic Reviews, 2019*(1), 1–44.

Nobre, J., & Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. *Expert Systems with Applications, 125*, 181–194.

Ouyang, F. S., Guo, B. L., Ouyang, L. Z., Liu, Z. W., Lin, S. J., Meng, W., Huang, X. Y., Chen, H. X., Qiu-Gen, H., & Yang, S. M. (2019). Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. *European Journal of Radiology, 113*, 251–257.

Palatucci, M., & Mitchell, T. M. (2007). Classification in very high dimensional problems with handfuls of examples. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 212–223).

Rahangdale, S., Singh, Y., Upadhyay, P. K., & Koutu, G. K. (2021). Principal component analysis of JNPT lines of rice for the important traits responsible for yield and quality. *Indian Journal of Genetics and Plant Breeding, 81*(1), 127–131.

Rukhsar, L., Bangyal, W. H., Ali Khan, M. S., Ag Ibrahim, A. A., Nisar, K., & Rawat, D. B. (2022). Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification. *Applied Sciences, 12*(4), 1–17.

Savargiv, M., Masoumi, B., & Keyvanpour, M. R. (2021). A new random forest algorithm based on learning automata. *Computational Intelligence and Neuroscience, 2021*(1), 1–19.

Shao, Y., & Lunetta, R. S. (2012). Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS Journal of Photogrammetry and Remote Sensing, 70*, 78–87.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1589–1604.

Sun, J., & Ongsomwang, S. (2023). Optimal parameters of random forest for land cover classification with suitable data type and dataset on Google Earth Engine. *Frontiers in Earth Science, 11*, 1–17.

Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies, 19*(1), 3–26.

Xia, Y., Ji, Z., Krylov, A., Chang, H., & Cai, W. (2017). Machine learning in multimodal medical imaging. *BioMed Research International, 2017*, 1–2.

Xu, B., Huang, J. Z., Williams, G., Wang, Q., & Ye, Y. (2012). Classifying very high-dimensional data with random forests built from small subspaces. *International Journal of Data Warehousing and Mining (IJDWM), 8*(2), 44–63.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends, 1*(1), 56–70.

Zekić-Sušac, M., Pfeifer, S., & Šarlija, N. (2014). A comparison of machine learning methods in a high-dimensional classification problem. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy, 5*(3), 82–96.

Zhang, Y. (2013). *Sparse principal component analysis: Algorithms and applications* (Doctoral dissertation, University of California, Berkeley).

Zhu, Y., Rakov, V. A., Tran, M. D., & Nag, A. (2016). A study of National Lightning Detection Network responses to natural lightning based on ground truth data acquired at LOG with emphasis on cloud discharge activity. *Journal of Geophysical Research: Atmospheres, 121*(24), 14–651.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265–286. https://doi.org/10.1198/106186006X113430

AUTHORS PROFILE

Zahwa Ahmed Jasim Obtained A Bachelor's Degree in Applied Statistics (Good Grade) in 2003 from the College of Administration and Economics - University of Baghdad, and a Master's Degree in Statistics (2026) from the Universiti Teknologi MARA includes her Master's thesis. Enhancing Classification Performance on High Dimensional Data Using an Improved Procedure of Sparse Principal Component Analysis (PCA) and Random Forest

LIST OF PUBLICATION:

Zahwa (2024). Comparison between Principal Component Analysis and Sparse Principal Component Analysis as Feature Dimension Techniques in High Dimensional Data Classification