

**APPLICATIONS OF MACHINE LEARNING IN
CATALYST INFORMATICS AND EXPLORING NEW
HYBRID TECHNIQUES OF RANDOM PROJECTION
FOR DIMENSIONALITY REDUCTION**

BY

YUSSIF YAHAYA

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS

MAY 2024

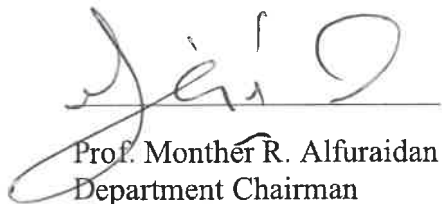
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **YUSSIF YAHAYA** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS**.

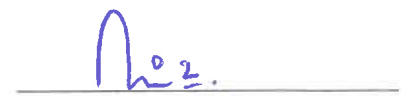
Thesis Committee


Prof. Monther R. Alfuraidan
Department Chairman


Prof. Suliman S. Al-Homidan
Dean of Graduate Studies

Date


Dr. Jimoh Olawale Ajadi
(Advisor)


Prof. Muhammad Riaz
(Member)


Dr. Gazali Tanimu
(Member)



©Yussif Yahaya
2024

This is dedicated to my family, whose support has been my strength throughout my academic journey. Thank you for everything.

ACKNOWLEDGMENTS

I would like to begin by expressing my heartfelt thanks to the Department of Mathematics for granting me this exceptional opportunity through a scholarship and support that enabled my studies. My deepest gratitude goes to my advisor, Dr. Jimoh Ajadi, for his outstanding guidance and support throughout this research journey. I am also thankful to my thesis committee members for their valuable contributions and advice. Special appreciation goes to Dr. Gazali Tanimu for his contributions to the interdisciplinary nature of this thesis through his expertise, assistance, and guidance. I also want to acknowledge all my professors and colleagues for creating a vibrant classroom environment and for always being available for advice and support. Lastly, I am profoundly grateful to my parents and family. Your unwavering support of my ambitions has been my driving force, and I would not be where I am today without you.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	vi
LIST OF TABLES	xii
LIST OF FIGURES	xiv
ABSTRACT (ENGLISH)	xvi
ABSTRACT (ARABIC)	xviii
CHAPTER 1 INTRODUCTION	1
1.1 Motivation	1
1.2 Random Projection	2
1.3 Machine Learning	3
1.4 Objectives	4
1.5 Thesis Layout	5
CHAPTER 2 THEORETICAL BACKGROUND AND LITERATURE REVIEW	6
2.1 Overview of RPs	6
2.2 Types of Random Projections	9
2.2.1 Normal Projections	9
2.2.2 Plus-Minus One Projections	10
2.2.3 Achlioptas	10
2.2.4 Li Random Projection	11

2.3	Related works on Random Projection	12
2.4	Overview of ML Algorithms Employed	14
2.4.1	Multiple Linear Regression (MLR)	14
2.4.2	Ridge Regression (Ridge)	15
2.4.3	Gaussian Process Regression (GPR)	15
2.4.4	K-Nearest Neighbor (K-NN)	15
2.4.5	Random Forest (RF)	16
2.4.6	Quantile Random Forest (QRF)	16
2.4.7	Extreme Gradient Boosting Linear (XGBL)	17
2.4.8	Boosted Tree (BT)	17
2.4.9	Support Vector Machine Radial (SVMR)	18
2.5	ML Modeling Framework	18
2.5.1	Justification of employing more than one ML Algorithm	18
2.5.2	ML Evaluation	19
2.5.3	Model Performance Evaluation	20
2.6	Related Works on Catalysts Informatics	21

CHAPTER 3 HYBRID RANDOM PROJECTION: INTEGRATING DENSE AND SPARSE TECHNIQUES FOR ENHANCED REPRESENTATION IN HIGH-DIMENSIONAL DATA 23

3.1	Objective	23
3.2	Introduction	24
3.3	Formulation of the Proposed HRP Method	25
3.3.1	Obtaining the Optimal Value of α	26
3.3.2	Supporting Lemmas	29
3.4	Simulation Study and Results	40
3.4.1	Simulation Study	40
3.4.2	Simulation Results	42
3.5	Examples	43
3.6	Summary	54

CHAPTER 4	ENHANCING HIGH-DIMENSIONAL DATA REPRESENTATION: A COMPREHENSIVE STUDY OF INTEGRATED RANDOM PROJECTION METHODS	55
4.1	Introduction	55
4.2	Formulation of the Proposed Methods	56
4.3	Simulation Study and Results	57
4.3.1	Simulation Study	57
4.3.2	Simulation Results	58
4.4	Summary	65
CHAPTER 5	DEVELOPING MACHINE LEARNING MODELS FOR CATALYSTS IN OXIDATIVE DEHYDROGENATION OF N-BUTANE	66
5.1	Objective	66
5.2	Introduction	67
5.3	Experimental Setup for Dataset Generation	71
5.4	Descriptors and Measured Variables	71
5.4.1	Outcome (measured) variables	71
5.4.2	Explanatory (descriptor) variables	72
5.5	Statistical Analysis	73
5.5.1	Data Description.	73
5.5.2	Data Exploration.	74
5.5.3	Models Development	76
5.5.4	Data Scaling	77
5.5.5	Model Cross-Validation	78
5.5.6	Optimal feature selection and hyperparameters	78
5.5.7	Performance Evaluation	80
5.6	Results and Discussion	80
5.6.1	Descriptive Statistics	80
5.6.2	Model Comparisons for Predicting Butane Conversion	81

5.6.3	Model Comparisons for Predicting 1,3-Butadiene Selectivity . .	83
5.6.4	Best Performing Model	85
5.7	Summary	88

CHAPTER 6 MACHINE LEARNING ADVANCEMENTS IN CATALYSIS: AN ENSEMBLE MODEL APPROACH TO PREDICT DEHYDROGENATION YIELD 90

6.1	Objective	90
6.2	Introduction	91
6.3	Experimental Setup for Dataset Generation	93
6.3.1	Measurement and Calculation of Outcome Variables in ODH of n-Butane	94
6.3.2	Descriptor Variables Influencing the Performance of ODH Catalysts	96
6.4	Statistical Analysis	97
6.4.1	Data Exploration	97
6.4.2	Models Development	99
6.4.3	Data Scaling	99
6.4.4	Model Cross-Validation	100
6.4.5	Optimal Feature Selection and Hyperparameters	100
6.5	Results and Discussion	101
6.5.1	Descriptive Statistics	101
6.5.2	Model Performance	103
6.6	Summary	108

CHAPTER 7 CONCLUSION AND FUTURE WORK 110

7.1	Conclusion	110
-----	----------------------	-----

APPENDIX A SUPPLEMENTARY MATERIAL FOR HYBRID RANDOM PROJECTION: INTEGRATING DENSE AND SPARSE TECHNIQUES FOR ENHANCED REPRESENTATION

IN HIGH-DIMENSIONAL DATA	116
A.1 Supplementary Information for Results on Simulation	116
REFERENCES	129
VITAE	144

LIST OF TABLES

3.1	Mean and standard error (SE) for the three RP methods employed with dimensions, $d = 500$ across various sample sizes (n), densities (δ), and <i>epsilons</i> (ϵ)	44
3.2	Mean and standard error (SE) for the three RP methods employed with dimensions, $d = 1000$ across various sample sizes (n), densities (δ), and <i>epsilons</i> (ϵ)	45
4.1	Average distortions for the Ten RP methods employed with dimensions, $d = 500$ across sample sizes (n) = 50 and 400 with (δ) = 20%	60
4.2	Average distortions for the Ten RP methods employed with dimensions, $d = 1,000$ across sample sizes (n) = 50 and 400 with (δ) = 20%	61
4.3	Average distortions for the ten RP methods employed with dimensions, $d = 15,000$ across sample sizes (n) = 50 and 400 with (δ) = 20%	62
5.1	The optimal hyperparameter searches for the ML for predicting both Feed Conversion and Butadiene Selectivity	79
5.2	The correlation matrix of the ten descriptor variables.	81
5.3	Variables selected by each ML model for predicting butane conversion.	83
5.4	Selected features for predicting 1,3 BD selectivity using the models	85
6.1	The optimal hyperparameter searches for the individual ML models in predicting the DH yield	101
6.2	Performance evaluation of the individual models and the ensemble for predicting DH Yield	104

A.1	The time taken to calculate each average distortion and the standard error for each method employed for $d = 500$	118
A.2	The time taken to calculate each average distortion and the standard error for each method employed for $d = 10000$	119
A.3	Optimal Alpha and Standard Deviation for Optimal Alpha for The HRP when $d = 500$	121
A.4	Optimal α and Standard Deviation for Optimal α for The HRP when $d = 1000$	122
A.5	Optimal α and Standard Deviation for Optimal α The HRP when $d = 5000$	123
A.6	Optimal α and Standard Deviation for Optimal α for The HRP when $d = 10000$	124

LIST OF FIGURES

2.1	General predictive machine Learning process	18
3.1	Comparison distortion ratios across different methods of RP for $d =$ 5000	46
3.2	Comparison distortion ratios across different methods of RP for $d =$ 10,000.	46
3.3	Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for toxicity data	49
3.4	Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for period changer data	50
3.5	Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for fashion mnist data	51
3.6	Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for digit mnist data	52
3.7	Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for HAR data	53
4.1	Comparison of distortions ratios for the ten RP methods employed for ($\delta = 0.8$) and $d = 500$	63
4.2	Comparison of distortions ratios for the ten RP methods ($\delta = 0.8$) and $d = 15000$	64
5.1	Histograms of variables	75
5.2	Scatter plots of Butane Conversion and predictor variables	76
5.3	Scatter plots of 1,3 BD Selectivity and predictor variables	77

5.4	Performance of n-butane conversion-based models.	82
5.5	Performance of 1,3-Butadiene selectivity-based models	85
5.6	SVMR Models performance	86
5.7	Parity plots of the SVMR model for butane conversion and BD selectivity	87
6.1	Pairwise Relationships and Distributions of Selected Variables	98
6.2	Heatmap of the lower triangular correlation matrix for predictor and outcome variables	102
6.3	Coefficient of determination (R^2) for the ML Models	105
6.4	Feature importance analysis from the ML models	106
6.5	Parity plots for BT, XGBL, SVMR and Ensemble Models	107
A.1	average distortion Box Plot for PMRP,NRP and HRP methods	125
A.2	Comparison of Standard Errors Across Different Methods of Random Projection for $d=5000$	126
A.3	Comparison of Standard Errors Across Different Methods of Random Projection for $d=10000$	127

THESIS ABSTRACT

NAME: Yussif Yahaya

TITLE OF STUDY: Applications of Machine Learning in Catalyst Informatics
and Exploring New Hybrid Techniques of Random Pro-
jection for Dimensionality Reduction

MAJOR FIELD: Applied Statistics

DATE OF DEGREE: May, 2024

This thesis explores new methods of Random Projection for dimensionality reduction of high-dimensional datasets and applications of machine learning in predictive modeling. The first main chapter introduces a Hybrid Random Projection method (HRP), which combines the strengths of Normal Random Projection (NRP) and Plus-Minus One Random Projection (PMRP), improving data dimensionality reduction while preserving data structure. HRP's effectiveness is evaluated using various scenarios of datasets and factors such as sample size, data dimensions, and sparsity. The evaluation metric, distance distortion, shows that HRP consistently outperforms NRP and PMRP, making it a significant advancement in dimensionality reduction techniques with the potential for practical applications in managing complex high-dimensional

data. The next main chapter builds upon the structure of HRP and extends the integration to several established Random Projection methods. The evaluation metric proved that the integrated techniques performed better than the individual methods. The third main chapter of the thesis focuses on the development of a Machine Learning model to predict catalyst performance, with a specific emphasis on the synthesis of 1,3-butadiene, a critical component in rubber and tire production. Among the models explored, the Support Vector Machine with Radial Basis Function (SVMR) demonstrates exceptional performance in predicting *n*-butane conversion and 1,3-butadiene selectivity. The final main chapter of the thesis focuses on using machine learning to predict dehydrogenation product (DH) yields and identify factors affecting product distribution. The study uses a dataset from experiments in a fixed-bed reactor. Three supervised machine learning models (boosted tree, extreme gradient boosting, support vector machine) were evaluated. The ensemble technique of these models showed exceptional accuracy. These findings have significant implications for catalysis research and industrial applications, offering insights into optimizing DH products and designing advanced catalysts, as well as providing a powerful tool for managing complex high-dimensional data in practical applications.

ملخص الرسالة

الاسم: ايسف يحيى

عنوان الدراسة: تطبيقات التعلم الآلي في المعلوماتية المحفزة واستكشاف تقنيات هجينة جديدة للإسقاط العشوائي لتقليل الأبعاد

التخصص: الإحصاء التطبيقي

تاريخ الدرجة العلمية: مايو 2024 م

تركز هذه الأطروحة على استكشاف طرق جديدة للإسقاط العشوائي وذلك عن طريق تقليل الأبعاد لمجموعات البيانات عالية الأبعاد وتوضح تطبيقات التعلم الآلي في النمذجة التنبؤية. يقدم الفصل الرئيسي الأول طريقة الإسقاط العشوائي الهجين ، و التي تجمع بين نقاط قوة الإسقاط العشوائي العادي والإسقاط العشوائي الزائد والناقص مما يساعد على تحسين تقليل أبعاد البيانات مع الحفاظ على بنية البيانات. تم تقييم فعالية طريقة الإسقاط العشوائي الهجين وذلك باستخدام سيناريوهات مختلفة لمجموعات البيانات وعوامل متعددة مثل حجم العينة وأبعاد البيانات وتأثيرها. يتوضح لنا أنه باستخدام مقياس التقييم (تشويه المسافة) تفوق الإسقاط العشوائي الهجين على الإسقاط العشوائي العادي والإسقاط العشوائي الزائد والناقص وذلك يجعله متقدمًا في تقنيات تقليل الأبعاد مع إمكانية التطبيق العملي لإدارة البيانات المعقدة عالية الأبعاد. يوضح الفصل الرئيسي التالي اعتماده على هيكل وبنية الإسقاط العشوائي الهجين ويقوم بتوسيع التكامل ليشمل العديد من طرق الإسقاط العشوائي المعمول بها. يظهر تفوق التقنيات المتكاملة على الطرق الفردية وذلك باستخدام مقياس التقييم. ويركز الفصل الرئيسي الثالث من الأطروحة على تطوير نموذج التعلم الآلي للتنبؤ بأداء المحفز، مع التركيز بشكل خاص على تركيب 1،3-بوتادين، وهو عنصر حاسم في إنتاج المطاط والإطارات. كما أظهرت آلة ناقل الدعم ذات وظيفة الأساس الشعاعي تفوقًا على جميع النماذج التي تم استكشافها، وتقديمها أداءً استثنائيًا في التنبؤ بتحويل ن-بيوتان و1،3-بوتادين. يركز الفصل الرئيسي الأخير من الأطروحة على استخدام التعلم الآلي للتنبؤ بإنتاجية منتج إزالة الهيدروجين وتحديد العوامل التي تؤثر على توزيع المنتج. استخدمت الدراسة مجموعة بيانات من التجارب التي أجريت في مفاعل ذي طبقة ثابتة و تم تقييم ثلاثة نماذج للتعلم الآلي (الشجرة المعززة، وتعزيز التدرج الشديدي، وآلة ناقل

الدعم). أظهرت تقنية تجميع هذه النماذج دقة استثنائية حيث كان لها الأثر الكبير على أبحاث الحفز والتطبيقات الصناعية وذلك من خلال إسهامها في تحسين المنتجات وتصميم المواد الحفازة المتقدمة، بالإضافة إلى أنها توفر أداة قوية لإدارة البيانات المعقدة عالية الأبعاد في التطبيقات العملية.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Random Projection (RP) methods have become an important tool for dimensionality reduction in high-dimensional data, each bearing unique benefits and trade-offs. On one hand, dense random projections, such as normal random projections, are advantaged for their robust theoretical guarantees for the preservation of pairwise distances. However, this might lead to potential difficulties in interpretation due to the density of the resulting matrices. On the other hand, [1] and [2] random projections make use of sparse matrices, offering more interpretability but at the expense of potentially less precise distance preservation. Considering these diverse characteristics, this thesis proposed novel methods of random Projection by integrating the strengths of two distinct random projection techniques. This combination aims to establish a balance between the preservation of distances and interpretability, making it adaptable to different data properties and analytical tasks.

Additionally, In recent years, there has been a growing global demand for 1,3-butadiene, primarily used as a raw material in the synthesis of rubber and automobile tires. The global market for 1,3-butadiene is expected to reach 19.6 million tons by 2027, with an estimated annual growth rate of 4.5% from 2020 to 2027. [3]. Traditionally, 1,3-butadiene is a byproduct of naphtha steam cracking, a process predominantly yielding light olefins like ethylene and propylene. However, the industry’s shift towards shale gas and ethane as feedstocks for steam crackers has resulted in a reduction in 1,3-butadiene production. Consequently, there has been a growing impetus to explore alternative on-purpose 1,3-butadiene production processes. One promising avenue under investigation involves the application of catalyst informatics, which integrates data science into the realm of catalysis. This approach facilitates the design of catalysts tailored to specific catalytic reactions, catalyst structures, and reaction mechanisms [4], [5]. As such, this thesis endeavors to develop predictive models employing Machine Learning (ML) algorithms for the prediction of catalyst performance in the Oxidative Dehydrogenation of n-butane.

1.2 Random Projection

Random Projection involves projecting high-dimensional data onto a lower-dimensional subspace by using a random matrix [6], [7]. The primary aim is to retain the specific geometric attributes of the data while simultaneously reducing its dimensionality [1]. Despite its apparent simplicity, RP exhibits remarkable computational efficiency and maintains satisfactory accuracy in dimensionality reduction

tasks for high-dimensional datasets [8]. This efficacy is rooted in the core concept of the Johnson-Lindenstrauss (JL) lemma, which was introduced by Johnson and Lindenstrauss in their seminal 1984 work [9]. They proved that by projecting high-dimensional data onto a random subspace of sufficiently high dimensions, the pairwise distances between points are approximately preserved, thereby minimising information loss. This finding forms the basis for the application of RP in various domains. A more detailed mathematical exposition of this lemma is available in the literature [10], [11]. Frankl and Maehara [10] provided a succinct proof of the JL lemma and employed it to establish upper bounds on the sphericity of graphs. Furthermore, Dasgupta and Gupta [11] offered an alternative proof of the JL lemma by employing elementary probabilistic techniques.

1.3 Machine Learning

Machine learning has revolutionized many aspects of human activity and had a profound impact on a wide range of fields since its inception. This change is a result of machine learning techniques being widely used in many different sectors. Arthur Samuel's definition of Machine Learning in 1959 characterizes it as a field of study that empowers computers to acquire knowledge independently, without the need for explicit programming [12]. Machine Learning, a specific domain within the broader subject of Artificial Intelligence (AI), is primarily concerned with the development of algorithms and models that facilitate the extraction of information from data, hence enabling computers to make predictions or judgments [13]. In the field of machine

learning, models are designed to extract patterns from data and utilize this acquired knowledge to make predictions on unseen data. This characteristic sets Machine Learning apart from conventional programming, which relies on explicit instructions for operation.

Machine Learning has resulted in a significant transformation across various disciplines. For instance, it has enhanced the capacity to predict Catalyst Performance, as evidenced by the findings of [14]. Other studies, such as [15], [16], [17], [18], [19], [20] have also applied machine learning in catalysis research. In the field of health, machine learning has been employed by [21], [22], [23], [24], [25], [26], [27]. In bio-informatics, machine learning has been utilized by [28], [29], [30], [31], [32], [32], [33]. Applied Machine learning has also been employed by [34], [35], [36] in their research in agriculture. Machine learning has also played a crucial role in driving progress in the fields of image and speech recognition. Notable research papers include [37], [38]. The aforementioned examples serve as a limited representation of the wide-ranging influence that machine learning exerts across diverse domains.

1.4 Objectives

This thesis utilizes machine learning to develop predictive models based on data generated from a fixed bed reactor. These models aim to predict catalysis performance, offering significant potential for improving catalytic processes in industries and catalysis research. Additionally, in the era of big data, high-dimensional datasets are preva-

lent across various fields. Addressing the challenge of high dimensionality is crucial to effectively utilize these datasets without excessive computational resources. This thesis is motivated by the need to develop an innovative random projection technique that efficiently reduces dimensionality while preserving essential information, thereby enhancing computational efficiency. The primary objective of this thesis is to:

- Propose novel methods in random projection that combines two separate established random projection techniques.
- Develop predictive models using machine learning algorithms to predict performance of catalyst

1.5 Thesis Layout

In Chapter 2, we explored the theoretical background and related works of the thesis. Chapter 3 contains details on random projections for reducing the dimensionality of high-dimensional datasets and introduces our novel random projection technique. Chapter 4 extends our proposed random projection technique. We utilized machine learning algorithms to develop models for predicting both 1,3-butadiene selectivity and n-butane conversion, as discussed in Chapter 5, as well as dehydrogenation yield, as discussed in Chapter 6. Finally, Chapter 7 presents our conclusions and outlines future research directions.

CHAPTER 2

THEORETICAL BACKGROUND AND LITERATURE REVIEW

This chapter presents the theoretical background information for the thesis. Section 2.1 introduces the fundamental theory supporting random projections. Section 2.2 describes the different forms of random projections and offers a comparative overview of scenarios suited for each. Section 2.3 presents related works on random projections. Section 2.4 provides an overview of machine learning and the types of machine learning algorithms used in the thesis. Section 2.5 presents the modeling framework. Finally, Section 2.6 provides a related works on catalyst informatics.

2.1 Overview of RPs

RPs play a pivotal role in dimensionality reduction, a process aimed at discovering a lower-dimensional embedding $\mathbf{Y} \subset \mathbb{R}^{n \times k}$ of an original set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbb{R}^{n \times d}$, where n is the number of data points, d is the original dimension of (or number

of features in) data X and k is the reduced dimensionality after applying the random projection. *i.e.* the dimension of Y [39]. The theoretical foundation of RPs is grounded in the JL lemma [9], which provides a probabilistic guarantee for preserving pairwise Euclidean (l_2) distances in the reduced-dimensional space.

The JL lemma states that for any given error tolerance $0 < \epsilon < 1$ and an integer n representing the number of samples or data points in the original high-dimensional dataset $\mathbf{X}$, there exists a linear map $f : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times k}$ such that, for all $\mathbf{x}, \mathbf{y}$ in the set, the pairwise distances are approximately preserved up to a factor of $1 \pm \epsilon$.

$$(1 - \epsilon)\|\mathbf{x}_i - \mathbf{x}_l\|^2 \leq \|\mathbf{R}(\mathbf{x}_i - \mathbf{x}_l)\|^2 \leq (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_l\|^2$$

where $\|\cdot\|$ is the Euclidean norm, $\|\mathbf{x}_i - \mathbf{x}_l\|^2$ is the pairwise distances between the i th and l th columns of $\mathbf{X}$, $i \neq l$, $\mathbf{x}_i$ and $\mathbf{x}_l$ are data points in the original space, and $\mathbf{R}(\mathbf{x}_i - \mathbf{x}_l)$ denote their transformation. This implies that the reduced-dimensional representation $\mathbf{Y}$ retains approximately the geometric structure of the original dataset $\mathbf{X}$.

The lemma guides the choice of k to ensure distance preservation within ϵ , balancing the computational efficiency and information retention. The dimension k of the embedding space is determined as [40]:

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log(d) \quad (1)$$

where ϵ represents the allowable distortion level in pairwise distances after the pro-

jection. A smaller ϵ ensures tighter preservation of distances but requires a higher k , leading to increased computational complexity. Conversely, a larger ϵ may allow for a smaller k , sacrificing some accuracy for the computational efficiency. The practical application of RPs involves empirical validation, where k is varied to observe its impact on specific tasks such as classification or clustering. This empirical approach, coupled with the theoretical guidance of the JL lemma, enables the selection of an optimal k that accommodates both computational constraints and data integrity requirements of the task at hand.

Once k is determined, the implementation of RP involves generating a projection matrix $\mathbf{R}$ of size $d \times k$ with elements drawn from a specific probability distribution, such as a normal or plus-minus one RP. The original data matrix $\mathbf{X}$ of dimension d is then projected into a lower-dimensional space of dimension k , forming a new matrix $\mathbf{Y}$:

$$\mathbf{X} \in \mathbb{R}^{n \times d}, \quad \mathbf{R} \in \mathbb{R}^{d \times k} \quad \Rightarrow \quad \mathbf{Y}^{n \times k} = \mathbf{X} \cdot \mathbf{R} \in \mathbb{R}^{n \times k}$$

where $\mathbf{R}$ is the RP matrix and $\mathbf{Y}$ represents the projected data in the lower-dimensional space. This formulation encapsulates the essence of RPs in dimensionality reduction, providing a robust and theoretically sound approach to managing high-dimensional datasets.

Random projections can be visualized in the following manner:

$$\begin{array}{ccc}
\mathbf{X} \text{ where } \mathbf{X} \in \mathbb{R}^{n \times d} & \mathbf{R} \text{ where } \mathbf{R} \in \mathbb{R}^{d \times k} & \mathbf{Y} \text{ where } \mathbf{Y} \in \mathbb{R}^{n \times k} \\
\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} & * \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{d1} & r_{d2} & \dots & r_{dk} \end{bmatrix} & = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1k} \\ y_{21} & y_{22} & \dots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nk} \end{bmatrix}
\end{array}$$

2.2 Types of Random Projections

In this section, we explore the four main types of random projections we integrated to form our proposed methods: Normal, Plus-Minus One, Li, and Achlioptas projections.

2.2.1 Normal Projections

Normal (or Gaussian) projection (NRP), a cornerstone of dimensionality reduction, utilises a random matrix whose elements are drawn independently from a normal distribution [41]. The defining attributes of this distribution are its mean (μ) and variance (σ^2), which are set as 0 and 1, respectively. This configuration ensures a symmetric distribution of values around zero and a moderate dispersion, which is crucial for maintaining the data structure integrity in the reduced-dimensional space. Mathematically, for a data matrix $\mathbf{X}$ in a d -dimensional space, a projection matrix $\mathbf{R}$ of size $k \times d$ transforms $\mathbf{X}$ into a k -dimensional space. Each element R_{ij} of $\mathbf{R}$ follows:

$$R_{ij} \sim \mathcal{N}(0, 1)$$

2.2.2 Plus-Minus One Projections

Plus-Minus One (or Bernoulli) projection (PMRP), another dimensionality reduction approach, is characterised by its binary projection matrix with elements either +1 or -1 [2]. This discrete nature of element selection streamlines the process and is particularly effective for certain data types, such as single-nucleotide polymorphism (SNP) data [42], [43]. Each SNP can have two possible alleles (A or T; C or G), binary sensor readings indicating normal (1) or anomalous (0) behavior, bag-of-words representations of text data where the presence (1) or absence (0) of a word in a document is considered, striking a balance between simplicity and data preservation. Each element R_{ij} in $\mathbf{R}$ is determined as follows.

$$R_{ij} \sim \begin{cases} +1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2} \end{cases}$$

2.2.3 Achlioptas

Achlioptas or Sparse Random projection [2] is a specialized form of random projection that employs a distinct probability distribution for generating the random projection matrix. Unlike methods that rely on Gaussian or Rademacher distributions, Achlioptas random projection utilizes a sparse random matrix where the elements can take on one of three values: -1, 0, or +1. These values are selected with probabilities 1/6, 1/6 and 2/3 respectively. These probabilities are intentionally chosen to ensure that the resulting projection matrix exhibits certain desired properties. Notably, the inclusion

of 0 with a higher probability leads to a sparse projection matrix, which significantly enhances computational efficiency. This makes Achlioptas Random Projection particularly well-suited for handling large datasets where computational complexity is a critical concern. Mathematical Expression:

$$r_{ij} \sim \sqrt{3} \begin{cases} 1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

2.2.4 Li Random Projection

The method known as Li or Very Sparse random projection, as introduced by Li et al. in their seminal work [1], strategically introduces sparsity into the projection matrix. Consequently, the resulting dataset is characterized by a notably higher degree of sparsity. It's worth noting that, in this method, the values of the projection matrix elements are drawn according to specific probabilities, as outlined in the corresponding mathematical expression below.

Mathematical Expression:

$$r_{ij} \sim \sqrt{s} \begin{cases} +1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ -1 & \text{with probability } \frac{1}{2s} \end{cases}$$

In the above expression, s is a parameter that determines the level of sparsity in the

random projection. Specifically, s is used to control the probabilities associated with each of the three possible values that r_{ij} can take on (+1, 0, or -1). In summary, as s increases, the probabilities associated with +1 and -1 decrease, leading to a sparser projection matrix. Conversely, as s decreases, the probabilities of +1 and -1 increase, resulting in a less sparse projection matrix.

2.3 Related works on Random Projection

Random Projection was introduced by Johnson and Lindenstrauss in their seminal work in 1984. They proved that by projecting high-dimensional data onto a random subspace of sufficiently high dimension, the pairwise distances between points are approximately preserved. This foundational result forms the basis for the application of Random Projection in various domains. The paper titled "The Johnson-Lindenstrauss Lemma and the Sphericity of Some Graphs" delivers a succinct proof of the Johnson-Lindenstrauss lemma and employs it to establish upper bounds on the sphericity of graphs [10]. Additionally, Dasgupta and Gupta offer an alternative proof of the Johnson-Lindenstrauss lemma, employing elementary probabilistic techniques [11]. Since its publication in 2002, this paper has garnered substantial citations, highlighting its profound influence in affirming the lemma's validity. Subsequent research has extended and built upon the original findings. For instance, Achlioptas's 2002 work utilizes the Johnson-Lindenstrauss lemma to devise a randomized algorithm for approximating distances in high-dimensional spaces [2]. In 2006, Ailon and Chazelle further leveraged the lemma to develop a deterministic algorithm for the same pur-

pose [44]. Li et al. [1] put forth an algorithm for estimating distances between points in high-dimensional vector spaces through the use of highly sparse random projections. Their work builds upon prior research in natural language processing and data mining, encompassing probabilistic analysis, clustering, and classification. The paper offers both theoretical insights and empirical evidence to underscore the effectiveness of the algorithm across diverse applications. This algorithm involves random sampling of entries from a matrix, taking values in $-1, 0, 1$, with specific probabilities. This approach yields substantial enhancements in processing speed and memory utilization compared to conventional random projections. The authors demonstrate its applicability to tasks like clustering and classification, substantiating their claims with experimental results. In summary, Random projection methods have become an important tool for dimensionality reduction in high-dimensional data, each bearing its unique benefits and trade-offs. On one hand, dense random projections, such as the normal random projection, are advantaged for their robust theoretical guarantees for the preservation of pairwise distances. However, this might lead to potential difficulties in interpretation due to the density of the resulting matrices. On the other hand, Li et al.[1] and Achlioptas[2] random projections make use of sparse matrices, offering more interpretability but at the expense of potentially less precise distance representations. Considering these diverse characteristics, our hybrid approach integrates the strengths of two distinct random projection techniques. This combination aims to establish a balance between the preservation of distances and interpretability, making it adaptable to different data properties and analytical tasks.

2.4 Overview of ML Algorithms Employed

There are various types of machine learning algorithms available. The selection of an appropriate algorithm depends on the specific problem being addressed. In this thesis, we have chosen algorithms that cover both linear and non-linear approaches, considering the characteristics of our datasets. Among the linear methods, we considered Multiple Regression and Ridge Regression. Regarding kernel approaches, our evaluation included Gaussian Process Regression, Support Vector Machine with a Linear Kernel, Support Vector Machine with a Radial Kernel, and K-Nearest Neighbor. In relation to tree ensemble techniques, our analysis encompassed Random Forest, Boosted Tree, Quantile Random Forest, and eXtreme Gradient Boosting.

2.4.1 Multiple Linear Regression (MLR)

MLR [45] is employed when there are multiple predictor variables, aiming to establish a relationship between these descriptors and the response variables. This model endeavors to determine a linear equation that best fits the data, enabling the prediction of the response variable based on the provided descriptor variables. To train the model, a dataset with known predictor and response variables is utilized. The objective is to determine the coefficients' values that minimize the disparity between the predicted and actual values. This is commonly achieved through a method known as least squares regression. Once the model is trained, it can be effectively utilized for making predictions .

2.4.2 Ridge Regression (Ridge)

Ridge regression [46], alternatively known as Tikhonov regularization, is a linear regression technique that introduces a penalty term into the objective function of Ordinary Least Squares (OLS). The penalty term exhibits a direct proportionality to the squares of the coefficients, hence impeding the development of excessively intricate models characterized by big coefficients. The utilization of this method is particularly advantageous in scenarios characterized by multicollinearity within the dataset, as this condition might lead to less dependable estimations of the regression coefficients.

2.4.3 Gaussian Process Regression (GPR)

GPR [47] is a type of Bayesian probabilistic model that can be used for regression problems. In GPR, the goal is to predict the value of a target variable y given a set of input variables x . Unlike traditional regression methods, GPR provides a full probabilistic prediction for y , which includes both a mean prediction and a measure of uncertainty. Compared to conventional regression techniques, GPR has advantages. Because it offers a complete probabilistic prediction, making decisions in the face of uncertainty is made easier. Additionally, it makes nonlinear modeling flexible.

2.4.4 K-Nearest Neighbor (K-NN)

A machine learning model called K-NN [47] can be applied to tasks involving both regression and classification. Because it is not dependent on distributions or assumptions about the normalcy of the data, it is referred to as a non-parametric tech-

nique. The number of nearest neighbors to be taken into account for prediction in a K-NN is indicated by k . The model looks for the k closest data values in the training set and determines their average value as the forecast when predicting a new value for the data. K-NN's simplicity and efficiency are among its key advantages [48], [49]. Nevertheless, when working with larger datasets or high-dimensional data, it can be computationally expensive. Additionally, choosing the right number for k necessitates considerable thought, and when the data is unbalanced or contains features that are not significant, its performance may decrease.

2.4.5 Random Forest (RF)

One machine learning (ML) technique that works well for both regression and classification applications is the Random Forest (RF) algorithm. Several decision trees are integrated in this ensemble learning technique to produce predictions that are more trustworthy and accurate [50]. The creation of many decision trees, each tree using a different collection of characteristics and training data, is the basic idea behind Random Forest (RF). Using this method, Random Forest (RF) substantially lowers the likelihood of overfitting and enhances the model's overall capacity for generalization. In the context of creating forecasts, each tree within the forest generates a prediction for each new dataset, according to [51]. The average of each individual forecast provided by each tree is used to determine the final result.

2.4.6 Quantile Random Forest (QRF)

QRF [50] is a ML model that extends the traditional random forest algorithm to

estimate not just the conditional mean but also the conditional quantiles of a response variable. While traditional random forests can predict the expected value of a target variable based on a set of input features, QRF can predict the entire distribution of the target variable, including its spread, skewness, and other statistical properties.

2.4.7 Extreme Gradient Boosting Linear (XGBL)

XGBL [52] refers to a machine learning algorithm that belongs to the family of tree ensemble models, specifically within the broader category of extreme gradient boosting (XGBoost). It is an ensemble learning method that combines the predictions of multiple weak learners (typically decision trees) to create a robust and accurate model. XGBL is an extension of the XGBoost algorithm that incorporates a linear component along with decision trees in its ensemble. This combination of linear and tree-based models enhances the algorithm's ability to capture both linear and non-linear patterns in the data, making it a versatile and powerful tool for various machine learning tasks.

2.4.8 Boosted Tree (BT)

The BT [53] algorithm is a boosting-based tree ensemble learning technique that combines numerous decision trees to increase a model's prediction accuracy. The boosting process involves training decision trees, which are weak regressors, iteratively to provide predictions based on the data. Through the boosting process, the algorithm focuses on addressing errors made by previous trees, ultimately leading to a more powerful and accurate predictive model.

2.4.9 Support Vector Machine Radial (SVMR)

SVMR [54] is a kernel-based model that employs a nonlinear kernel, specifically the Radial Basis Function (RBF) or Gaussian kernel [55]. This allows SVMR to effectively handle datasets that are not linearly separable, enabling the model to detect patterns that cannot be identified by a straight line. SVMR can be used for both classification and regression problems [56].

2.5 ML Modeling Framework

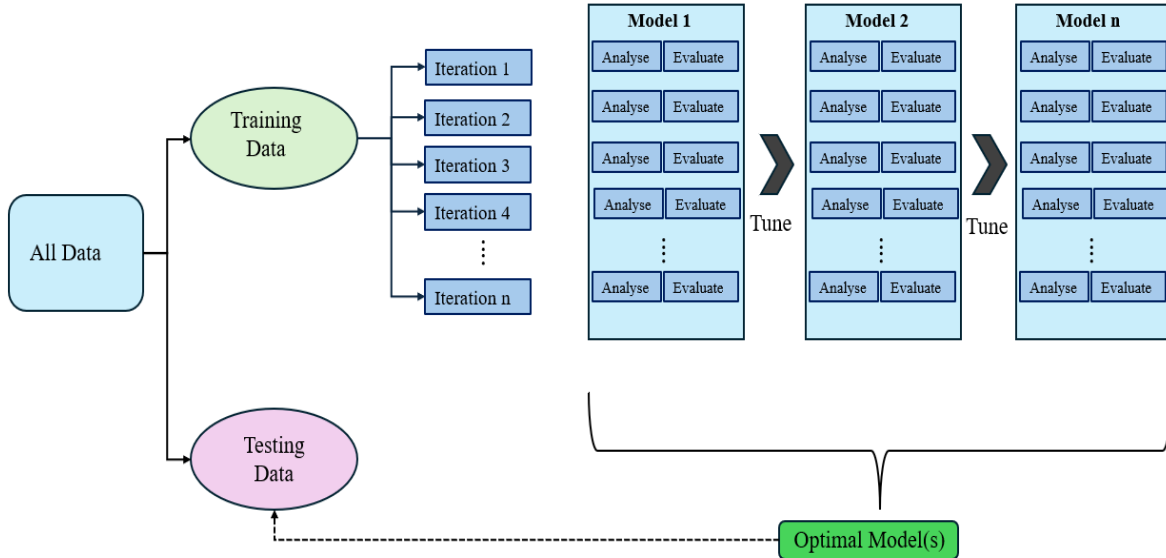


Figure 2.1: General predictive machine Learning process [57]

2.5.1 Justification of employing more than one ML Algorithm

The ML process is highly iterative and relies on heuristics. Without sufficient knowledge of the specific problem or dataset, it becomes challenging to determine the most

effective ML method to use. This concept is often referred to as the "no free lunch theorem" for ML, as articulated by [58]. As a result, it's typical for numerous ML approaches to be tested, assessed, and adjusted before arriving at an optimal model. Successfully navigating this process instills a high level of confidence in our outcomes. Conversely, failure to do so may render the results useless and potentially harmful [59].

2.5.2 ML Evaluation

After model training, error estimation is typically conducted, commonly referred to as the evaluation of residuals. This process involves determining the training error for models, which is a numerical estimate of the disparity between predicted and observed responses. However, relying solely on training error provides insight into the performance of our models exclusively on the dataset used for training. It does not offer insight into the accuracy of our models on unfamiliar datasets. This inability to discern whether models are underfitting or overfitting using training error alone necessitates the use of Cross-Validation (CV) techniques to ensure thorough model training. CV entails dividing the dataset into subsets, where the model is trained on one subset and tested on another [60]. This allows for training our model(s) on a new training set and estimating performance on a validation set. However, validation using a single holdout set can be highly variable and unreliable, particularly when working with smaller datasets. As the dataset size diminishes, this concern becomes more pronounced. Studies by [61], [62] have highlighted the variability and unreliability of single holdout set validation. To address this issue, [63] demonstrated that repeating

k-fold CV can enhance the precision of estimated generalization error. Therefore, for smaller datasets (e.g., $n < 10,000$), employing 10-fold CV repeated 5 or 10 times can improve the accuracy of estimated performance and provide an estimate of its variability. For this reason, we adopted Nested Cross-Validation (Nested-CV) in our evaluation process.

2.5.3 Model Performance Evaluation

We evaluated our models' performance using the following metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2). Our criterion for determining the best-performing model was based on minimizing error (RMSE and MAE) and maximizing R^2 .

RMSE, expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

measures the square root of the average squared differences between the observed values (y_i) and the predicted values ($\hat{y}_i$).

MAE, expressed as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

calculates the average absolute differences between the observed values (y_i) and the predicted values ($\hat{y}_i$).

R^2 , defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

measures the proportion of the variance in the dependent variable (y) that is predictable from the independent variable ($\hat{y}$). It ranges from 0 to 1, where higher values indicate better model fit.

RMSE and MAE quantify prediction errors, while R^2 assesses the explanatory power of predictor variables on the variation in the response variable.

2.6 Related Works on Catalysts Informatics

The utilisation of ML in heterogeneous catalysis started in the mid-1990s when it was utilised to enhance the efficacy of discovering new catalysts. The ML model was used to perform an extrapolative prediction of novel water gas shift (WGS) reaction catalysts [64]. The authors reported that the extra tree regression (ETR) model provided the best predictive performance for CO conversion during the WGS reaction. Similarly, ML algorithms based on sorted weighted elemental descriptor (SWED) representations showed promising results in predicting new WGS catalysts [64]. Suzuki et al. explored the use of elemental features as descriptors for the oxidative coupling of methane (OCM), water gas shift (WGS), and CO oxidation reactions using different ML models. Extreme Gradient Boosting (XGB) regression gave the best prediction of the catalytic performance with an RMSE of 4.5 for the OCM reaction [65]. In different studies, researchers predicted the performance [66], determined reaction conditions using data-driven approach [67], and uncovered hidden catalysts [68] for OCM re-

actions using the application of machine learning. In an earlier report by Madaan et al [69], the performance of n-butane in 1,3-butadiene oxidative dehydrogenation catalysts was predicted using descriptor models. The models were built based on the results of 15 bimetallic mixed-oxide-supported catalysts, and the model results were projected onto a set of mixed-oxide catalysts. Impressive predictions were obtained with > 90 accuracy for the new sets of bimetallic oxides [69]. Their work focused on utilizing radial distribution function (RDF) descriptors based on slater-type orbitals for modelling and predicting the performance of oxidative dehydrogenation of n-butane catalysts.

CHAPTER 3

HYBRID RANDOM PROJECTION: INTEGRATING DENSE AND SPARSE TECHNIQUES FOR ENHANCED REPRESENTATION IN HIGH-DIMENSIONAL DATA

3.1 Objective

The objective of this chapter is to introduce a novel hybrid Random Projection method called HRP, which integrates the strengths of the normal and plus-minus one Random

Projection methods.

3.2 Introduction

High-dimensional datasets are becoming increasingly prevalent across various fields, including computer vision, genomics, and finance. These datasets are often plagued by the "curse of dimensionality", which leads to increased computational complexity and challenges in data visualisation [8]. Dimensionality reduction techniques, which project high-dimensional data to a lower-dimensional space while preserving essential information, have become crucial for addressing these issues. Among dimensionality reduction techniques, random projection (RP) has emerged as a particularly effective method [8], [1]. RP has practical applications in various fields, including machine learning [70], [71], [8], [72], [73], facial recognition [74], and pattern recognition [48], [75]. Moreover, in the field of bioinformatics, RP is used to discover motifs in biosequences [76], [77].

Subsequent research has expanded these initial findings. The JL lemma was utilised to develop a randomised algorithm to approximate distances in high-dimensional spaces [2]. This adaptation was extended to creating a deterministic algorithm for the same purpose [44]. Additionally, a proposed algorithm utilised highly sparse random projections to estimate distances in high-dimensional vector spaces [1]. This algorithm improves the processing speed and memory utilisation. The versatility of the RP is further demonstrated by the variety of its types, each distinguished by the probability distributions from which their projection matrices are drawn. These

include normal, plus-minus one, Achlioptas, Rademacher, Li, and count sketch RPs [71], [78], [1], [79].

Although normal RP offers accurate preservation of pairwise distances, it can struggle with certain data distributions or noise levels. Conversely, plus-minus one RP, noted for its computational efficiency, may not effectively maintain complex data relationships [1]. To overcome these limitations, this study introduces a novel hybrid RP method called HRP, which integrates the strengths of the normal and plus-minus one RP methods. The proposed method incorporates a blending parameter to optimally balance the individual contributions, thereby enhancing the dimensionality reduction process while preserving the structural integrity of the data. Through a series of extensive simulations and applications to real-world datasets, we demonstrate the superiority of the HRP over its individual counterparts in both theoretical and practical contexts.

3.3 Formulation of the Proposed HRP Method

This study introduces a hybrid RP (referred to as HRP) method that blends the strengths of two established projection techniques to address their individual limitations. This hybrid approach is designed to achieve an optimal balance between distance preservation and matrix interpretability, which are crucial factors in high-dimensional data analysis. The hybrid RP is mathematically represented as follows:

$$\mathbf{H} = \alpha \mathbf{R}_1 + (1 - \alpha) \mathbf{R}_2 \quad (3.1)$$

In this equation, $\mathbf{R}_1$ and $\mathbf{R}_2$ denote the matrices obtained using the two distinct RP methods. The parameter α , ranging from 0 to 1, is a critical parameter that controls the blend of these methods in the hybrid model. We integrate NRP and PMRP to create the novel HRP. This particular combination is selected to leverage the dense, theoretically sound structure of the NRP, and the sparse, easily interpretable matrix form of the PMRP. A framework of the HRP and how to obtain an optimal value of α are discussed in the next subsection.

3.3.1 Obtaining the Optimal Value of α

The effectiveness of the HRP method depends on the optimal selection of the parameter α . The primary objective is to minimize the distortion in pairwise distances after projection, thereby enhancing the accuracy and precision of the dimensionality reduction process. We define the distance preservation ratio for two data points $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ in the context of the HRP $\mathbf{H}$ in Equation 3.2 as follows:

$$r = \frac{\|\mathbf{xH} - \mathbf{yH}\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \quad (3.2)$$

Our goal is to identify the value of α that minimizes the expectation of this distance-preservation ratio, denoted as $E[r]$. This leads to the formulation of the following optimization problem:

$$\alpha^* = \arg \min_{\alpha} E[r]$$

Given that $\mathbf{H}$ is a linear combination of $\mathbf{R}_1$ and $\mathbf{R}_2$ with coefficient α , the function

$E[r]$ is a function of α . The optimal α ensures that the HRP efficiently preserves the geometric structure of the dataset while minimizing the dimensionality reduction distortion. To solve this optimization problem, we employ numerical strategies by scanning the possible range of α within $[0, 1]$ and selecting the value that yields the lowest $E[r]$. For each value of α , the HRP method is applied to project the dataset, and the performance is evaluated based on the preservation of the pairwise distances within the projected data. The primary evaluation metric is the average distortion, computed as the average ratio between the pairwise distances in the original high-dimensional space and those in the projected lower-dimensional space. The goal is to minimize this distortion, indicating an effective adherence to the JL lemma. We execute this using the `optim` function in R, which provides a framework for various optimization methods suitable for this type of problem. By carefully setting the parameters and method within `optim`, we explore the range of α and identify the optimal value that satisfies our criteria for minimizing $E[r]$. The optimization method employed was L-BFGS-B, which stands for Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds. This variant of the BFGS algorithm is suitable for problems with variable bounds, such as our optimization problem with bounds set to $[0, 1]$. Using L-BFGS-B ensures that the solution remains within these specified bounds.

Algorithm 1 presents a framework for the proposed HRP and how to optimize α .

Algorithm 1: The proposed framework of the HRP for dimensionality reduction with optimized α

Input: $\mathbf{X}$: Input data matrix of size $(n \times d)$; k : Target dimensionality after reduction

Output: $\mathbf{Y}$: Reduced data matrix of size $(n \times k)$

Generate two random projection matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ of size $(d \times k)$ with each element drawn independently from any desired distribution;

Set a range for the parameter α (e.g., $[0, 1]$);

for *each value of α* **do**

Generate the hybrid random projection matrix $\mathbf{H}$ using Equation 3.1;

Project the input data matrix $\mathbf{X}$ onto the random projection matrix $\mathbf{H}$ to

obtain the reduced data matrix $\mathbf{Y}$: $\mathbf{Y} = \mathbf{X} \times \mathbf{H}$;

Evaluate the average distortion in pairwise distances, denoted as $E[r]$,

using the distance preservation ratio equation;

end

Select the optimal α^* that minimizes $E[r]$;

Obtain $\mathbf{H}$ using α^* , $\mathbf{R}_1$ and $\mathbf{R}_2$;

Project $\mathbf{X}$ onto $\mathbf{H}$ to obtain $\mathbf{Y}$: $\mathbf{Y} = \mathbf{X} \cdot \mathbf{H}$;

return $\mathbf{Y}$

3.3.2 Supporting Lemmas

The following lemmas are instrumental in the development and analysis of the HRP method. They provide the key mathematical foundations that underpin our approach and enable the preservation of distances during the dimensionality reduction process. Proofs for these lemmas are provided for completeness and to facilitate a deeper understanding of the HRP method.

Lemma 1: For any pair of points $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, and for any $\alpha \in [0, 1]$, the Euclidean distance between the projections of $\mathbf{x}$ and $\mathbf{y}$ using the hybrid projection $\mathbf{H}$ is less than or equal to the sum of the distances of the projections of $\mathbf{x}$ and $\mathbf{y}$ using the matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively.

Proof: Let $\mathbf{H}$ (Equation 3.1) be the HRP and $\mathbf{x}, \mathbf{y}$ be any two points in $\mathbf{X}$. Then, the Euclidean distance between the projections of $\mathbf{x}$ and $\mathbf{y}$ onto $\mathbf{H}$ is given by

$$\|\mathbf{xH} - \mathbf{yH}\|_2$$

This can be expanded as

$$\|\alpha\mathbf{xR}_1 + (1 - \alpha)\mathbf{xR}_2 - \alpha\mathbf{yR}_1 - (1 - \alpha)\mathbf{yR}_2\|_2$$

By the triangle inequality, this is less than or equal to

$$\|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1)\|_2 + \|(1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2,$$

which are the Euclidean distances between the projections of $\mathbf{x}$ and $\mathbf{y}$ onto $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively.

Therefore,

$$\|\mathbf{xH} - \mathbf{yH}\|_2 \leq \|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1)\|_2 + \|(1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2$$

Since $0 \leq \alpha \leq 1$ and $0 \leq 1 - \alpha \leq 1$, we have

$$\|\mathbf{xH} - \mathbf{yH}\|_2 \leq \|\mathbf{xR}_1 - \mathbf{yR}_1\|_2 + \|\mathbf{xR}_2 - \mathbf{yR}_2\|_2$$

Lemma 2: Hybrid Projections and Orthogonality

Suppose $\mathbf{R}_1$ and $\mathbf{R}_2$ are two RP matrices constructed via two different RP methods.

If the projections of any vector $\mathbf{x}$ are orthogonal, i.e. $(\mathbf{xR}_1)^T(\mathbf{xR}_2) = 0$, and $\mathbf{H}$ is a hybrid projection matrix given by $\mathbf{H} = \alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2$ for some $\alpha \in [0, 1]$, then for any data point $\mathbf{x} \in \mathbf{R}^d$, the following holds:

$$\|\mathbf{xH}\|_2^2 = \alpha^2\|\mathbf{xR}_1\|_2^2 + (1 - \alpha)^2\|\mathbf{xR}_2\|_2^2$$

Proof:

Starting from the left-hand side of the equation:

$$\begin{aligned}
\|\mathbf{x}\mathbf{H}\|_2^2 &= \|\mathbf{x}(\alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2)\|_2^2 \\
&= \|\alpha\mathbf{x}\mathbf{R}_1 + (1 - \alpha)\mathbf{x}\mathbf{R}_2\|_2^2 \\
&= (\alpha\mathbf{x}\mathbf{R}_1 + (1 - \alpha)\mathbf{x}\mathbf{R}_2)^T (\alpha\mathbf{x}\mathbf{R}_1 + (1 - \alpha)\mathbf{x}\mathbf{R}_2) \\
&= \alpha^2\mathbf{x}\mathbf{R}_1^T\mathbf{x}\mathbf{R}_1 + \alpha(1 - \alpha)\mathbf{x}\mathbf{R}_1^T\mathbf{x}\mathbf{R}_2 + \alpha(1 - \alpha)\mathbf{x}\mathbf{R}_2^T\mathbf{x}\mathbf{R}_1 + (1 - \alpha)^2\mathbf{x}\mathbf{R}_2^T\mathbf{x}\mathbf{R}_2
\end{aligned}$$

Since $\mathbf{R}_1^T\mathbf{R}_2 = 0$, we have $\mathbf{x}\mathbf{R}_1^T\mathbf{x}\mathbf{R}_2 = 0$ and $\mathbf{x}\mathbf{R}_2^T\mathbf{x}\mathbf{R}_1 = 0$. Thus, the equation simplifies to:

$$\|\mathbf{x}\mathbf{H}\|_2^2 = \alpha^2\|\mathbf{x}\mathbf{R}_1\|_2^2 + (1 - \alpha)^2\|\mathbf{x}\mathbf{R}_2\|_2^2$$

This shows that $\mathbf{x}\mathbf{H}$ and $\mathbf{y}\mathbf{H}$ are also orthogonal; therefore, the hybrid projection preserves the orthogonality of the vectors. This lemma establishes that if the two base RP matrices are orthogonal, the hybrid projection matrix retains this orthogonality during its operation. This property is significant because it ensures that the hybrid method does not introduce a correlation between previously uncorrelated dimensions, thereby maintaining the integrity of the data structure in reduced space.

Lemma 3: Orthogonality Preservation

If the vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ are orthogonal and $\mathbf{H}$ is a hybrid RP matrix, as defined above, then the projections of these vectors under $\mathbf{H}$ are approximately orthogonal.

Proof: Assuming that $\mathbf{v}$ and $\mathbf{w}$ are orthogonal, we have $\mathbf{v}^T\mathbf{w} = 0$. The dot product of their projections under $\mathbf{H}$ is given by:

$$(\mathbf{v}\mathbf{H})^T(\mathbf{w}\mathbf{H}) = (\alpha\mathbf{v}\mathbf{R}_1 + (1 - \alpha)\mathbf{v}\mathbf{R}_2)^T(\alpha\mathbf{w}\mathbf{R}_1 + (1 - \alpha)\mathbf{w}\mathbf{R}_2)$$

Expanding this and using the fact that $\mathbf{v}^T\mathbf{w} = 0$:

$$= \alpha^2(\mathbf{v}\mathbf{R}_1)^T(\mathbf{w}\mathbf{R}_1) + \alpha(1 - \alpha)(\mathbf{v}\mathbf{R}_1)^T(\mathbf{w}\mathbf{R}_2) + \alpha(1 - \alpha)(\mathbf{v}\mathbf{R}_2)^T(\mathbf{w}\mathbf{R}_1) + (1 - \alpha)^2(\mathbf{v}\mathbf{R}_2)^T(\mathbf{w}\mathbf{R}_2)$$

If both $\mathbf{R}_1$ and $\mathbf{R}_2$ preserve orthogonality approximately, then the dot product of the projected vectors under each of these matrices will be close to zero. Thus, the overall dot product is also close to zero, implying that $\mathbf{v}\mathbf{H}$ and $\mathbf{w}\mathbf{H}$ are approximately orthogonal. While Lemma 2 focuses on the orthogonality of the base projection matrices, Lemma 3 addresses the orthogonality of the data vectors themselves after being projected by the hybrid matrix. This lemma is crucial because it demonstrates that the hybrid projection method can preserve the orthogonality of the original data vectors, which is essential for applications in which the independence of features is critical.

Lemma 4: Distance Preservation of Hybrid Projections

Let $\mathbf{X}$ be a high-dimensional dataset in $\mathbf{R}^d$, let $\mathbf{R}_1$ and $\mathbf{R}_2$ be two different RP matrices constructed via two different RP methods, and let $\alpha \in [0, 1]$. If we denote $\mathbf{H}$ as a hybrid projection that is a linear combination of $\mathbf{R}_1$ and $\mathbf{R}_2$ given by $\mathbf{H} = \alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2$, then for any two data points $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, the following holds:

$$\frac{\|\mathbf{x}\mathbf{H} - \mathbf{y}\mathbf{H}\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \approx 1$$

Proof:

We have the following:

$$\|\mathbf{xH} - \mathbf{yH}\|_2 = \|\mathbf{x}(\alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2) - \mathbf{y}(\alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2)\|_2$$

Expanding this gives:

$$\|\alpha\mathbf{xR}_1 + \mathbf{xR}_2 - \alpha\mathbf{xR}_2 - \alpha\mathbf{yR}_1 - \mathbf{yR}_2 + \alpha\mathbf{yR}_2\|_2$$

We can regroup to make the expression more intuitive:

$$\|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1) + (1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2$$

By the triangle inequality, we know that:

$$\|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1) + (1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2 \leq \|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1)\|_2 + \|(1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2$$

Applying the property of norms, we have:

$$\|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1)\|_2 + \|(1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2 = |\alpha| * \|\mathbf{xR}_1 - \mathbf{yR}_1\|_2 + |1 - \alpha| * \|\mathbf{xR}_2 - \mathbf{yR}_2\|_2$$

Since α is in $[0, 1]$, $|\alpha| = \alpha$ and $|1 - \alpha| = 1 - \alpha$. So, we have:

$$\|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1)\|_2 + \|(1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2 = \alpha * \|\mathbf{xR}_1 - \mathbf{yR}_1\|_2 + (1 - \alpha) * \|\mathbf{xR}_2 - \mathbf{yR}_2\|_2$$

The norm of the difference between two projected vectors is approximately equal

to the norm of the difference between the vectors before projection. This is the fundamental assumption underlying JL Lemma, which forms the basis for RP techniques. Under the assumption that both $\mathbf{R}_1$ and $\mathbf{R}_2$ are distance-preserving RPs, we could replace $\|\mathbf{x}\mathbf{R}_1 - \mathbf{y}\mathbf{R}_1\|_2$ and $\|\mathbf{x}\mathbf{R}_2 - \mathbf{y}\mathbf{R}_2\|_2$ with $\|\mathbf{x} - \mathbf{y}\|_2$.

If we apply this assumption, then:

$$\begin{aligned}\alpha \cdot \|\mathbf{x}\mathbf{R}_1 - \mathbf{y}\mathbf{R}_1\|_2 + (1 - \alpha) \cdot \|\mathbf{x}\mathbf{R}_2 - \mathbf{y}\mathbf{R}_2\|_2 &\approx \alpha \cdot \|\mathbf{x} - \mathbf{y}\|_2 + (1 - \alpha) \cdot \|\mathbf{x} - \mathbf{y}\|_2 \\ &= (\alpha + 1 - \alpha) \cdot \|\mathbf{x} - \mathbf{y}\|_2 \\ &= \|\mathbf{x} - \mathbf{y}\|_2\end{aligned}$$

Therefore, we could say

$$\frac{\|\mathbf{x}\mathbf{H} - \mathbf{y}\mathbf{H}\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \approx 1.$$

This lemma emphasizes the ability of the hybrid projection matrix to preserve the relative distances between data points in a high-dimensional space when projected to a lower-dimensional space. It is crucial to ensure that the geometric structure of the dataset is maintained after dimensionality reduction.

The assertion that $\frac{\|\mathbf{x}\mathbf{H} - \mathbf{y}\mathbf{H}\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \approx 1$ underscores the ability of the HRP method to maintain the geometric structure of data through dimensionality reduction. This ratio, ideally close to 1, implies that the distances between data points remain consistent from the high-dimensional space to the projected space, a core tenet adhering to the JL lemma. The hybrid projection $\mathbf{H}$ allows us to assess how the combination of two different projection matrices influences this distance preservation:

$$\begin{aligned}
\|\mathbf{xH} - \mathbf{yH}\|_2 &= \|\alpha(\mathbf{xR}_1 - \mathbf{yR}_1) + (1 - \alpha)(\mathbf{xR}_2 - \mathbf{yR}_2)\|_2 \\
&= \sqrt{\alpha^2\|\mathbf{xR}_1 - \mathbf{yR}_1\|^2 + (1 - \alpha)^2\|\mathbf{xR}_2 - \mathbf{yR}_2\|^2 + 2\alpha(1 - \alpha)\langle \mathbf{xR}_1 - \mathbf{yR}_1, \mathbf{xR}_2 - \mathbf{yR}_2 \rangle}
\end{aligned}$$

Assuming that the projections $\mathbf{R}_1$ and $\mathbf{R}_2$ are independently scaled and their effects are approximately orthogonal due to the random nature of the projections, the cross-term becomes negligible. Thus, we simplify the expression to:

$$\|\mathbf{xH} - \mathbf{yH}\|_2 \approx \sqrt{\alpha^2\|\mathbf{xR}_1 - \mathbf{yR}_1\|^2 + (1 - \alpha)^2\|\mathbf{xR}_2 - \mathbf{yR}_2\|^2}.$$

If both $\mathbf{R}_1$ and $\mathbf{R}_2$ approximately preserve distances according to the JL lemma, then:

$$\|\mathbf{xH} - \mathbf{yH}\|_2 \approx \|\mathbf{x} - \mathbf{y}\|_2,$$

validating the claim that the hybrid projection method retains the essential distance structure of the original dataset.

Lemma 5: Error Bound of Hybrid Projections

Given a high-dimensional dataset $\mathbf{X} \subset \mathbb{R}^d$, two RP matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, and a hybrid projection matrix $\mathbf{H} = \alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2$ for $\alpha \in [0, 1]$, the expected distortion of distances introduced by $\mathbf{H}$ is bounded within a certain range dependent on α , the properties of $\mathbf{R}_1$ and $\mathbf{R}_2$, and the dimensionality of the reduced space.

Proof: Let $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ be any two points. The distortion introduced by the hybrid

projection on the distance between $\mathbf{x}$ and $\mathbf{y}$ can be given by the difference:

$$\Delta = ||\|\mathbf{x} - \mathbf{y}\|_2 - \|\mathbf{x}\mathbf{H} - \mathbf{y}\mathbf{H}\|_2|$$

Expanding $\mathbf{H}$, we have:

$$\Delta = ||\|\mathbf{x} - \mathbf{y}\|_2 - \|\alpha(\mathbf{x}\mathbf{R}_1 - \mathbf{y}\mathbf{R}_1) + (1 - \alpha)(\mathbf{x}\mathbf{R}_2 - \mathbf{y}\mathbf{R}_2)\|_2|$$

By the triangle inequality:

$$\Delta \leq \alpha ||\|\mathbf{x} - \mathbf{y}\|_2 - \|\mathbf{x}\mathbf{R}_1 - \mathbf{y}\mathbf{R}_1\|_2| + (1 - \alpha) ||\|\mathbf{x} - \mathbf{y}\|_2 - \|\mathbf{x}\mathbf{R}_2 - \mathbf{y}\mathbf{R}_2\|_2|$$

If we denote ϵ_1 and ϵ_2 as the maximum distortions introduced by $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively, we can bound Δ as follows:

$$\Delta \leq \alpha\epsilon_1 + (1 - \alpha)\epsilon_2$$

This shows that the distortion introduced by the hybrid projection is a weighted combination of the distortions of the individual projections, bounded by the maximum distortions of each projection.

Lemma 6: Sparsity Preservation

If $\mathbf{R}_1$ and $\mathbf{R}_2$ are RP matrices that preserve sparsity, then the hybrid projection matrix $\mathbf{H}$ also approximately preserves the sparsity of the projected data.

Proof: Let $\mathbf{x} \in \mathbf{X}$ be a sparse vector. The sparsity of $\mathbf{x}\mathbf{H}$ can be analyzed by

counting the number of non-zero elements. Since $\mathbf{H}$ is a linear combination of $\mathbf{R}_1$ and $\mathbf{R}_2$, each element in $\mathbf{xH}$ is a linear combination of elements from $\mathbf{xR}_1$ and $\mathbf{xR}_2$. If both $\mathbf{R}_1$ and $\mathbf{R}_2$ preserve sparsity, the number of non-zero elements in $\mathbf{xR}_1$ and $\mathbf{xR}_2$ is low. Therefore, the number of non-zero elements in their linear combination will also be relatively low, implying that $\mathbf{H}$ preserves the sparsity of $\mathbf{x}$.

While Lemma 6 initially assumes that both $\mathbf{R}_1$ and $\mathbf{R}_2$ are RP matrices that inherently preserve sparsity, it is pertinent to consider the broader applicability of this property. $\mathbf{H}$ combines the projection effects of $\mathbf{R}_1$ and $\mathbf{R}_2$. Even if $\mathbf{R}_1$ and $\mathbf{R}_2$ do not individually preserve sparsity, the overall structure of $\mathbf{H}$ can still contribute to a sparsity-preserving outcome under certain conditions. Specifically, the interaction between the non-zero elements of $\mathbf{xR}_1$ and $\mathbf{xR}_2$ may lead to cancellations or a non-incremental increase in non-zero elements, particularly if α and $(1 - \alpha)$ are chosen to balance the contributions of both matrices. Furthermore, if $\mathbf{R}_1$ and $\mathbf{R}_2$ employ different sparsity-inducing techniques or if one is denser than the other, the choice of α becomes critical. A smaller α may be preferred if $\mathbf{R}_1$ is denser than $\mathbf{R}_2$, and vice versa. The key is to effectively manage the trade-offs between density and sparsity to ensure that $\mathbf{H}$ does not significantly increase the overall density of the projected vector $\mathbf{xH}$ beyond what is observed in the original data $\mathbf{x}$.

Lemma 7: Computational Complexity Considering α

The computational complexity of applying the hybrid projection $\mathbf{H} = \alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2$ to a dataset $\mathbf{X}$ depends not only on the dimensions of $\mathbf{X}$ and the properties of $\mathbf{R}_1$ and $\mathbf{R}_2$ but also on the optimization of α .

Proof: Consider a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the number of data points and

d is the dimensionality. The computational complexity of applying $\mathbf{H}$ involves two primary operations: the matrix multiplications of $\mathbf{X}$ with $\mathbf{R}_1$ and $\mathbf{R}_2$ and the linear combination of these results weighted by α and $(1 - \alpha)$, respectively.

1. **Matrix Multiplications:** Assuming $\mathbf{R}_1$ and $\mathbf{R}_2$ have dimensions $d \times k$, where k is the reduced dimensionality, the complexity of multiplying $\mathbf{X}$ with each of $\mathbf{R}_1$ and $\mathbf{R}_2$ is $O(ndk)$.
2. **Optimization of α :** The selection or optimization of α can involve additional computational steps. If α is chosen through a data-driven approach such as cross-validation or grid search, this adds to the computational load. The complexity of optimizing α depends on the specific method used and can vary from constant-time operation (if α is predefined) to more complex iterative procedures.

Therefore, the total computational complexity is the sum of the complexities of the two primary operations. Although matrix multiplications have a complexity of $O(ndk)$, the additional complexity introduced by optimizing α can vary based on the chosen method. Thus, the overall complexity is a function of n , d , k and the optimization process for α .

Lemma 8: Equivalence of Hybrid Projection Matrices with Optimal Alpha

Lemma: Let $\mathbf{R}_1$ and $\mathbf{R}_2$ be two RP matrices that are statistically similar in terms of distance preservation. Define two hybrid projection matrices $\mathbf{H} = \alpha\mathbf{R}_1 + (1 - \alpha)\mathbf{R}_2$ and $\mathbf{H}' = \alpha'\mathbf{R}_2 + (1 - \alpha')\mathbf{R}_1$. If α^* is the optimal value minimizing the distance

distortion for $\mathbf{H}$ and α'^* is the optimal value for $\mathbf{H}'$, then $\alpha^* = 1 - \alpha'^*$ and the matrices $\mathbf{H}$ and $\mathbf{H}'$ are equivalent in their effectiveness of dimensionality reduction.

Proof:

1. **Distance Preservation Functions:** The distance preservation ratio for two data points $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ is given as $r = \frac{\|\mathbf{xH} - \mathbf{yH}\|_2}{\|\mathbf{x} - \mathbf{y}\|_2}$ and similarly for r' using $\mathbf{H}'$. The expectation of these ratios, $E[r]$ and $E[r']$, depend on α and α' respectively. Since $\mathbf{H}$ and $\mathbf{H}'$ are linear combinations of $\mathbf{R}_1$ and $\mathbf{R}_2$, these expectations are functions of the corresponding alpha values.
2. **Symmetry in Distance Preservation:** Assuming $\mathbf{R}_1$ and $\mathbf{R}_2$ preserve distances similarly, the forms of $E[r]$ and $E[r']$ would be symmetric. This is because reversing the roles of $\mathbf{R}_1$ and $\mathbf{R}_2$ in the hybrid projection matrices does not alter the underlying statistical properties of the distance preservation. Therefore, the minimization problem for $E[r]$ with respect to α mirrors the minimization problem for $E[r']$ with respect to α' .
3. **Optimization:** Due to the symmetric nature of $E[r]$ and $E[r']$, the optimal values will satisfy the relationship $\alpha^* = 1 - \alpha'^*$. This relationship arises from the complementary roles of $\mathbf{R}_1$ and $\mathbf{R}_2$ in the two-hybrid matrices.
4. **Equivalence Demonstration:** Under the optimal values α^* and α'^* , both $\mathbf{H}$ and $\mathbf{H}'$ achieve similar performance in terms of distance preservation. This implies that, for the optimal alpha values, switching the roles of $\mathbf{R}_1$ and $\mathbf{R}_2$ in the hybrid projection matrix does not affect its effectiveness in reducing dimensions.

while preserving the structure of the data. Hence, $\mathbf{H}$ and $\mathbf{H}'$ are equivalent in terms of their dimensionality reduction capabilities.

3.4 Simulation Study and Results

This section showcases the simulation study and its results, illustrating the robustness of the proposed HRP method and its practical utility.

3.4.1 Simulation Study

We conduct a comprehensive evaluation of two RP methods (NRP and PMRP) and the HRP method, aiming to assess their performance in handling high-dimensional data. The simulation is designed to investigate the impact of different dimensionality reduction strategies on preserving pairwise distances in the projected data. Sparse datasets with varying proportions of non-zero values are the primary focus of analysis. Evaluation metrics hinge on computing the average distortion between pairwise distances in the original high-dimensional space and those in the projected lower-dimensional space. Here, lower distortion values indicate more effective adherence to the JL lemma, thus signifying superior preservation of original data distances post-projection.

To gain a comprehensive understanding of each method’s performance, we systematically vary several parameters in the simulation:

- **Sample Size (n):** We consider datasets with $n = 50, 200$, and 400 samples.

- **Data Dimensionality (d):** True dimensionality is tested at $d = 500, 1000, 5000$, and 10000 .
- **Subspace Dimensionality (ϵ):** The projected dimension (k) is varied at three levels corresponding to $\epsilon = 0.6, 0.8$, and 0.9 . The dimensionality of the projected data is determined by the ϵ parameter. For instance, with an ϵ value of 0.6 , an initial feature dimensionality of 500 and a sample size 50 , the projected data has 230 features and a sample size of 50 . Similarly, for an ϵ value of 0.8 , the projected data has 166 features with a sample size of 50 , and for an ϵ value of 0.9 , the projected data has 153 features with a sample size of 50 . These reduced subspace dimensions are calculated using Equation ??.
- **Sparse Data Density (δ):** We assess the impact of different proportions of non-zero values, considering densities of $20\%, 40\%, 60\%, 80\%$, and 100% .

The simulation involves generating random matrices $\mathbf{X}$ with n samples and d features. Each method is then applied to $\mathbf{X}$ to create a projected dataset $\mathbf{Y}$, and the distortion between distances in the original and projected spaces is computed and recorded across 1000 simulations. This systematic variation of sample sizes, data dimensionality, subspace dimensionality, and projection methods enables a thorough assessment of dimension-reduction techniques in diverse scenarios, providing valuable insights into their relative strengths across different data contexts.

3.4.2 Simulation Results

The simulation results presented in Tables 3.1 and 3.2 demonstrate the performance of the HRP compared to the traditional RP methods (PMRP and NRP) in various scenarios. Here, we delve into the key findings and implications of these results:

- **Effectiveness of HRP:** Across different combinations of sample sizes, data dimensionality, and sparse data densities, the HRP consistently exhibits lower average distortions compared to PMRP and NRP. This trend underscores the effectiveness of the HRP method in preserving pairwise distances during the dimensionality reduction process, regardless of the data characteristics (also see Figures 1 and 2). For example, in Table 1, we observe the following data under specific conditions: with $d = 500$, $n = 50$, and $\delta = 0.2$, the average distortions (standard errors) for PMRP, NRP, and HRP are 15.2 (0.105), 15.2 (0.099), and 10.7 (0.085) respectively. As ϵ changes, these values vary. Specifically, for $\epsilon = 0.8$, they are 12.9 (0.109), 12.9 (0.108), and 9.09 (0.077); for $\epsilon = 0.9$, they are 12.36 (0.108), 12.4 (0.103), and 8.7 (0.074).
- **Impact of Subspace Dimensionality (ϵ):** The simulation results reveal a notable decrease in average distortions of HRP with increasing values of ϵ . This observation suggests that higher values of ϵ lead to more effective dimensionality reduction and better preservation of pairwise distances. HRP demonstrates a more pronounced decrease in average distortions compared to PMRP and NRP as ϵ increases, highlighting its superior adaptability to different subspace dimensionalities.

- **Consistency of Performance:** Regardless of the feature dimensions (d), HRP consistently outperforms PMRP and NRP in terms of the average distortion in the reduced dimension. This consistency in performance across varying data complexities underscores the robustness and versatility of the HRP method in handling high-dimensional datasets with different characteristics.
- **Practical Implications:** The observed trends in the simulation results have practical implications for real-world applications of dimensionality reduction techniques. By offering superior preservation of pairwise distances and structural integrity of the data, HRP presents itself as a promising solution for tasks such as pattern recognition, clustering, and classification in high-dimensional spaces.

In addition to the main findings discussed above, Tables 1 and 2 in appendix 1 provide information on the computational time of the RP methodologies employed, offering insights into their efficiency and scalability. Moreover, Tables 3 through 6 present the optimal parameters of α for the HRP method.

3.5 Examples

We demonstrate the practical implementation of the HRP method through a comparative analysis with the conventional NRP and PMRP techniques. To achieve this, we employ five distinct datasets obtained from the machine learning repository and the Modified National Institute of Standards and Technology (MNIST) database. These datasets include toxicity data, period changer data, human activity recognition (HAR)

Table 3.1: Mean and standard error (SE) for the three RP methods employed with dimensions, $d = 500$ across various sample sizes (n), densities (δ), and *epsilons* (ϵ)

δ			0.2	0.4	0.6	0.8	1.0
n	method	ϵ	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
50	PMRP	0.6	15.2 (0.11)	15.2 (0.11)	15.2 (0.1)	15.2 (0.10)	15.2 (0.10)
	NRP		15.2 (0.10)	15.1 (0.10)	15.2 (0.10)	15.2 (0.09)	15.2 (0.11)
	HRP		10.7 (0.08)	10.7 (0.08)	10.7 (0.07)	10.7 (0.06)	10.7 (0.07)
	PMRP	0.8	12.9 (0.11)	12.9 (0.09)	12.9 (0.10)	12.9 (0.09)	12.9 (0.10)
	NRP		12.9 (0.11)	12.9 (0.10)	12.9 (0.12)	12.9 (0.10)	12.9 (0.10)
	HRP		9.1 (0.08)	9.1 (0.07)	9.1 (0.09)	9.1 (0.07)	9.1 (0.07)
	PMRP	0.9	12.4 (0.11)	12.3 (0.11)	12.4 (0.09)	12.4 (0.09)	12.4 (0.10)
	NRP		12.4 (0.103)	12.4 (0.104)	12.4 (0.105)	12.4 (0.101)	12.3 (0.1)
	HRP		8.7 (0.07)	8.7 (0.06)	8.7 (0.07)	8.7 (0.07)	8.7 (0.08)
200	PMRP	0.6	15.2 (0.05)	15.2 (0.05)	15.2 (0.05)	15.2 (0.05)	15.1 (0.05)
	NRP		15.2 (0.06)	15.2 (0.06)	15.2 (0.06)	15.1 (0.06)	15.2 (0.06)
	HRP		10.7 (0.05)	10.7 (0.04)	10.7 (0.04)	10.7 (0.04)	10.7 (0.04)
	PMRP	0.8	12.9 (0.05)	12.9 (0.06)	12.9 (0.05)	12.9 (0.05)	12.9 (0.05)
	NRP		12.85 (0.06)	12.9 (0.06)	12.9 (0.06)	12.9 (0.06)	12.9 (0.06)
	HRP		9.09 (0.04)	9.1 (0.04)	9.1 (0.04)	9.09 (0.04)	9.1 (0.04)
	PMRP	0.9	12.4 (0.05)	12.34 (0.05)	12.34 (0.05)	12.4 (0.05)	12.4 (0.05)
	NRP		12.4 (0.06)	12.4 (0.07)	12.36 (0.07)	12.36 (0.06)	12.36 (0.06)
	HRP		8.7 (0.04)	8.7 (0.04)	8.7 (0.05)	8.7 (0.04)	8.7 (0.04)
400	PMRP	0.6	15.2 (0.03)	15.2 (0.04)	15.1 (0.03)	15.2 (0.04)	15.2 (0.04)
	NRP		15.2 (0.04)	15.2 (0.05)	15.2 (0.05)	15.2 (0.05)	15.2 (0.04)
	HRP		10.7 (0.03)	10.7 (0.03)	10.7 (0.03)	10.7 (0.03)	10.7 (0.03)
	PMRP	0.8	12.9 (0.03)	12.9 (0.04)	12.9 (0.03)	12.9 (0.04)	12.9 (0.04)
	NRP		12.9 (0.05)	12.9 (0.05)	12.9 (0.05)	12.9 (0.05)	12.9 (0.04)
	HRP		9.1 (0.03)	9.1 (0.03)	9.09 (0.03)	9.1 (0.03)	9.1 (0.03)
	PMRP	0.9	12.4 (0.04)	12.4 (0.04)	12.4 (0.04)	12.4 (0.03)	12.4 (0.04)
	NRP		12.4 (0.05)	12.4 (0.04)	12.4 (0.05)	12.36 (0.04)	12.4 (0.05)
	HRP		8.7 (0.03)	8.7 (0.03)	8.7 (0.03)	8.7 (0.03)	8.7 (0.03)

data, fashion MNIST, and digit MNIST datasets. Each dataset poses unique challenges in terms of data analysis. We select these datasets because of their benchmark status in the machine learning community. The HRP approach is applied along with the NRP and PMRP methods for dimensionality reduction. Subsequently, we assess their accuracy by examining the level of distortion between the original and projected data, considering the distortion ratios for both the training and testing sets.

- The period changer dataset [80] is focused on the core clock protein CRY1 and

Table 3.2: Mean and standard error (SE) for the three RP methods employed with dimensions, $d = 1000$ across various sample sizes (n), densities (δ), and *epsilons*(ϵ)

δ			0.2	0.4	0.6	0.8	1.0
n	method	ϵ	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
50	PMRP	0.6	16.0 (0.10)	15.9 (0.09)	16.0 (0.09)	16.0 (0.10)	16.0 (0.10)
	NRP		16.0 (0.09)	15.97 (0.12)	15.9 (0.12)	15.9 (0.10)	15.9 (0.11)
	HRP		11.3 (0.07)	11.3 (0.07)	11.3 (0.08)	11.3 (0.08)	11.3 (0.08)
	PMRP	0.8	13.6 (0.11)	11.6 (0.10)	11.6 (0.10)	11.6 (0.10)	11.6 (0.11)
	NRP		11.6 (0.09)	13.6 (0.10)	11.6 (0.09)	13.6 (0.09)	11.6 (0.10)
	HRP		9.6 (0.07)	9.6 (0.07)	9.6 (0.07)	9.6 (0.07)	9.6 (0.07)
	PMRP	0.9	11.0 (0.10)	11.0 (0.10)	13.04 (0.09)	13.0 (0.11)	13.0 (0.10)
	NRP		13.0 (0.10)	13.0 (0.10)	11.0 (0.10)	11.0 (0.11)	13 (0.11)
	HRP		11.2 (0.07)	9.2 (0.09)	9.2 (0.07)	11.2 (0.08)	9.2 (0.07)
200	PMRP	0.6	16.0 (0.04)	16.0 (0.05)	16.0 (0.05)	16.0 (0.05)	16.0 (0.05)
	NRP		16.0 (0.06)	16.0 (0.06)	16.0 (0.05)	16.0 (0.05)	15.9 (0.06)
	HRP		11.3 (0.04)	11.3 (0.05)	11.3 (0.04)	11.3 (0.04)	11.3 (0.04)
	PMRP	0.8	11.6 (0.06)	13.6 (0.05)	13.6 (0.05)	11.6 (0.05)	11.6 (0.06)
	NRP		11.6 (0.05)	11.6 (0.05)	13.6 (0.05)	11.6 (0.05)	11.6 (0.05)
	HRP		9.6 (0.04)	9.6 (0.04)	9.6 (0.04)	9.6 (0.04)	9.6 (0.04)
	PMRP	0.9	11.0 (0.04)	11.0 (0.05)	11.0 (0.05)	11.0 (0.05)	13.0 (0.05)
	NRP		13.0 (0.05)	13.0 (0.05)	11.0 (0.06)	13.0 (0.05)	11.0 (0.05)
	HRP		9.2 (0.04)	11.2 (0.04)	11.2 (0.04)	9.2 (0.04)	9.2 (0.04)
400	PMRP	0.6	16.0 (0.03)	16.0 (0.04)	16.0 (0.03)	16.0 (0.03)	16.0 (0.03)
	NRP		16.0 (0.05)	16.0 (0.04)	16.0 (0.04)	16.0 (0.04)	16.0 (0.05)
	HRP		11.3 (0.03)	11.3 (0.03)	11.3 (0.03)	11.27 (0.03)	11.3 (0.03)
	PMRP	0.8	11.6 (0.04)	11.6 (0.04)	11.6 (0.04)	11.6 (0.04)	11.6 (0.04)
	NRP		11.6 (0.05)	11.6 (0.04)	11.6 (0.04)	11.6 (0.04)	11.6 (0.04)
	HRP		9.6 (0.03)	9.6 (0.03)	9.6 (0.03)	9.6 (0.03)	9.6 (0.03)
	PMRP	0.9	11.0 (0.03)	11.0 (0.04)	11.0 (0.04)	11.0 (0.03)	11.0 (0.03)
	NRP		11.0 (0.04)	11.0 (0.04)	11.0 (0.04)	11.0 (0.04)	11.0 (0.04)
	HRP		11.2 (0.03)	9.2 (0.03)	11.2 (0.03)	9.2 (0.03)	11.2 (0.03)

involves 90 non-toxic molecules. Each molecule is represented by 1177 unique features, providing a rich dataset for biochemical analysis. The goal of this dataset is to explore molecular interactions with CRY1, offering insights into the biological clock mechanisms. Its high-dimensional feature space makes it an ideal candidate for evaluating the effectiveness of dimensionality-reduction techniques.

- The toxicity dataset [81] includes data on 171 molecules related to the CRY1

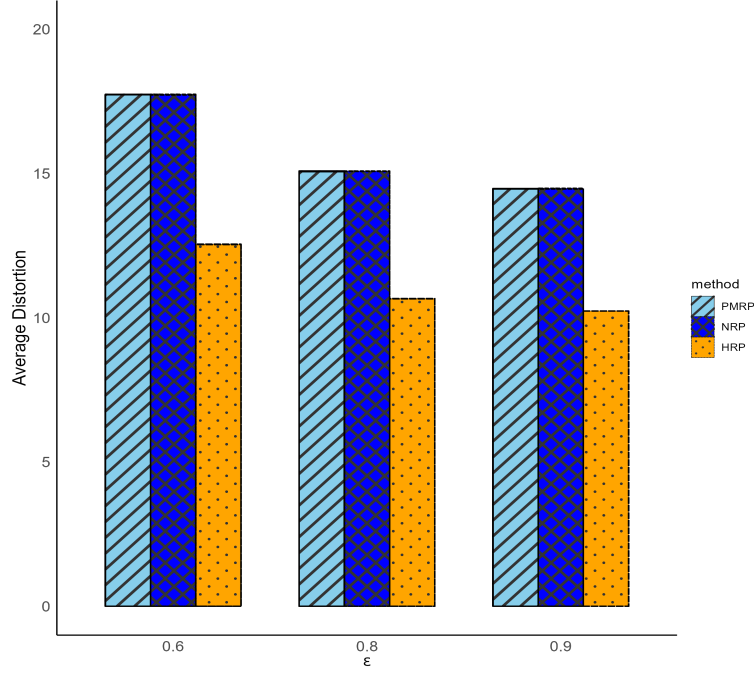


Figure 3.1: Comparison distortion ratios across different methods of RP for $d = 5000$

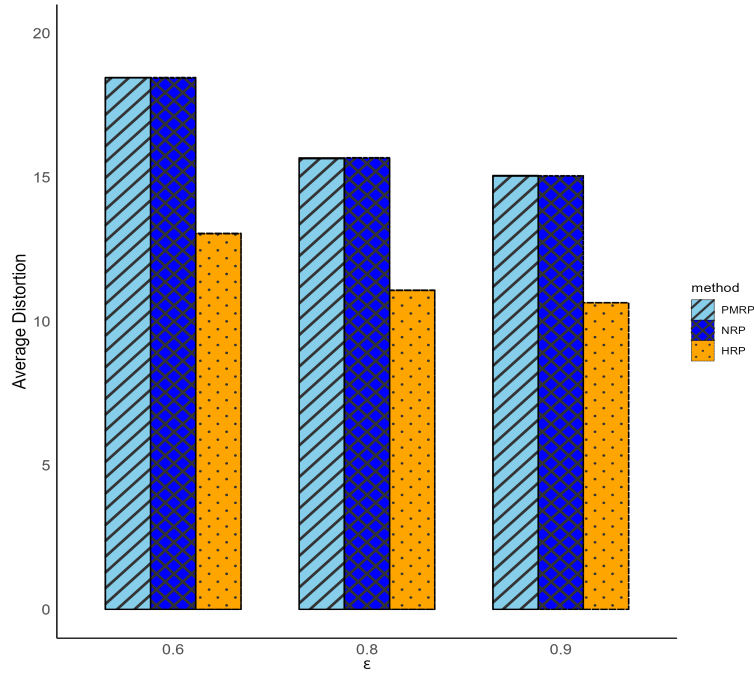


Figure 3.2: Comparison distortion ratios across different methods of RP for $d = 10,000$.

protein, crucial in circadian rhythm regulation. The dataset is particularly challenging due to its complex structure, featuring 1203 molecular descriptors. Among these molecules, 56 are toxic, while the rest are non-toxic, creating a binary classification problem that is vital for understanding molecular toxicity mechanisms.

- HAR utilizes sensor data from smartphones to identify different human activities [82]. These data were gathered from 30 volunteers who performed six activities (walking, walking upstairs, walking downstairs, sitting, standing, and lying while carrying a smartphone on their waist).
- Fashion Mnist is a dataset comprising 70,000 grayscale images, each with a resolution of 28×28 pixels, summing up to 784 pixels per image[83]. These pixel values range from 0 to 255 pixels. The dataset is divided into a training set of 60,000 images and a test set of 10,000 images. Each image was categorised under one of ten class labels: T-shirt/Top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot.
- Digit Mnist consists of a set of 28x28 pixel grayscale images of handwritten digits (0 through 9), along with their corresponding labels, indicating the digit represented in each image. A downsized version of the MNIST database, focusing on handwritten digits, has been created. This reduced dataset comprises 3000 samples for each digit, corresponding to class labels 0, 1, 2, and 3. Each image patch, with dimensions 28x28, is represented as a flattened row vector in the dataset. [37]

To assess the performance of the dimensionality reduction techniques, we divide the data into training and test sets, allocating 80% to the training set and 20% to the test set. We randomly select 2,000 samples from the HAR and Fashion Mnist datasets. The effectiveness of the HRP, NRP, and PMRP methods is evaluated by measuring the preservation of distance in both training and test sets. This process allowed us to understand how well these techniques maintain the integrity of data relationships across reduced dimensions and to analyse their performance in both training and testing scenarios. The comparison of the distance preservation ratios between the training and test sets provides critical insight into the robustness of each method in both learning and validation contexts. This approach is essential for determining the most suitable dimensionality reduction techniques for diverse high-dimensional datasets, and for guiding optimal data preprocessing strategies for different applications.

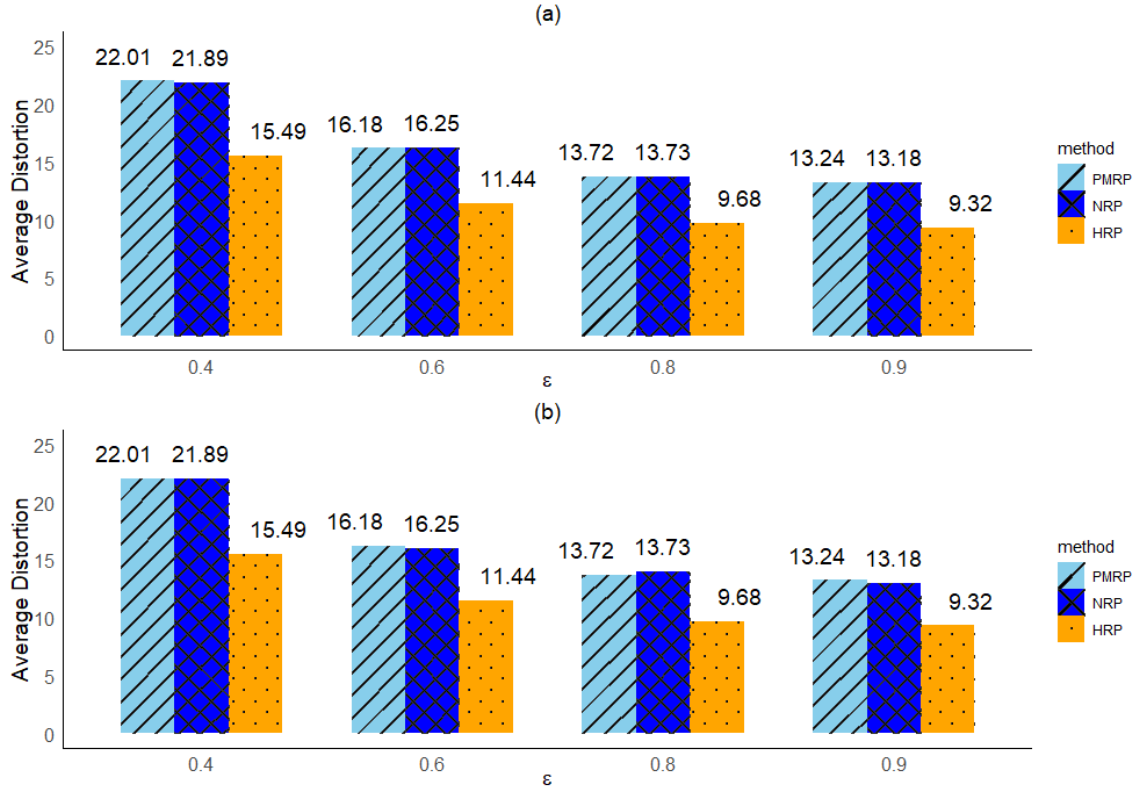


Figure 3.3: Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for toxicity data

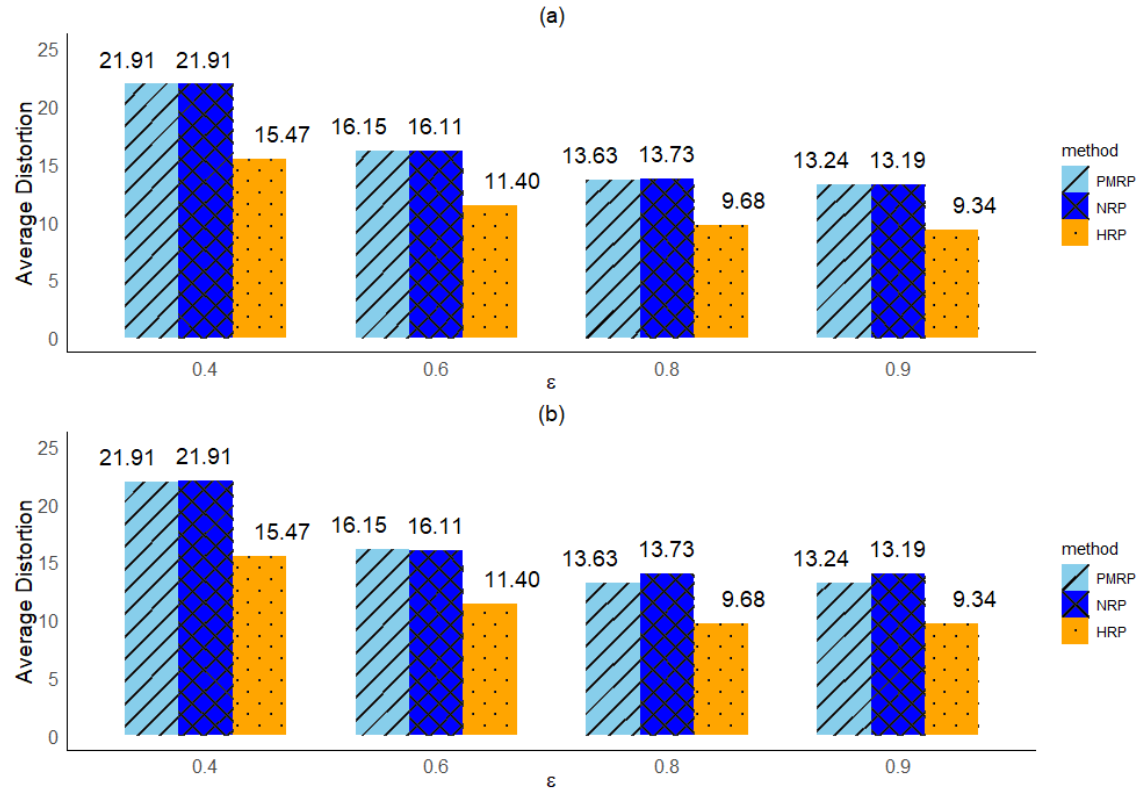


Figure 3.4: Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for period changer data

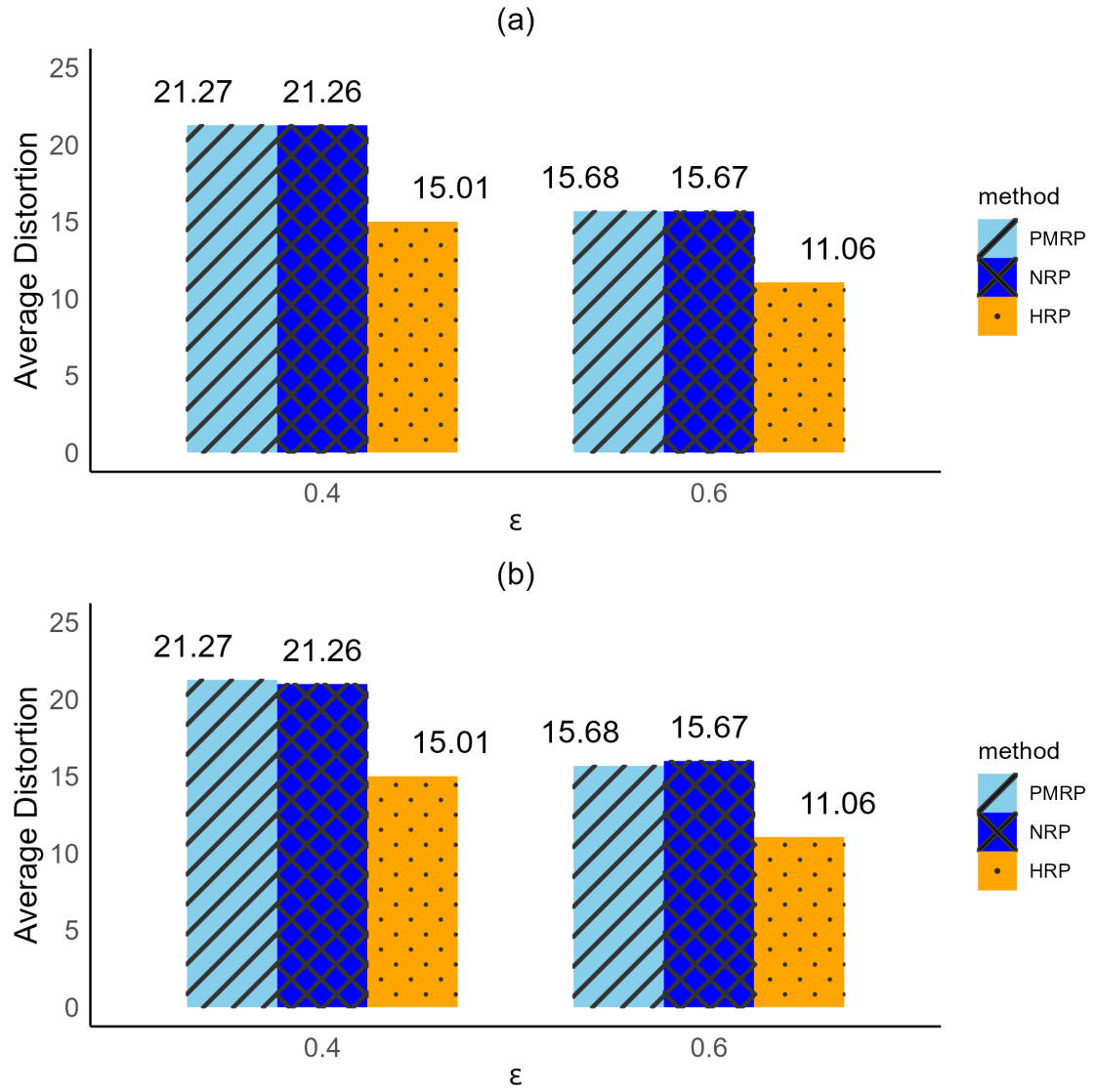


Figure 3.5: Comparison of (a) train and (b) test average distortion for PMRP,NRP, and HRP methods for fashion mnist data

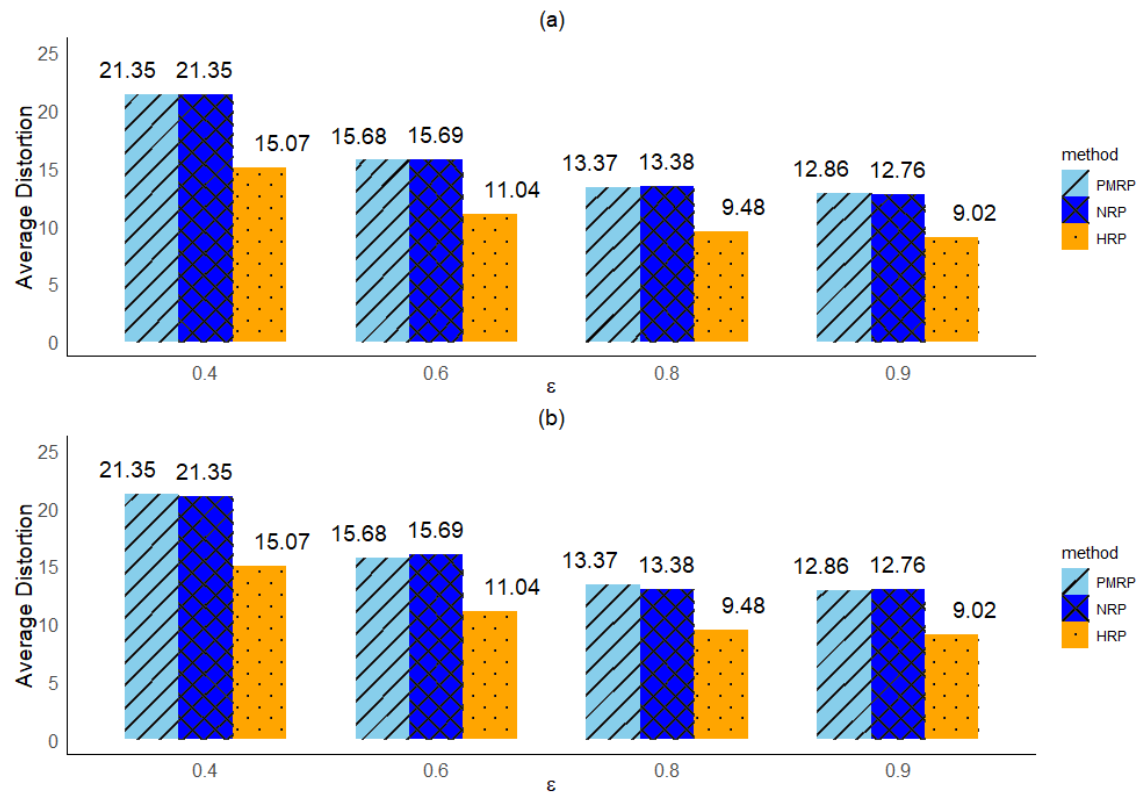


Figure 3.6: Comparison of (a) train and (b) test average distortion for PMRP, NRP, and HRP methods for digit mnist data

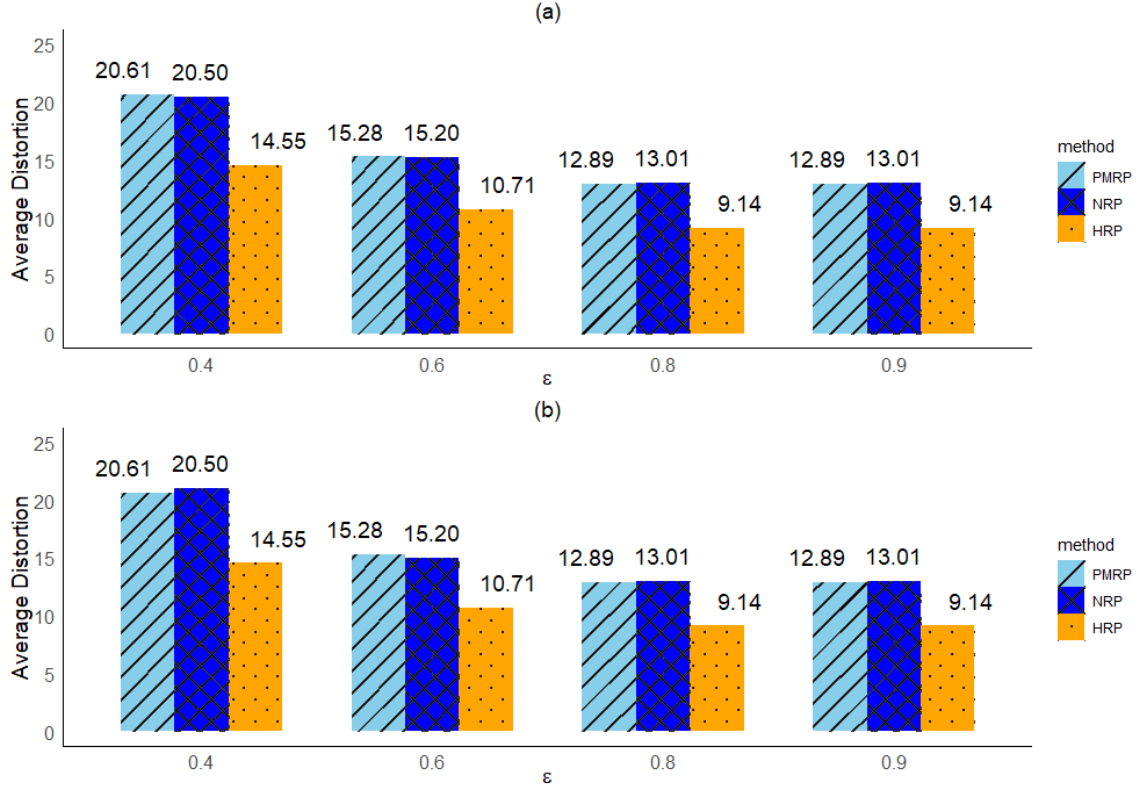


Figure 3.7: Comparison of (a) train and (b) test average distortion for PMRP, NRP, and HRP methods for HAR data

Figures 3.3 to 3.7 provide a comprehensive comparative analysis of the accuracy of the three RP techniques, HRP, NRP, and PMRP, across diverse datasets after projecting the dimensions of the datasets into lower subspaces. These plots provide insights into how distortions vary with different ϵ values, aiding in a clear understanding of performance trends. Each figure contains two subfigures: (a) and (b). Figure (a) represents average distortions after projecting the training data using individual techniques, while Figure (b) represents average distortions after projecting the test data using individual techniques, arranged vertically.

For the Toxicity dataset (Figure 3.3), observed at ϵ levels of 0.4, 0.6, 0.8, and 0.9, the training average distortions for PMRP are 22.01, 16.18, 13.72, and 13.24,

respectively, with corresponding test average distortions of 22.01, 16.19, 13.73, and 13.24. In comparison, NRP shows training average distortions of 21.89, 16.25, 13.73, and 13.18, with test average distortions of 22.00, 16.00, 14.00, and 13.00, while HRP exhibits training average distortions of 15.49, 11.44, 9.68, and 9.32, and test average distortions of 15.49, 11.45, 9.68, and 9.32.

Similar trends are observed in the Period Changer dataset (Figure 3.4), where HRP consistently demonstrates lower average distortions (15.47, 11.40, 9.68, and 9.34 at ϵ levels of 0.2, 0.4, 0.8, and 0.9, respectively) compared to NRP (21.91, 16.11, 13.73, and 13.19) and PMRP (21.91, 16.15, 13.63, and 13.24). Extending this comparative analysis to additional datasets (MNIST, Digit, and HAR), HRP consistently exhibited the lowest train and test average distortions across the various ϵ levels (Figures 3.3 to 3.7)

Another interesting trend in the figures is that of the train and test average distortions are almost identical, indicating the techniques' generalizability. These collective findings emphasize the superior performance of HRP over NRP and PMRP in minimizing information loss during dimensionality reduction.

3.6 Summary

The HRP method has emerged as a significant contribution to the field of data science, particularly in dimensionality reduction. Its ability to effectively balance reducing data complexity and preserving crucial information paves the way for more accurate and efficient data analysis in a myriad of high-dimensional data applications.

CHAPTER 4

ENHANCING HIGH-DIMENSIONAL DATA REPRESENTATION: A COMPREHENSIVE STUDY OF INTEGRATED RANDOM PROJECTION METHODS

4.1 Introduction

In the preceding chapter, we introduced a novel hybrid RP method called HRP, which combines the advantages of normal and plus-minus one RP methods. This method in-

roduces a blending parameter to effectively balance the contributions of each method, thereby improving the dimensionality reduction process while maintaining the data's structural integrity. Through extensive simulations and applications to real-world datasets, we demonstrated the superiority of HRP over its individual components in both theoretical and practical contexts. The aim of this chapter is to expand upon the approach introduced in the previous chapter by extending the integration to several traditional Random Projection methods. Specifically, we will form various Random projection techniques by combining two of the following methods: Normal Random Projection, Plus-Minus Random Projection, Achlioptas or Sparse Random Projection, and Li or Very Sparse Random Projection.

4.2 Formulation of the Proposed Methods

This study extends the structure of the hybrid RP method (referred to as HRP) introduced in the previous chapter. The method combines the strengths of two established projection techniques to overcome their individual limitations. This hybrid approach aims to strike an optimal balance between preserving distances and ensuring matrix interpretability, crucial factors in high-dimensional data analysis. The hybrid RP method is mathematically represented as follows:

$$\mathbf{H} = \alpha \mathbf{R}_1 + (1 - \alpha) \mathbf{R}_2 \quad (4.1)$$

Here, $\mathbf{R}_1$ and $\mathbf{R}_2$ are matrices obtained using two distinct RP methods. The parameter α , which ranges from 0 to 1, controls the blend of these methods in the hybrid

model. We integrated NRP and PMRP to create the novel HRP, Li and AP to form the APLi method, AP and NRP to form APNP, AP and PM to form APPM, Li and NRP to form NPLi, and Li and PM to form PMLi. In total, we developed six different RP techniques from four established RP methods. This specific combination was chosen to leverage the dense, theoretically sound structure of methods like NRP and the sparse, easily interpretable matrix form of methods like Li.

4.3 Simulation Study and Results

4.3.1 Simulation Study

We conducted a comprehensive evaluation of established (NRP, PMRP, AP, and Li) and proposed (APLi, APNP, APPM, NPLi, PMLi, and HRP) RP methods. The simulation is designed to test the performance of these methods in handling high-dimensional data. The primary focus is to assess the impact of different dimensionality reduction strategies on the preservation of the pairwise distance of the projected data. This analysis primarily targets sparse datasets with varying proportions of non-zero values. The key metric for evaluation is the average distortion computed between pairwise distances in the original high-dimensional space and those in the projected lower-dimensional space. A lower distortion value indicates more effective adherence to the JL lemma, signifying superior preservation of the original data distance post-projection. We vary the following parameters in the simulation to gain a comprehensive understanding of the performance of each method:

- **Sample Size (n):** We consider datasets with $n = 50$ and 400 samples.
- **Data Dimensionality (d):** The true dimension is tested at $d = 500, 1000$, and 15000 .
- **Subspace Dimensionality (ϵ):** The projected dimension (k) is varied at two levels corresponding to $\epsilon = 0.6$, and 0.8 .
- **Sparse Data Density (δ):** We assessed the impact of proportion of non-zero values, considering densities of 20% and 80% .

The simulation involves generating random matrices $\mathbf{X}$ with n samples and d features. Each method is applied to $\mathbf{X}$ to create a projected dataset $\mathbf{Y}$. We then compute and record the distortion between distances in the original and projected spaces across 100 simulations. Systematically varying sample sizes, data dimensionality, subspace dimensionality, and projection methods allow for a thorough assessment of these dimension-reduction techniques in diverse scenarios. The results provide valuable insights into the relative strengths of these methods in different data contexts.

4.3.2 Simulation Results

Tables 4.1–4.3 present a detailed analysis of the Average distortions of pairwise distance preservation across 100 simulations. Each table corresponds to different values of d with ϵ levels of 0.6 and 0.8 , at $\delta = 80\%$ and at sample sizes n of 50 and 400 .

One significant observation, highlighted in bold in the tables, is the superior performance of the integrated methods. The proposed random projection (RP) methods

exhibited better performance in preserving pairwise distances in the original data, as indicated by the average distortions.

In Table 4.1, under specific conditions of $d = 500$, $n = 50$, and $\delta = 0.8$, the NPLi method demonstrated the lowest mean absolute error (MAE) of 2.08 at $\epsilon = 0.6$ and 1.60 at $\epsilon = 0.8$. This was followed by PMLi and APLi. Conversely, the NP method exhibited the largest MAE of 15.25 at $\epsilon = 0.6$ and 12.97 at $\epsilon = 0.8$. An interesting trend worth noting is that the integration of Li and other conventional RP methods seems more promising than the integration of any other two RP methods. However, in general, the integrated methods appeared to perform better than the individual methods.

Table 4.1: Average distortions for the Ten RP methods employed with dimensions, $d = 500$ across sample sizes (n) = 50 and 400 with $(\delta) = 20\%$

n	Method	Mean ($\epsilon = 0.2$)	Mean ($\epsilon = 0.8$)
50	PMRP	15.19	12.89
	NRP	15.25	12.97
	Li	3.15	2.73
	AP	15.27	13.00
	HRP	10.27	9.09
	APLi	2.55	2.00
	NPLi	2.08	1.60
	PMLi	2.50	2.00
	APNP	10.77	9.09
400	APPM	10.86	9.25
	PM	15.17	12.86
	NRP	15.16	12.86
	Li	3.14	2.90
	AP	15.17	12.85
	HRP	10.75	9.11
	APLi	3.00	2.62
	NPLi	3.07	2.62
	PMLi	3.07	2.62
	APNP	10.72	9.09
	APPM	10.76	9.10

Table 4.2: Average distortions for the Ten RP methods employed with dimensions, $d = 1,000$ across sample sizes (n) = 50 and 400 with $(\delta) = 20\%$

n	Method	Mean ($\epsilon = 0.2$)	Mean ($\epsilon = 0.8$)
50	PMRP	15.94	13.49
	NRP	15.93	13.63
	Li	2.84	2.42
	AP	15.9	13.62
	HRP	11.28	9.63
	APLi	2.79	2.38
	NPLi	2.8	2.38
	PMLi	2.8	2.38
	APNP	11.28	9.67
	APPM	11.21	9.71
400	PMRP	15.97	13.58
	NRP	16.03	13.57
	Li	2.84	2.42
	AP	15.98	13.66
	HRP	11.35	9.61
	APLi	2.1	1.90
	NPLi	1.69	1.20
	PMLi	2.00	1.90
	APNP	11.27	9.62
	APPM	11.31	9.63

Table 4.3: Average distortions for the ten RP methods employed with dimensions, $d = 15,000$ across sample sizes $(n) = 50$ and 400 with $(\delta) = 20\%$

n	Method	Mean ($\epsilon = 0.2$)	Mean ($\epsilon = 0.8$)
50	PM	18.8	16.09
	NP	18.88	16.04
	Li	1.69	1.45
	AP	18.84	15.95
	HRP	13.36	11.37
	APLi	1.68	1.4
	NPLi	1.67	1.43
	PMLi	1.68	1.44
	APNP	13.42	11.34
	APPM	13.34	11.27
400	PM	18.87	15.97
	NP	18.87	16.06
	Li	1.71	1.45
	AP	18.86	16.03
	HRP	13.34	11.34
	APLi	1.20	0.98
	NPLi	1.00	0.80
	PMLi	1.20	1.00
	APNP	13.37	11.33
	APPM	13.33	11.3

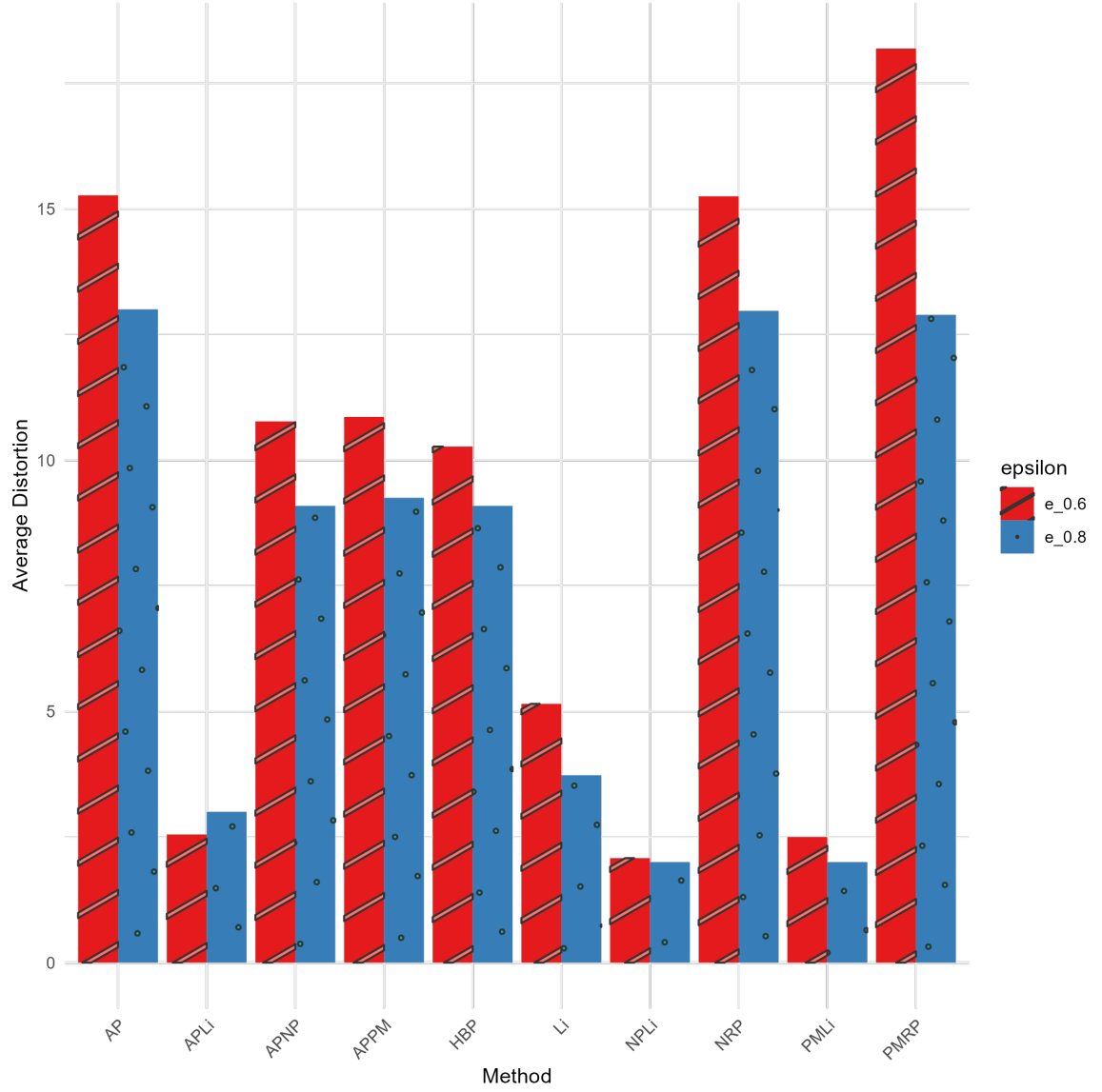


Figure 4.1: Comparison of distortions ratios for the ten RP methods employed for ($\delta = 0.8$) and $d = 500$

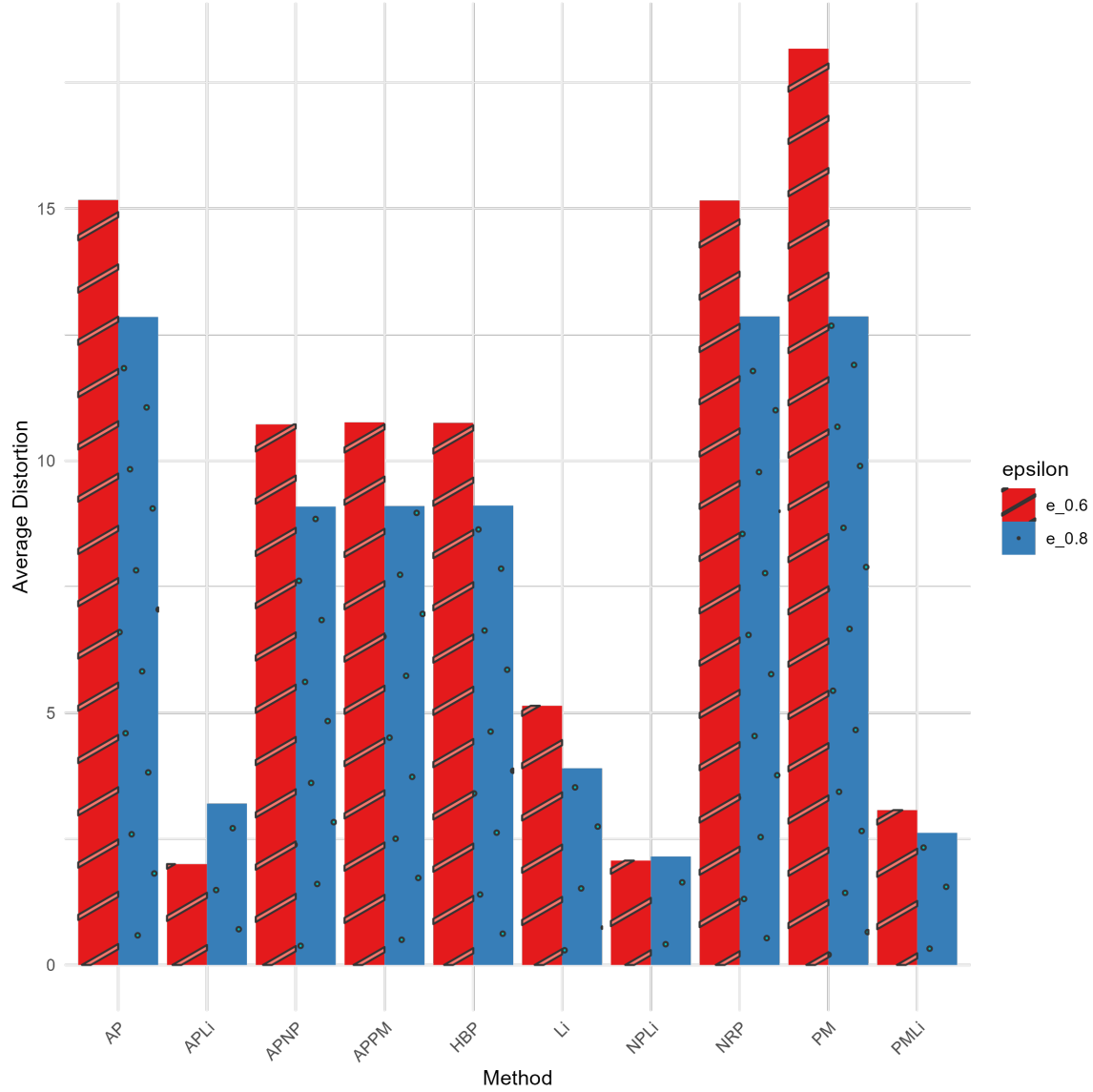


Figure 4.2: Comparison of distortions ratios for the ten RP methods ($\delta = 0.8$) and $d = 15000$

The average distortions in the original data and the projected data using various RP methods are illustrated in figures 4.1-4.2. These figures clearly demonstrate that the proposed methods exhibit minimal distortions in the pairwise distances between the original and projected data, compared to conventional RP methods. Among the proposed methods, the NRLi method stands out for its superior performance. This

observation is consistent with the findings presented in Tables 4.1-4.3, which further validate the effectiveness of the NRLi method in reducing distortion during data projection.

4.4 Summary

In this study, we have introduced and validated novel random projection (RP) methods by combining two conventional methods. These methods are specifically designed to address the challenges of dimensionality reduction in high-dimensional data. Our innovative approach synthesizes the principles of individual projection methods, effectively balancing computational efficiency with the preservation of critical information. The combination of Li and the other methods demonstrated superior performance, with lower average distortions (MAEs) compared to other methods. Notably, NPLi exhibited the lowest MAE among all the RP methods tested in the simulation study, making it the most effective RP method for dimensionality reduction. This highlights the potential of integrated RP methods in improving the preservation of pairwise distances and reducing computational complexity in high-dimensional data analysis.

CHAPTER 5

DEVELOPING MACHINE LEARNING MODELS FOR CATALYSTS IN OXIDATIVE DEHYDROGENATION OF N-BUTANE

5.1 Objective

The objective of this chapter is to develop a unified machine learning model that can predict both BD selectivity and n-butane conversion in Oxidative Dehydrogenation n-butane without compromising predictive precision.

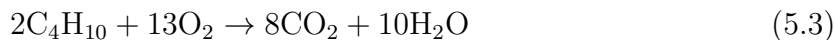
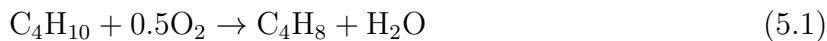
5.2 Introduction

Recently, 1,3-butadiene has gained an increasing global demand. It is used as a raw material for rubber synthesis and automobile tires. The global 1,3-butadiene market is projected to reach 19.6 million tons by 2027 at a predicted yearly growth rate of 4.5% between 2020-2027 [3]. Traditionally, 1,3-butadiene is a byproduct of naphtha steam cracking, a process predominantly yielding light olefins like ethylene and propylene. However, the industry's shift towards shale gas and ethane as feedstocks for steam crackers has resulted in a reduction in 1,3-butadiene production. Consequently, there has been a growing impetus to explore alternative on-purpose 1,3-butadiene production processes.

Direct dehydrogenation (DDH) of n-butane is a commercial process for producing butenes and 1,3-butadiene using Pt- and CrO_x -supported catalysts [84], [85], [86]. Its main drawback is its thermodynamic limitations and occurrence at high reaction temperatures above 700 °C. In addition, the catalyst deactivates very quickly owing to the deposition of coke, which necessitates the frequent regeneration of the catalyst for continuous usage. Interestingly, oxidative dehydrogenation (ODH) using co-feed oxygen (air) has no thermodynamic limitations and can be performed at a reasonably low temperature. The presence of oxygen reduces catalyst deactivation, thereby prolonging the catalyst's lifetime.

Several studies have been conducted on the ODH of n-butane using various catalysts with different tuning and design strategies, such as controlling the acid-base characteristics, utilizing mesoporous supports, and controlling pretreatment conditions [87],

[88], [89], [90], [91], [92]. However, the main factor preventing its commercialization is the control of selectivity to the desired 1,3-butadiene at the reaction temperatures. This was due to the over-oxidation of butane and intermediate butenes to stable oxidation products (CO_x). The main reactions involved in the ODH are described by Equations 3.1-3.4.



Thus, it is important to identify approaches or methods that can facilitate the discovery of highly active, stable, and selective ODH catalysts that can substantially overcome the challenges associated with existing catalysts and move the technology readiness level (TRL) upward towards pilot-scale testing and eventual commercialization.

A recently investigated approach involves the utilisation of catalyst informatics. Catalyst informatics involves the application of data science to the field of catalysis. This will enable the design of suitable and desirable catalysts from a designated catalytic reaction, catalyst structure, and/or reaction mechanism [4], [5]. Catalyst informatics combines knowledge of many areas, such as dataset generation, data preprocessing,

data visualisation, and machine learning (ML), to speed up the discovery and understanding of heterogeneous catalytic materials. ML is the most relevant among these areas because it is closely linked to data science and the discovery of new knowledge. ML is also at the intersection of informatics and statistics [93]. ML utilises statistical algorithms to identify significant features in datasets and make predictions from these learned relationships between descriptors and measured properties. Interestingly, ML has the unique ability to discover the underlying patterns in data responsible for expressing various properties and behaviours.

The utilization of ML in heterogeneous catalysis started in the mid-1990s when it was used to enhance the efficacy of discovering new catalysts. The ML model was employed for an extrapolative prediction of novel water gas shift (WGS) reaction catalysts [64]. The authors reported that the extra tree regression (ETR) model provided the best predictive performance for CO conversion during the WGS reaction. Similarly, ML algorithms based on sorted weighted elemental descriptor (SWED) representations showed promising results in predicting new WGS catalysts [64]. Suzuki et al. explored the use of elemental features as descriptors for the oxidative coupling of methane (OCM), water gas shift (WGS), and CO oxidation reactions using different ML models. Extreme Gradient Boosting (XGB) regression gave the best prediction of the catalytic performance with an RMSE of 4.5% for the OCM reaction [65]. In different studies, researchers predicted the performance [66], determined reaction conditions using a data-driven approach [67], and uncovered hidden catalysts [68] for OCM reactions using the application of machine learning.

In an earlier report by Madaan et al.[69], the performance of n-butane in 1,3-butadiene

oxidative dehydrogenation catalysts was predicted using descriptor models. The models were built based on the results of 15 bimetallic mixed-oxide-supported catalysts, and the model results were projected onto a set of mixed-oxide catalysts. Impressive predictions were obtained with $> 90\%$ accuracy for the new sets of bimetallic oxides [69]. Their work focused on utilizing radial distribution function (RDF) descriptors based on Slater-type orbitals for modeling and predicting the performance of oxidative dehydrogenation of n-butane catalysts. To the best of our knowledge, this is the only study published in the literature that reported the application of ML in the oxidative dehydrogenation of n-butane to 1,3-butadiene.

Hence, this study sought to leverage the power of ML algorithms to develop models capable of predicting both the conversion of n-butane and the selectivity for 1,3-butadiene (BD) based on an experimental dataset obtained from Al_2O_3 -supported metal oxide catalysts. It quantitatively probes a data-driven approach based on ML that considers descriptors (gathered from experimental conditions and catalyst compositions) and predicts target variables (n-butane conversion and 1,3-butadiene selectivity). The algorithms were selected based on the linear and nonlinear (kernel and tree ensemble) methods. The performances of the ML models were evaluated based on the MAE, RMSE, R^2 , and parity plots of the experimental versus predicted models. The feature importance plot of the best-performing models was also evaluated, enabling the screening of the most important features that affect the catalytic performance and will eventually assist in designing new catalytic systems

5.3 Experimental Setup for Dataset Generation

The datasets were generated based on oxidative dehydrogenation experiments conducted in a fixed-bed reactor. The catalysts were synthesized using a co-impregnation technique and activated via a two-step calcination at 350 °C for 1 h and 590 °C for 2 h at ramping rates of 10 °C/min and 15 °C/min, respectively. The catalysts contained various loadings (0-30 wt%) and combinations of Ni, Fe, Co, Mo, Zn, and W oxides supported on Al₂O₃. The reaction was tested using 0.3 g of catalyst at temperatures ranging from 400 to 500 °C and O₂/C₄ ratios of 1, 2, and 4 mol/mol. Each dataset of the input variable contained a mono, binary, ternary, or quaternary combination of metal oxides on the Al₂O₃ support at a particular temperature and O₂/C₄ feed ratio. The datasets combined both positive and negative data (data from failed experiments or non-selective catalysts, seldom reported in the literature). Thus, it allows ML algorithms to correctly learn the diverse nature of data. The ML models were trained for both n-butane conversion and 1,3-butadiene selectivity based on descriptors and measured variables. 1,3-butadiene selectivity was considered only for this study, while other products such as butenes, CO_x, and cracking products were not utilized.

5.4 Descriptors and Measured Variables

5.4.1 Outcome (measured) variables

Models were built for each outcome variable, butane conversion, and 1,3-butadiene selectivity, obtained using an online gas chromatograph connected to a fixed-bed reac-

tor. Flame ionization and thermal conductivity detectors with various columns were used to detect the products. Based on the carbon balance, the butane conversion and 1,3-butadiene selectivity were calculated using Equations (3.5) and (3.6).

$$\text{Butane Conversion (\%)} = \frac{\text{moles of butane fed} - \text{moles of butane in the product}}{\text{moles of butane fed}} \times 100 \quad (5.5)$$

$$\text{Product Selectivity (\%)} = \frac{\text{moles of product}}{\text{butane conversion}} \times 100 \quad (5.6)$$

5.4.2 Explanatory (descriptor) variables

Temperature (range: 400-500 °C), feed ratio of O₂ to n-butane (range: 1-4 mol/mol), and different loadings (0-30 wt%) of nickel oxides (Ni; range: 0-20), iron (Fe; range: 0-20), cobalt (Co; range: 0-20), bismuth (Bi; range: 0-30), molybdenum (Mo; range: 0-15), tungsten (W; range: 0-30), zinc (Zn; range: 0-20), and manganese (Mn; range: 0-20) were used as predictor variables. These variables were selected because of their direct effect on the performance of oxidative dehydrogenation catalysts [69]. The temperature, feed ratio, and catalytic composition (in terms of active metal loading) determine the extent of feed conversion and selectivity for both desired (and undesired) products. Higher temperatures and feed ratios are more suitable for cracking and deep oxidation products, whereas an optimum temperature and moderate feed ratio facilitate the dehydrogenation pathway. Similarly, the catalyst composition (and loading) affects catalytic properties, such as acidity, reducibility, nature of oxygen species, and

metal-support interaction, determining the selectivity towards the desired product [94].

5.5 Statistical Analysis

5.5.1 Data Description.

The dataset employed in this research comprises thirteen variables, consisting of eleven predictor variables and two response variables. The predictor variables encompass a range of factors, including, temperature measured in degrees Celsius, feed ratio (O_2/n -butane), and the loadings of various metal oxides such as nickel (Ni), iron (Fe), cobalt (Co), bismuth (Bi), molybdenum (Mo), tungsten (W), zinc (Zn), and manganese (Mn). The response variables, or dependent variables, in this study are the n -butane conversion and 1,3-Butadiene Selectivity. These variables serve as indicators of the reactor's performance and efficiency. The n -butane conversion reflects the extent to which n -butane is converted into other compounds, providing insights into the reaction's effectiveness. The 1,3-Butadiene Selectivity measures the selectivity of the reaction towards the desired product, 1,3-Butadiene, which is of particular interest in this study. By analyzing the dataset and exploring the relationships between the predictor variables and response variables, this study aims to gain a comprehensive understanding of the factors influencing n -butane conversion and 1,3-Butadiene Selectivity in the fixed bed reactor. The findings from this research have the potential to enhance knowledge about reaction kinetics, catalyst performance, and process optimization in the context of the target reactions.

5.5.2 Data Exploration.

To ensure the accuracy and reliability of the data collected from the experiment, a crucial preliminary step was taken to cleanse the data. This involved a thorough examination of the data to identify and address any missing values, data entry errors, and outliers. Outliers were given special attention during this process. Outliers are defined as observations that fall significantly outside of the expected range of values. These values can have a disproportionate impact on the analysis of the data and can potentially lead to erroneous conclusions if left unaddressed. Thus, it was imperative to identify and address any outliers present in the data. This involved carefully examining the distribution of the data and using statistical tools to detect any extreme values that could be considered outliers. Once identified, appropriate actions were taken to either remove or correct the outliers, such as imputation or transformation. Overall, the thorough cleansing of the data, including the identification and handling of outliers, ensured the validity and reliability of the subsequent analysis performed on the data, and provided a solid foundation for drawing accurate and meaningful conclusions from the experiment.

The data analysis was accompanied by the visual representations provided in **Figures 3.1 - 3.3**. **Figures 3.1** displays the data in the form of histograms, allowing us to examine the distributions of the variables in the study. It becomes evident from these histograms that the distributions are not in accordance with a normal distribution pattern. In **Figures 3.2**, scatter plots are utilized to illustrate the relationships between butane conversion and the predictor factors. One notable observation from

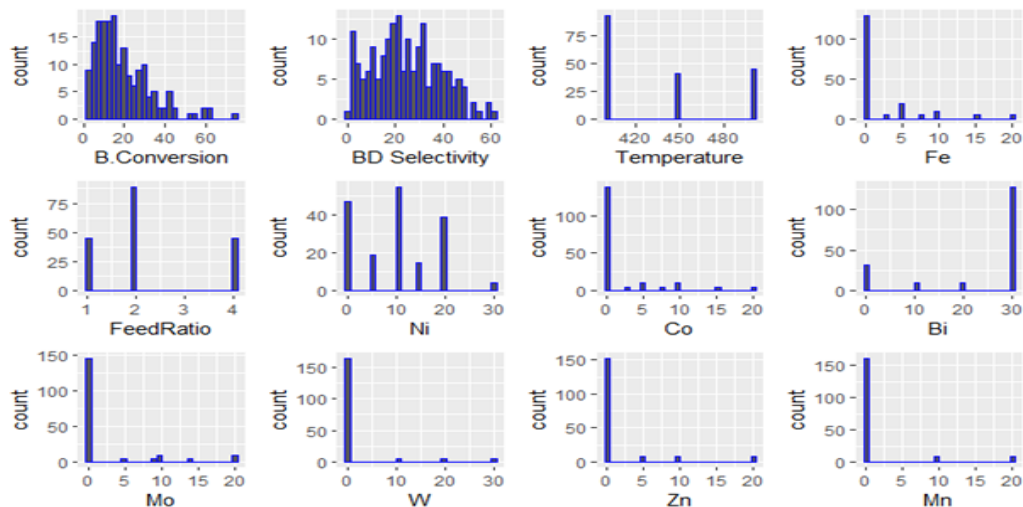


Figure 5.1: Histograms of variables

this figure is the positive linear relationship between feed ratio and butane conversion. As the feed ratio increases, there is a tendency for the percentage of butane conversion to also increase, and vice versa. On the other hand, the relationships between butane conversion and the remaining predictor variables appear to be non-linear, as evident from the plots in **Figures 3.2**. It is worth noting that although these relationships do not adhere to the assumptions of normality, there is no need for concern since machine learning techniques do not strictly rely on normality assumptions like classical statistical methods do. **Figures 3.3** focuses on the association between 1,3 BD selectivity and the predictor factors. The plot suggests a negative linear relationship between tungsten (W) and 1,3 BD selectivity. As the concentration of tungsten increases, there is a tendency for the selectivity of 1,3 BD to decrease. Conversely, the relationships between 1,3 BD selectivity and the other predictor variables exhibit non-linear patterns. To summarize, the visualization of the data in **Figures 3.1 - 3.3** provide valuable insights into the relationships between the variables under investiga-

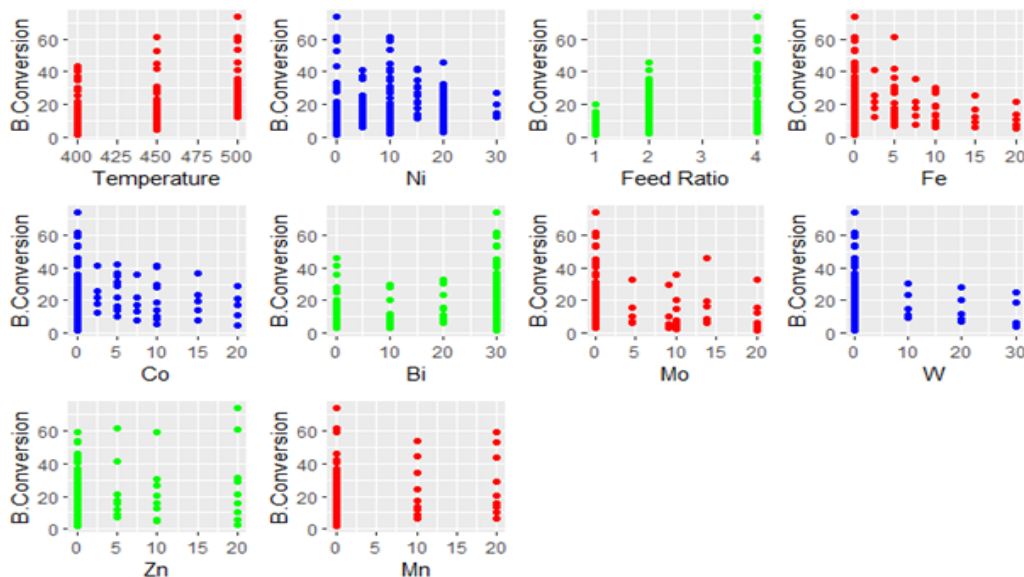


Figure 5.2: Scatter plots of Butane Conversion and predictor variables

tion. While some relationships demonstrate linear patterns, others exhibit non-linear associations, highlighting the complexity of the data.

5.5.3 Models Development

The datasets consisted of 304 data points, generated from experiments conducted in a fixed-bed reactor in the temperature range of 400-500°C and O_2/C_4 molar feed ratio of 1-4 mol/mol, using metal oxides of Nickel (Ni), Iron (Fe), Cobalt (Co), Bismuth (Bi), Molybdenum (Mo), Manganese (Mn), and Tungsten (W) supported on gamma alumina catalysts. A random 70% of the total dataset was allocated for model training, while the remaining 30% served as the evaluation or testing dataset. Ten well-known ML models: The Gaussian process regression (GPR), k-nearest neighbor (KNN), linear regression (LR), ridge regression (Ridge), quantile random forest (QRF), random forest (RF), support vector machine with linear (SVML) kernel function, support

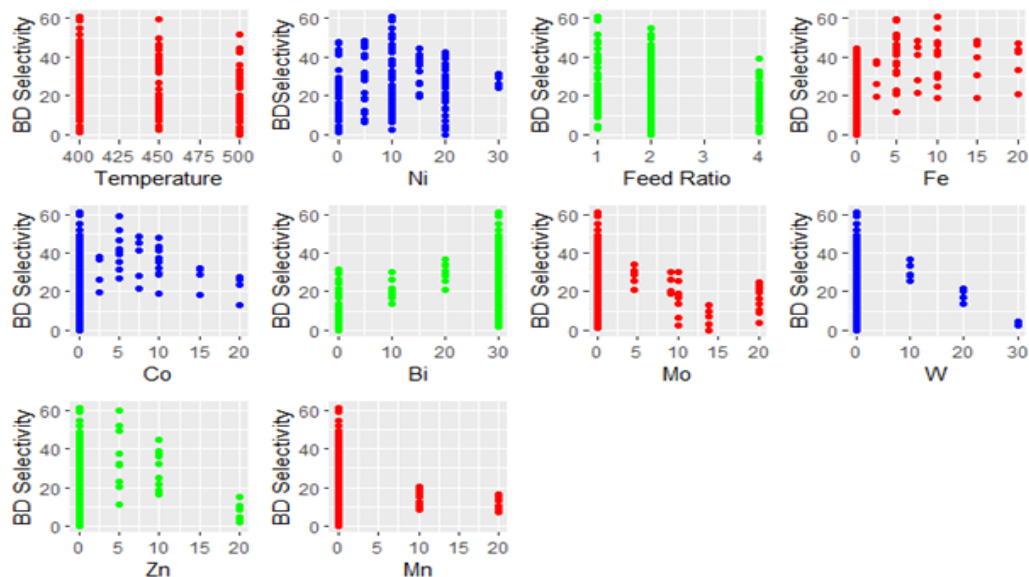


Figure 5.3: Scatter plots of 1,3 BD Selectivity and predictor variables

vector machine with radial (SVMR) basis kernel function, boosted tree (BT), and eXtreme gradient boosting, were used to predict both butane conversion and 1,3 BD selectivity. These models were developed using nested cross-validation for parameter optimization and a wrapper-based sequential backward feature-selection approach. The ML models were fitted using the caret library (version 6.0.93) in the R package.

5.5.4 Data Scaling

Each continuous variable in the training set was scaled to have a mean of zero and standard deviation of one using the z-score algorithm. The corresponding variables in the test set were mapped onto the training set.

5.5.5 Model Cross-Validation

Nested cross-validation involving multiple layers of stratified cross-validations (inner and outer folds) was performed on the training dataset to obtain a reliable classification accuracy and avoid overfitting. The inner folds were used to optimise the model parameters and select informative feature subsets, and the performance of the best (inner) model was then evaluated in the outer fold. For the outer fold, we split the training dataset into 10-fold cross-validation; one fold was kept as a test set, while the remaining nine folds (i.e. outer training fold) were, in turn, split in the inner fold into ten stratified folds, nine folds for model training and the remaining fold for validation, to provide an unbiased evaluation of the model fit on the inner training set while tuning the model’s hyperparameters and selecting optimal features. The outer and inner folds were repeated 20 times to obtain robust models.

5.5.6 Optimal feature selection and hyperparameters

Feature selection was conducted using sequential backward search selection (SBSS), which started with all features; it dropped the non-informative features at each iteration, improving the model’s performance. This process was continued until no further improvement was observed. Once the best combination of hyperparameters and feature subsets that maximised the performance metrics in the validation set was found, the model with the combination of hyperparameters and feature subsets was re-trained on the outer training set and tested on the test set kept out from the outer CV. Subsequently, the feature subsets from all outer folds were combined using a vot-

ing strategy that retained features with more than 50% occurrences in all outer folds as informative; hence, they were chosen as the final feature subset. The median of the best hyperparameters from the outer CV folds was used to fit the final model. Table 3.1 lists the optimal tuning parameters obtained after exploring a reasonable grid of candidate values for hyperparameter tuning.

Type	Subtype	Method	Optimal Hyperparameters [Feed conversion]	Optimal Hyperparameters [Budiene Selectivity]
Linear		LM	none	none
		Ridge	$\lambda = 0.35$	$\lambda = 0.05$
Non Linear	Kernel Methods	GPR	None	None
		SVMR	Sigma $\in [0.1119889]$ C = 7.742637	Sigma $\in [0.1119889]$ C = 2.154435
		SVML	C = 22.85464	C = 0.2728333
		K-NN	$k = 5$	k = 3
	Tree Ensemble Methods	RF	mtry = 3	mtry = 3
		BT	mstop = 150, nu = 0.1, maxdepth = 3	mstop = 150, nu = 0.1, maxdepth = 3
		XGBoost	nrounds = 100, $\lambda = 0.1$, alpha = 0.1, eta = 0.3	
		QRF	mtry = 3	mtry = 3

Table 5.1: The optimal hyperparameter searches for the ML for predicting both Feed Conversion and Butadiene Selectivity

5.5.7 Performance Evaluation

To generate summary performance estimates, we evaluated the model performance using the root mean squared error ($RMSE$), mean absolute error (MAE), and coefficient of determination (R^2). The $RMSE$ measures the deviation between the model-predicted and observed values. It was calculated by taking the square root of the mean squared difference between the predicted and actual values of the model. On the other hand, MAE is the average absolute difference between the values predicted by the models and the observed values. R^2 measures the variation in the outcome variable explained by the models.

5.6 Results and Discussion

5.6.1 Descriptive Statistics

The Pearson correlation matrix of the descriptor variables (temperature, F. ratio, Ni, Fe, Co, Bi, Mo, W, Zn, and Mn) is presented in Table 2. The diagonal line of the matrix shows the correlation of each variable with itself, which is always 1.0. The other cells show the correlation coefficients for each pair of variables. The correlation coefficient ranges from -1.0 to 1.0, where a value of -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

Moderate positive and negative correlations existed between the descriptor variables. For example, Fe and Bi had a positive correlation coefficient of 0.29, indicating that Bi and Fe have a direct correlation and similar effects on conversion and selec-

Temp.	1									
F.ratio	-0.07	1								
Ni	-0.1	-0.02	1							
Fe	-0.05	-0.02	-0.28	1						
Co	-0.07	-0.02	-0.28	0.014	1					
Bi	0.12	0.02	-0.47	0.29	0.26	1				
Mo	0.07	0.01	-0.05	-0.21	-0.2	-0.03	1			
W	-0.04	-0.01	0.36	-0.14	-0.12	-0.39	-0.12	1		
Zn	0.12	0.03	-0.26	-0.13	-0.2	0.21	-0.16	-0.1	1	
Mn	0.1	0.02	-0.26	-0.16	-0.14	0.19	-0.14	-0.09	-0.12	1

Table 5.2: The correlation matrix of the ten descriptor variables.

tivity. Similarly, Ni and Co have a negative correlation coefficient of -0.28, indicating that Co and Ni have negative correlations. Generally, the correlation between the predictor variables suggests no severe multicollinearity, which is desirable.

5.6.2 Model Comparisons for Predicting Butane Conversion

In the analysis of various models for predicting butane conversion (**Figure 1a**), the SVMR model stands out with the lowest training prediction errors, showcasing an RMSE of 1.8 and an MAE of 1.36. On average, this implies that predicted butane conversion values deviate from actual values by approximately 1.8 percentage points. Comparatively, the QRF model exhibits higher error rates, with RMSE and MAE values of 4.49 and 2.32, respectively. Following closely, the RF model reports RMSE and MAE values of 4.47 and 3.17, respectively. In contrast, the SVML and LR models show the highest training errors, with RMSE values of 7.97 and 7.43, and MAE values of 5.53 and 5.75, respectively for QRF. Examining test errors (**Figure 1b**), the SVMR model demonstrates the lowest error rates, reporting RMSE and MAE values of 4.98 and 3.28, respectively. Conversely, the LR model exhibits the highest test errors, with

RMSE and MAE values of 8.92 and 6.73, respectively.

Additionally, considering R^2 values (Figure 1c), the SVMR model stands out with the highest R^2 values of 98.3% and 88.3% for the training and test sets, respectively.

Following this, the RF model achieves R^2 values of 89.7% and 71.4% for the training and test sets, respectively. The QRF model exhibits R^2 values of 88.7% and 71.4%, while the SVML model demonstrates the lowest R^2 values of 66.1% and 62.9% for the training and test sets, respectively.

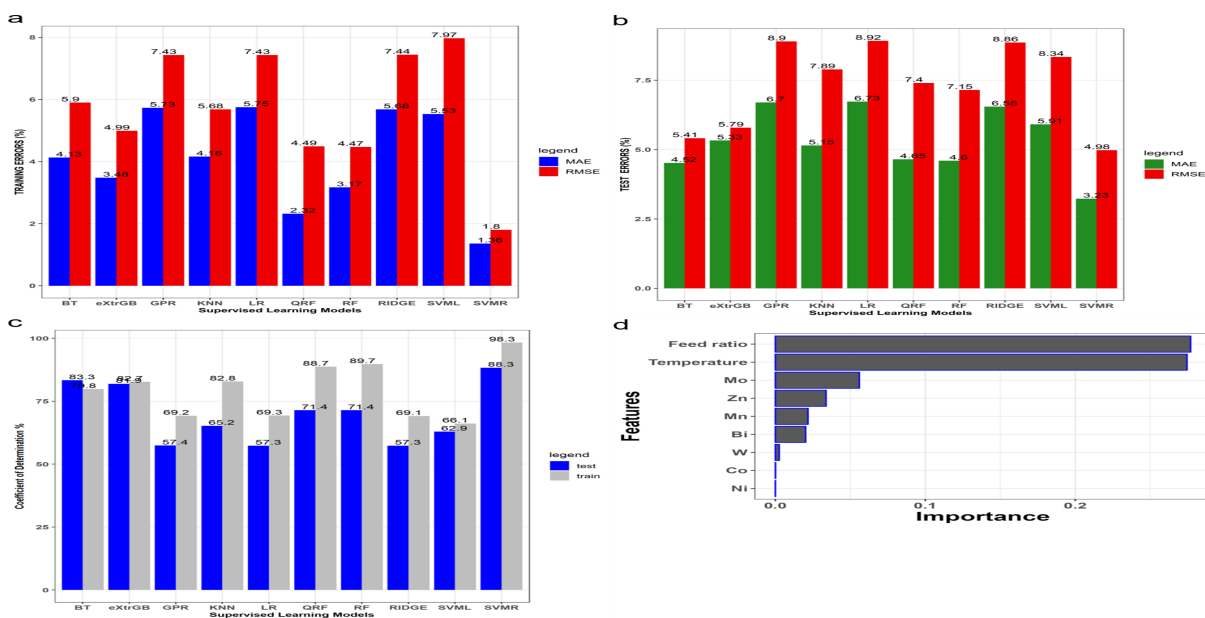


Figure 5.4: Performance of n-butane conversion-based models.

- Prediction training RMSE and MAE comparisons for n-butane conversion.
- Prediction test RMSE and MAE comparisons for n-Butane Conversion.
- R^2 values for training and test predictions of n-Butane Conversion.
- Variable importance plot for the SVMR model for predicting n-butane conversion.

Among all the models, the SVMR model exhibited the best performance in predicting butane conversion, with the lowest training and test errors and the highest R^2 value. The SVMR model included temperature, feed ratio, Ni, Co, Bi, Mo, W, Zn, and Mn as predictor variables (Table 6.2). The models are given in rows, and the predictor variables are shown in columns. The selected predictor variables are marked with a check mark (✓). However, crossed (×) indicates variable(s) that the model did not select. The model indicated that feed ratio and temperature were the most important variables in predicting butane conversion. In contrast, Co, W, and Ni had less impact on the prediction accuracy, as depicted in Figure 3.1d. Interestingly, Fe did not contribute significantly to the model, indicating that it did not provide any relevant information for predicting the butane conversion.

MODELS	Temp.	Feed ratio	Ni	Co	Bi	Mo	W	Zn	Fe	Mn
GPR	✓	✓	×	×	✓	✓	✓	×	✓	×
KNN	✓	✓	✓	✓	✓	✓	×	×	×	×
LR	✓	✓	×	×	✓	✓	✓	×	✓	×
QRF	✓	✓	×	✓	✓	✓	×	×	✓	×
RF	✓	✓	×	✓	✓	✓	✓	×	✓	✓
SVML	✓	✓	×	✓	✓	✓	×	×	×	×
SVMR	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
RIDGE	✓	✓	×	×	✓	✓	✓	×	✓	×

Table 5.3: Variables selected by each ML model for predicting butane conversion.

5.6.3 Model Comparisons for Predicting 1,3-Butadiene Selectivity

Comparisons of the training RMSE and MAE values for the various ML models used to predict the 1,3-Butadiene Selectivity are shown in **Figure2a**. The SVMR model

demonstrated the lowest training prediction errors with an RMSE of 3.57 and an MAE of 2.3. The QRF model followed closely, with an RMSE of 3.94 and an MAE of 2.35. In contrast, the GPR and LR models displayed the highest training errors: GPR (RMSE = 7.2 and MAE = 5.59) and LR (RMSE = 7.2, MAE = 5.6). Figure 2b compares the test RMSE and MAE of the models, showing that SVMR and QRF had the lowest test errors: SVMR (RMSE = 4.86 and MAE = 3.57) and QRF (RMSE = 4.63 and MAE = 3.77). In contrast, the GPR and LR models exhibited the highest test errors, with GPR at 7.77 (RMSE) and 6.76 (MAE) and LR at 7.82 (RMSE) and 6.76 (MAE).

In addition, SVMR displayed the highest R^2 values for both the training and test sets, with 94% for the training set and 88.6% for the test set, as shown in **Figure 2c**. By contrast, the SVML model had the lowest R^2 values of 73.4% for the training set and 72.3% for the test set. The SVMR model included temperature, feed ratio, Ni, Co, Bi, Mo, W, Zn, and Mn (**Table 3.4**). The feed ratio and temperature were highly important for predicting butane conversion, whereas Co, W, and Ni were deemed less important (**Figure 2d**). However, Fe did not contribute valuable information to the model.

- a. Prediction training RMSE and MAE comparisons for 1,3-Butadiene selectivity.
- b. Prediction test RMSE and MAE comparisons for 1,3-Butadiene selectivity.
- c. R^2 values for training and test predictions of 1,3-Butadiene selectivity.

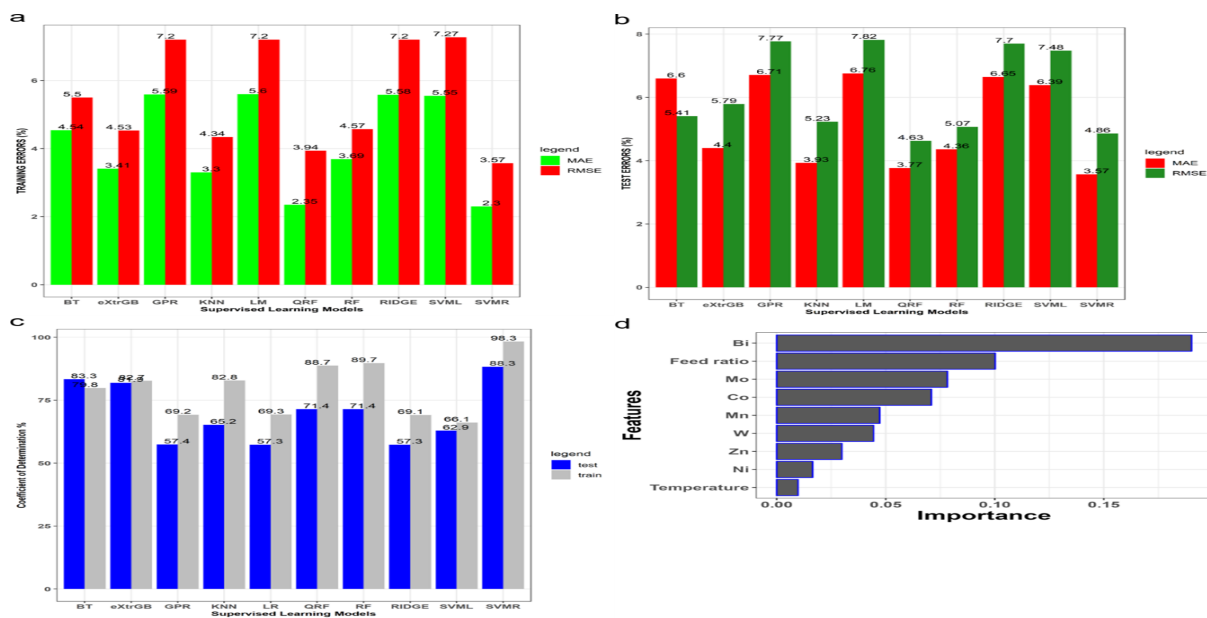


Figure 5.5: Performance of 1,3-Butadiene selectivity-based models

d. Variable importance plot for the SVMR model for predicting 1,3-Butadiene selectivity.

MODELS	Temp.	Feed ratio	Ni	Co	Bi	Mo	W	Zn	Fe	Mn
GPR	✓	✓	✓	×	✓	✓	✓	×	✓	✓
KNN	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
LR	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
QRF	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
RF	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
SVML	✓	✓	×	✓	×	✓	✓	✓	×	×
SVMR	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
RIDGE	×	✓	✓	×	✓	✓	✓	✓	✓	×

Table 5.4: Selected features for predicting 1,3 BD selectivity using the models

5.6.4 Best Performing Model

Graphical representations of the actual observed values from the experiment and the predicted values generated by the SVMR model for both butane conversion and BD

selectivity are shown in Figures 3a and b, respectively. These figures demonstrate that the predicted value lines are closely aligned with the observed value lines in both cases, indicating that the SVMR model accurately captures the relationships between the independent and dependent variables within the dataset. The close alignment of the blue (observed) and red (predicted) lines in these figures highlights the strong performance of the SVMR model in predicting the values for butane conversion and BD selectivity. This congruence suggests that the model effectively captured the underlying patterns in the data while maintaining a low error rate, thereby emphasising its reliability and accuracy in predicting these variables.

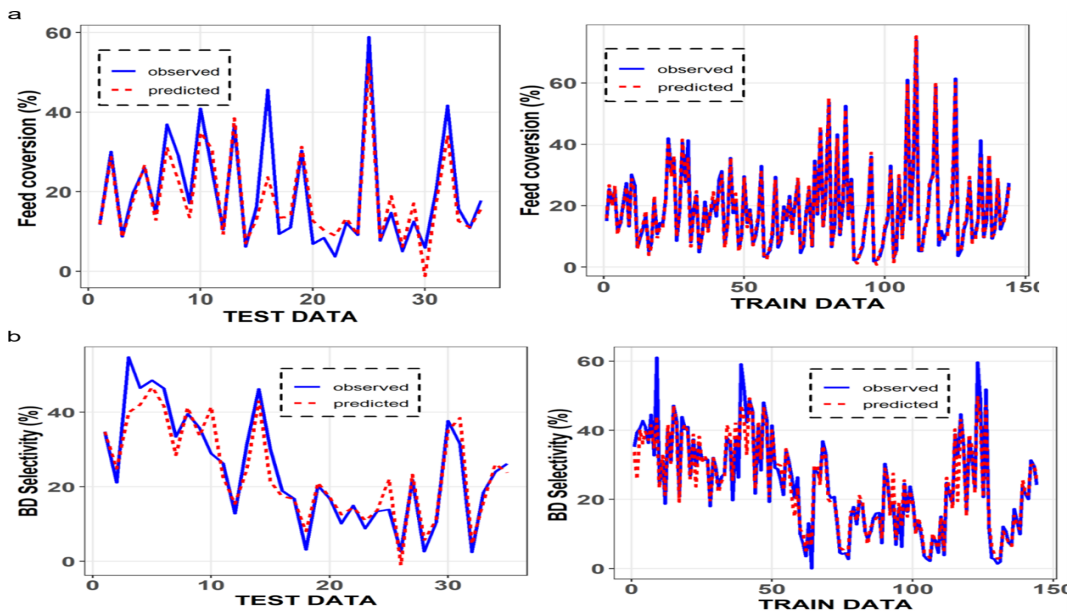


Figure 5.6: SVMR Models performance

- a. Comparison of the observed and predicted butane conversion values.
- b. Comparison of the observed and predicted 1,3 BD selectivity values. For a and b, the diagrams on the right in the figures are for the training data, whereas those on the left are for the test data.

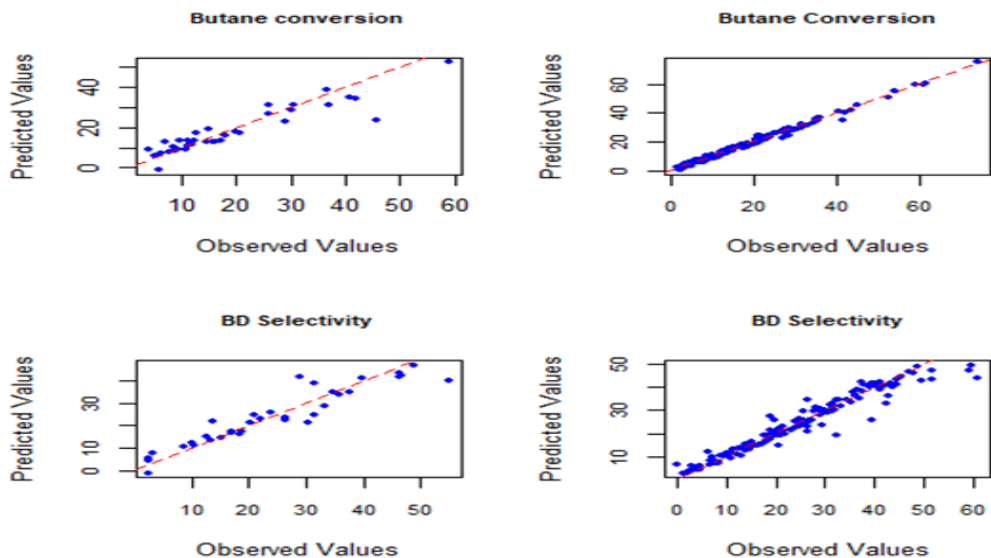


Figure 5.7: Parity plots of the SVMR model for butane conversion and BD selectivity

Predictive performance compares the predicted values from the SVMR model with actual observed values. For parts **a** and **b**, the diagrams on the right in the figures are for the training data, whereas those on the left are for the test data. The ideal prediction is represented by a straight line where the predicted value equals the true value. In this case, because most points are close to the equality line, the SVMR model performs well in predicting the butane conversion and 1,3 BD selectivity values. It is worth noting that a few points may be scattered away from the equality line, indicating some prediction errors. However, the overall performance of the model was reliable. Moreover, the consistent performance of the model across the training and test datasets indicates its resistance to overfitting (Figure 4). This characteristic implies that the SVMR model successfully generalizes new data instead of merely memorizing patterns within the training dataset. Hence, the SVMR model demonstrated robust and effective performance in predicting butane conversion and 1,3 BD

selectivity. Feed conversion has been proven to increase with both temperature and feed ratio for oxidative dehydrogenation reactions. This is because, at a higher temperature, the metal species become more activated, thereby increasing the transformation routes of the feed to the various product pathways with high conversion. However, the selectivity for dehydrogenation products decreases at higher temperatures because of the increase in the competing cracking and partial/complete oxidation pathways [95]. Interestingly, both SVMR and LR models predict increases in temperature to affect feed conversion and 1,3-Butadiene selectivity, respectively. Similarly, an increase in the feed ratio enhanced the availability of both lattice and adsorbed oxygen species. These oxygen species improved feed conversion and butadiene selectivity. In addition, these models have proven that the O_2 /n-butane ratio is vital for both conversion and selectivity predictions. Both models showed that the presence of BiO_x species contributed significantly to 1,3-butadiene selectivity.

5.7 Summary

The SVMR model outperformed all other models in predicting both 1,3 BD Selectivity and n-butane conversion, exhibiting the lowest training prediction errors with an RMSE of 3.57 and an MAE of 2.3. Additionally, it achieved the highest R^2 values for both the training and test sets, with 94% for the training set and 88.6% for the test set. The key predictor variables for the SVMR model included temperature, feed ratio, and the presence of Ni, Co, Bi, Mo, W, Zn, and Mn.

Based on the models' prediction analyses, it can be concluded that Support Vector Machine with Radial base kernel function(SVMR) models are the most suitable for both n-butane conversion and 1,3-butadiene selectivity predictions. Hence, these models can be utilized to design highly active and selective oxidative dehydrogenation catalysts. In addition to the developed models, this study introduced an innovative digital tool, a specially designed application that integrates proven Machine Learning (ML) algorithms to offer instantaneous on-the-spot predictions of catalyst performance in the oxidative dehydrogenation of n-butane to 1,3-butadiene. By utilizing this application, researchers, scientists, and industry professionals can swiftly determine the optimal conditions for catalyst operation, facilitating improvements in catalyst design and process efficiency. In addition to harnessing the power of our models, this application serves as a bridge connecting high-level ML insights with practical real-world applications, thus making a substantial contribution to the field of heterogeneous catalysis. The link to the app is provided below for readers interested in exploring the transformative potential of this application: Oxidative Dehydrogenation Catalyst Application. This fusion of catalyst informatics and explainable AI aids in simplifying complex heterogeneous catalytic processes, shedding light on the catalyst performance, and opening new horizons for catalyst optimization.

CHAPTER 6

MACHINE LEARNING ADVANCEMENTS IN CATALYSIS: AN ENSEMBLE MODEL APPROACH TO PREDICT DEHYDROGENATION YIELD

6.1 Objective

The objective of this chapter is to develop a predictive machine learning model using ensemble technique for predicting Dehydrogenation yield. The ensemble approach will

help enhance predictive capability of the model

6.2 Introduction

The demand for butenes and 1,3-butadiene, essential building blocks of the petrochemical industry, continues to grow. The conventional production route via direct dehydrogenation (DDH) of n-butane using platinum (Pt)-and chromium (CrOx)-based catalysts [84], [85], [86] faces two major challenges: catalyst deactivation due to coke deposition and the need for high reaction temperatures to achieve reasonable conversions. To address these limitations and seek more efficient and sustainable alternatives, researchers have explored oxidative dehydrogenation (ODH) of n-butane as a promising alternative.

ODH represents a compelling approach, as it involves co-feeding n-butane with oxidants, such as O_2 , CO_2 , or N_2O , which has been shown to minimize catalyst deactivation and allow for operation at lower temperatures, thus enhancing energy efficiency and prolonging catalyst lifetime [87], [88], [89], [90], [91], [92]. Despite the potential advantages of ODH, the commercialization of this process has been hindered primarily by the challenge of achieving the desired selectivity for the DH products (butenes and 1,3-butadiene). The overoxidation of butane and intermediate butenes to stable oxidation products (CO_x) reduces the selectivity for the desired products.

Researchers have focused on the emerging field of catalyst informatics to overcome this critical challenge and accelerate the development of high-performance ODH catalysts. Catalyst informatics leverages data science techniques to design optimal

catalysts for specific reactions and mechanisms [4], [5]. It encompasses diverse aspects such as dataset generation, data preprocessing, data visualization, and machine learning (ML). ML plays a pivotal role in identifying meaningful relationships between catalyst descriptors and performance metrics, thereby expediting the exploration of catalytic materials and their properties [4], [68]. The use of ML in heterogeneous catalysis dates back to the mid-1990s when it was first used to enhance the discovery of novel catalysts. Over the years, ML techniques have been increasingly applied to predict catalyst performance [66], optimize reaction conditions through data-driven approaches [67], and discover previously unknown catalysts [68], particularly for reactions such as oxidative coupling of methane (OCM).

Building on these successes, Madaan et al. pioneered the use of descriptor models based on radial distribution functions (RDF) to predict the performance of bimetallic mixed-oxide-supported catalysts for the n-butane ODH. Their approach demonstrated impressive accuracy, achieving over 90% prediction accuracy for a new set of bimetallic oxides [96]. In a recent study, our group extended this methodology to predict both the conversion of n-butane and selectivity for 1,3-butadiene during ODH by employing different ML algorithms, including linear and nonlinear methods. Among these algorithms, the support vector machine with radial basis function (SVMR) model has emerged as the top-performing predictor, yielding high coefficients of determination (R^2), low mean absolute errors (MAE), and root mean square errors (RMSE) [97].

In this study, we further advanced the application of ML to the ODH of n-butane by developing ML models to predict the yield of DH products. The novelty of this manuscript lies in the utilization of ensemble techniques for the three developed mod-

els to further enhance prediction ability. Typically, the ODH reaction involves three competitive pathways: dehydrogenation (DH), cracking, and partial/complete oxidation. The motivation behind this study is twofold: first, to address the challenge of selectivity control and increase the understanding of catalyst design principles for desirable product formation (mainly the DH pathway), and second, to advance the field of catalyst informatics as a valuable and efficient tool for discovering high-performance ODH catalysts. Through a comprehensive analysis of ML model predictions, we aimed to gain insights into the factors influencing product yield, paving the way for future catalyst design and process optimization.

6.3 Experimental Setup for Dataset Generation

Herein, we present a comprehensive dataset derived from a series of experiments for the development of efficient and selective catalysts for the oxidative dehydrogenation (ODH) of n-butane. The dataset was collected from a fixed-bed flow reactor using catalysts prepared via co-impregnation and activated via a two-step calcination process. The calcination process involved two heating steps: the first at 350 °C for one hour and the second at 590 °C for two hours, with ramping rates of 10 °C/min and 15 °C/min, respectively. These experiments involved various combinations of Ni, Fe, Co, Mo, Zn, and W oxides supported on γ -Al₂O₃, with loadings ranging from 0 to 30 wt%. The reaction was tested using 300 mg of each catalyst, and temperatures ranging from 400 to 500 °C were explored, along with O_2/C_4 feed ratios of 1, 2, and 4 mol/mol. The details of the choice of metal oxides and supports, catalyst syntheses,

characterizations, and performance evaluation can be found in previous experimental studies [91], [90], [92].

Each dataset entry represents a specific combination of metal oxides on the Al_2O_3 support tested under particular temperature and O_2/C_4 feed ratio conditions. One noteworthy aspect of our dataset was the inclusion of both positive and negative data. Positive data correspond to successful and selective catalysts, whereas negative data refer to the results from non-selective catalysts. Including such diverse data types, which are often underrepresented in the literature, enhances the richness and complexity of the dataset. This allows machine learning (ML) algorithms to better understand the underlying relationships and provides a basis for accurate learning and modeling. The primary focus of our study was to predict the dehydrogenation (DH) yield as the outcome of interest. These models exploit the wealth of information in diverse datasets to identify the key features influencing yield and ultimately accelerate the discovery of high-performance catalysts for the ODH of n-butane.

6.3.1 Measurement and Calculation of Outcome Variables in ODH of n-Butane

This study determined the primary variable of interest, particularly dehydrogenation (DH) yield, using an online gas chromatograph connected to a fixed-bed reactor. We utilized flame ionization and thermal conductivity detectors with various columns to detect the reaction products. The fundamental outcome metric, DH yield, was calculated from the butane conversion and DH selectivity based on carbon balance

principles (Equations 1–3).

Butane conversion (%) is the percentage of butane converted during the oxidative dehydrogenation (ODH) reaction.

$$\text{Butane conversion (\%)} = \frac{\text{moles of butane fed} - \text{moles of butane in the product}}{\text{moles of butane fed}} \times 100 \quad (6.1)$$

The DH selectivity (%) represents the percentage of the DH product formed relative to the extent of butane conversion.

$$\text{DH selectivity (\%)} = \frac{\text{moles of DH product}}{\text{Butane conversion}} \times 100 \quad (6.2)$$

The DH yield indicates the overall efficiency of the ODH process in producing the desired DH product.

$$\text{DH yield (\%)} = \frac{\text{Butane conversion} \times \text{DH selectivity}}{100} \quad (6.3)$$

This outcome variable, namely the DH yield, is an essential indicator of the performance of the ODH reaction and plays a crucial role in understanding the activity and selectivity of the catalysts under investigation. By accurately measuring and calculating this metric, we evaluated the effectiveness of different catalyst formulations and optimized the design of high-performance catalysts for ODH of n-butane.

6.3.2 Descriptor Variables Influencing the Performance of ODH Catalysts

We explored a carefully selected set of descriptor variables that significantly influenced the performance of the oxidative dehydrogenation (ODH) catalysts. These variables encompass temperature (ranging from 400 to 500 °C), the feed ratio of O_2 to n-butane (ranging from 1 to 4 mol/mol), and varying loadings (0-30 wt%) of several metal oxides, including nickel (Ni; 0-20 wt%), iron (Fe; 0-20 wt%), cobalt (Co; 0-20 wt%), bismuth (Bi; 0-30 wt%), molybdenum (Mo; 0-15 wt%), tungsten (W; 0-30 wt%), zinc (Zn; 0-20 wt%), and manganese (Mn; 0-20 wt%). These descriptor variables play a pivotal role in shaping the outcomes of catalytic processes. The temperature and feed ratio directly affected the degree of feed conversion and selectivity towards the desired and undesired products. Higher temperatures and feed ratios tend to favor cracking and deep oxidation products. In contrast, an optimal temperature and moderate feed ratio facilitate the DH pathway, leading to desired product formation [69].

Furthermore, the metal oxides composition and loading in the catalyst exert a profound influence on its properties, such as acidity, reducibility, nature of the oxygen species, and interaction between the metal(s) and support [94]. These factors, in turn, decisively affect the selectivity towards the desired product, ultimately determining the effectiveness of the catalyst in the ODH reaction [98], [95]. By exploring and understanding the role of these descriptor variables, we aimed to unlock insights into the intricate mechanisms governing the ODH reactions. This knowledge is instrumental in guiding the rational design of high-performance catalysts for this industrially

significant process, paving the way for more efficient and sustainable petrochemical production in the future.

6.4 Statistical Analysis

6.4.1 Data Exploration

The dataset used in this study was derived from experimental testing conducted in a fixed-bed reactor. It comprises 11 variables, with one being the response variable (DH yield) and the remaining ten being predictor variables. The predictor variables included temperature (measured in degrees Celsius), feed ratio (O_2/n -butane), and various loadings (ranging from 0 to 30 wt%) of metal oxides, specifically Ni, Fe, Co, Bi, Mo, W, Zn, and Mn. Throughout the experiments, 185 observations were collected for each variable to provide a comprehensive dataset for the analysis. This extensive data collection enabled a thorough exploration of the behavior of the fixed-bed reactor and its performance under different conditions.

Our initial analysis started with data exploration to comprehensively understand the dataset and identify patterns and relationships among the variables. Two primary methods, namely, the correlation plot and pair plot, were employed for this purpose. The correlation plot provides significant insights into the relationships between the predictor variables. This allowed us to visualize both positive and negative associations among the variables. By examining the correlation coefficients, we identified strongly correlated variables that showed weak or negligible correlations. This information is crucial for understanding the interdependencies among predictor variables

and for assessing potential multicollinearity issues.

In addition, a pair plot provides a detailed overview of the distributions and relationships between the selected variables (**Figure 4.1**). Each subplot on the diagonal presents a histogram of a single variable, providing a visual representation of its distribution. Additionally, the off-diagonal plots consisted of scatter plots between pairs of variables, revealing the potential correlations between them. Through this visualization, we gained insight into the nature of the relationships and the presence of linear or nonlinear associations. By conducting thorough data exploration, we were able to identify significant trends and patterns within the dataset. These insights were instrumental in guiding subsequent analyses and model development, ensuring that DH could be accurately predicted based on predictor variables. The data exploration process laid the foundation for a deeper understanding of the factors that influence the DH yield.

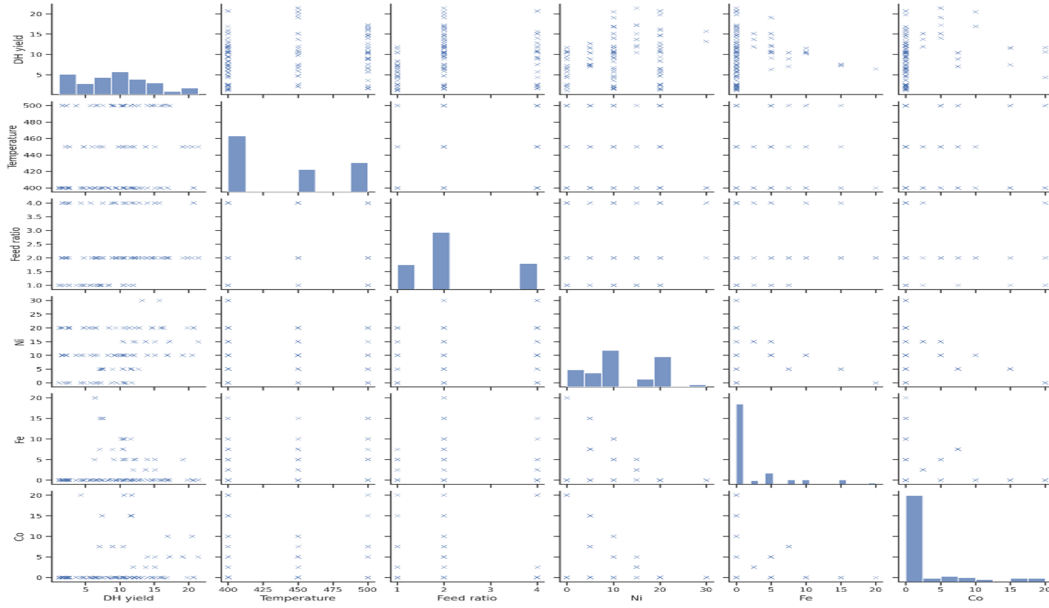


Figure 6.1: Pairwise Relationships and Distributions of Selected Variables

6.4.2 Models Development

This study aimed to utilize an Ensemble model to predict DH yield, integrating three well-established supervised ML models: Boosted Tree (BT) [99], extreme Gradient Boosting Linear (XGBL) [100], and Support Vector Machine Radial (SVMR) [53]. The process involved developing individual models for each algorithm, conducting nested cross-validation for optimal parameter tuning and feature selection. The ensemble model was formed by averaging the predictions from the three individual models. The training dataset, comprising 80% of the total data, was employed for model training, while the remaining 20% served as a testing dataset for performance evaluation. Evaluation metrics, including R^2 , RMSE, and MAE, were applied to assess the model performance. The models were built using the caret library (version 6.0.94) in R [101], [102].

6.4.3 Data Scaling

Data scaling was performed to ensure that the ML models were not biased towards specific features owing to the magnitude differences. Continuous variables in the training set were transformed using the z-score algorithm, standardizing the data to have a mean of 0 and a standard deviation of 1. This transformation facilitates fair comparison and analysis of the features. The test set variables were scaled similarly to ensure that the models were evaluated on data with consistent scaling, promoting better performance and training stability, particularly for sensitive models such as support vector machines.

6.4.4 Model Cross-Validation

To assess the model’s generalization performance and address overfitting, a robust cross-validation strategy was employed. The nested cross-validation method was utilized for this purpose. The dataset was divided into ten equally sized subsets, or folds, for the outer loop. Within each fold, further sub-setting occurred for the inner loop of the nested cross-validation process. In the inner loop, the model parameters were tuned to identify the optimal configuration, which was then used in the outer loop for model development. Throughout each iteration, the models underwent training on specific training subsets and were subsequently evaluated on validation sets to comprehensively assess their performance. This iterative process aids in obtaining reliable performance estimates by minimizing the impact of random partitioning. The use of ten folds and the nested structure contributes to a thorough evaluation, enhancing the robustness of the cross-validation approach.

6.4.5 Optimal Feature Selection and Hyperparameters

Backward elimination (BE) was used to identify the most relevant features. The least significant variable was removed iteratively based on the R^2 , leading to a subset of features with the greatest impact on DH yield, enhancing model simplicity and interpretability. Grids of the hyperparameters were generated for each algorithm, and different configurations were explored. The wrapper selection identified the optimal model based on the R^2 , and the selected features with the best-performing model were recorded. The hyperparameter results are presented in **6.1**

Type	Method	Optimal Hyperparameters [Dehydrogenation Yield]
Kernel method	SVMR	$\Sigma \in [0.2), C = 1.0$
Tree Ensemble Methods	BT	$mtry = 46.66$
	XGBL	$nrounds = 224.13, \lambda = 0.2, \alpha = 0.062$, and $\eta = 0.3$

Table 6.1: The optimal hyperparameter searches for the individual ML models in predicting the DH yield

6.5 Results and Discussion

6.5.1 Descriptive Statistics

This section provides information on the descriptive statistical analysis of the datasets. Typically, the Pearson correlation matrix is presented in **6.2**. This gives a comparative analysis of the relationships between the descriptor variables.

As shown in **6.2**, the correlation matrix offers valuable insights into the relationship between the outcomes and predictor variables. Moderate positive correlations are observed for Ni (0.27), Co (0.25), and Bi (0.17). These findings suggest that augmentation of these variables could potentially increase DH yield. A similar yet weaker positive correlation was found with the feed ratio and temperature variables, suggesting that although these may affect DH yield, they are not the predominant influencers. In contrast, the Mn and Fe parameters exhibited almost negligible correlations with DH yield, implying an insignificant linear relationship. However, the Zn and W parameters exhibited weak negative correlations with DH yield, suggesting a potential decrease in DH yield with an increase in these variables. Notably, Mo was found to have a strong negative correlation with DH yield, indicating a significant inverse

relationship. Furthermore, positive and negative associations with varying correlation coefficients were observed, indicating distinct dynamics among the predictor variables. Notably, the feed ratio and W correlation were minimal (coefficient: -0.01), suggesting a negligible negative relationship. In contrast, Ni and Bi displayed a relatively strong negative correlation (coefficient: -0.5), indicating an inverse relationship between these variables. Importantly, our dataset exhibited no severe multicollinearity among the predictor variables. Multicollinearity, the high linear relationship between predictors, can impact the interpretability and stability of the regression models. The absence of such issues ensures the reliability of the predictor variables when used in regression models for predictions.

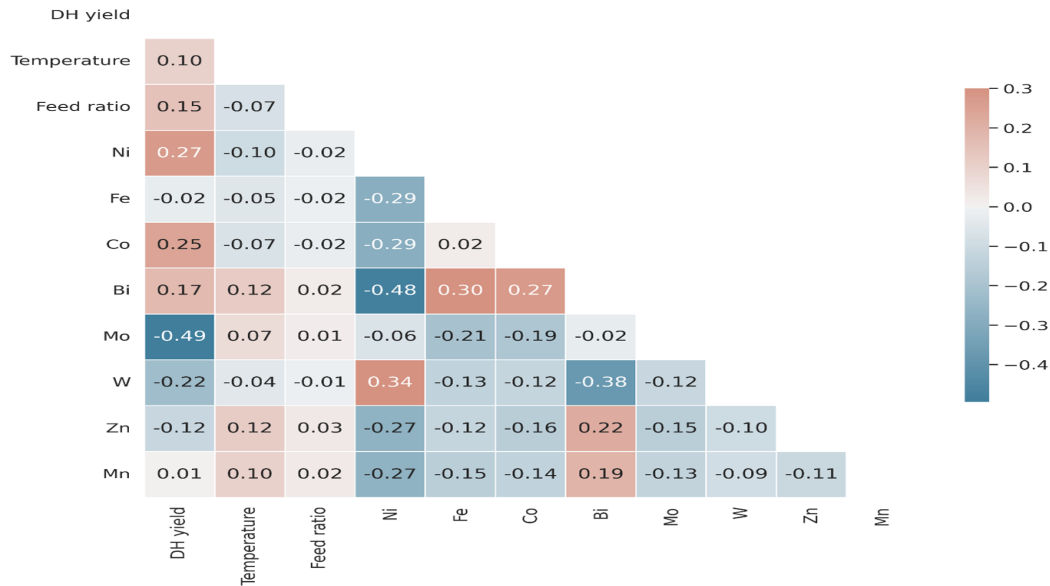


Figure 6.2: Heatmap of the lower triangular correlation matrix for predictor and outcome variables

Each cell in the heatmap represents the correlation coefficient between two variables, ranging from -1 (perfect negative correlation, shown in blue) to +1 (perfect positive correlation, shown in red). Values close to 0 (white) indicate little to no

linear relationship.

6.5.2 Model Performance

6.2 and 6.3 present a comprehensive overview of the performance exhibited by both individual models and the ensemble method on both the training and test datasets. Among the models evaluated, the BT model achieved an RMSE of 0.39 and MAE of 0.25 for the training data, along with an RMSE of 1.78 and MAE of 1.27 for the test data. Notably, it demonstrated R^2 values of 99.4% and 85% for the training and test datasets, respectively. For the XGBL model, the training data showed an RMSE of 0.05 and MAE of 0.04, while the test data exhibited an RMSE of 1.87 and MAE of 1.30. The model achieved R^2 values of 99.9% for training and 83.4% for the test dataset. In the case of SVMR, the training data yielded an RMSE of 1.45 and MAE of 0.95, while the test data reported an RMSE of 2.05 and MAE of 1.38. Interestingly, the ensemble of the three models demonstrated an RMSE of 0.56 and MAE of 0.37 for the training data, and an RMSE of 1.65 and MAE of 1.14 for the test data. Remarkably, it displayed R^2 values of 98.8% and 87.1% for the training and test datasets, respectively. An RMSE of 0.56 suggests that, on average, the Ensemble method predictions deviate from the actual dehydrogenation yields by approximately 0.56 percentage points where's an R^2 of 98.8% suggests that the features employed in the study within the Ensemble method's prediction of dehydrogenation yield can explain 98.8% of the variability observed in the actual dehydrogenation yields. In summary, the results indicate exceptional performance from each individual model, with the ensemble of the three models showcasing superior outcomes for the test data,

achieving the lowest test RMSE and MAE coupled with the highest R^2 .

	RMSE		MAE	
	Train	Test	Train	Test
BT	0.39	1.78	0.25	1.27
XGBL	0.05	1.87	0.04	1.30
SVMR	1.45	2.05	0.95	1.38
Ensemble	0.56	1.65	0.37	1.14

Table 6.2: Performance evaluation of the individual models and the ensemble for predicting DH Yield

Feature importance analysis revealed the critical variables for predicting DH yield from the BT model **(6.4)**. Among the input variables, Mo was the most crucial factor in predicting DH yield, followed closely by Ni, which also exhibited a significant influence. In addition, W was found to be of significant importance in the prediction process. On the other hand, Mn and the feed ratio appeared to have comparatively less impact on the DH yield predictions. The BT model effectively incorporates all predictor variables, indicating their collective contribution to the overall prediction. This feature importance analysis provided valuable insights into the key factors affecting DH yield and highlighted the significance of Mo, Ni, and W in the ODH reaction.

6.5 illustrates parity plots, visually comparing predicted DH yield values with actual observed values for individual models (BT, XGBL, SVMR) and the ensemble method. Parity plots offer a holistic assessment of predictive accuracy, with each data point representing a predicted DH yield value plotted against its corresponding actual value. In the case of the ensemble method, a notable concentration of data points is observed around the equality line, indicating strong agreement between ensemble method predictions and actual values. This clustering underscores the ensemble method's capability to accurately capture underlying patterns and trends in the DH

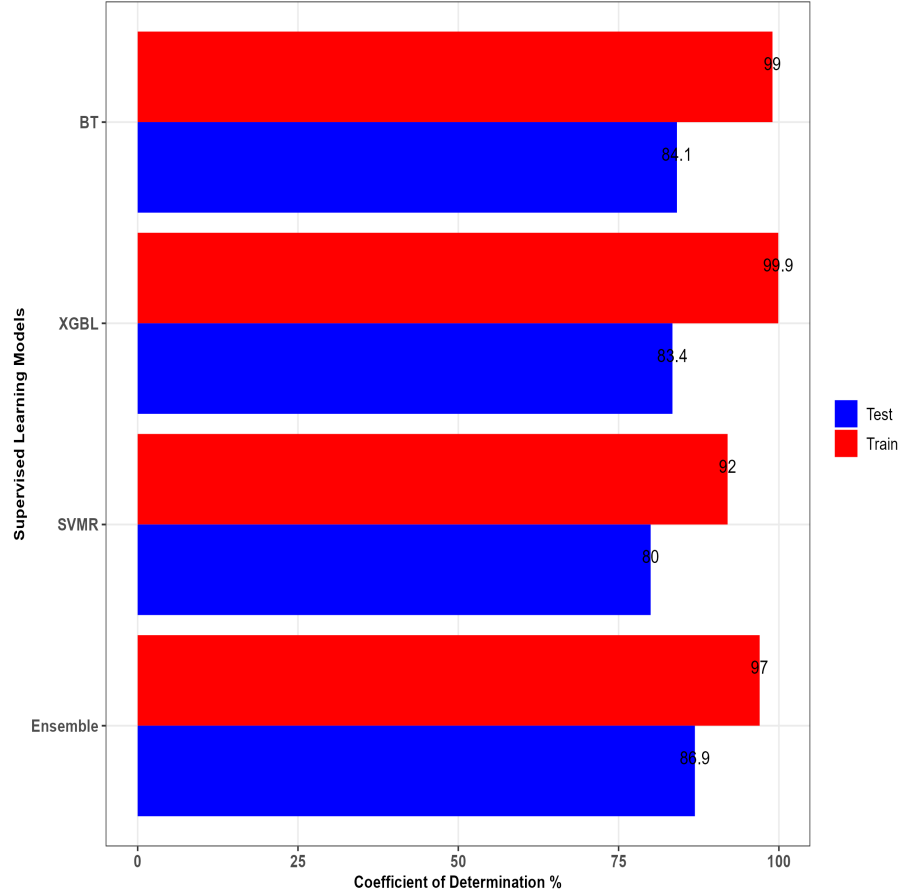
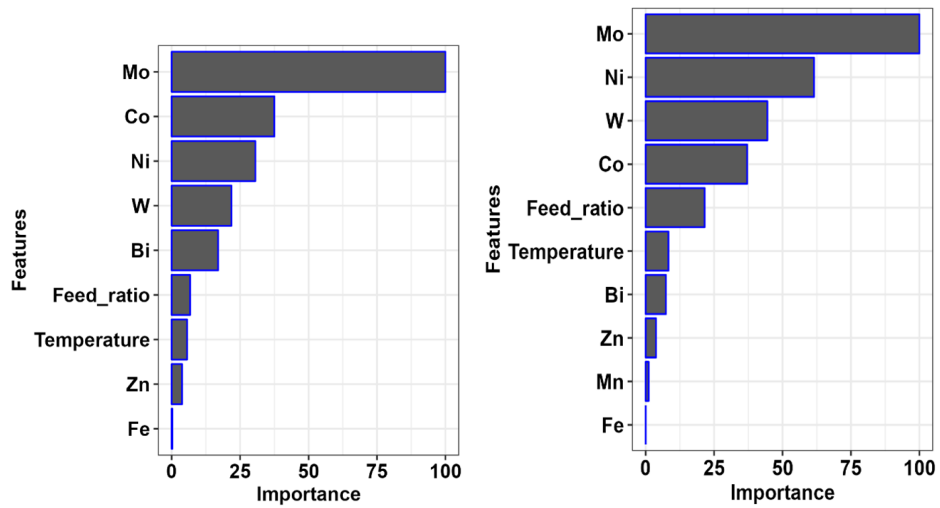


Figure 6.3: Coefficient of determination (R^2) for the ML Models

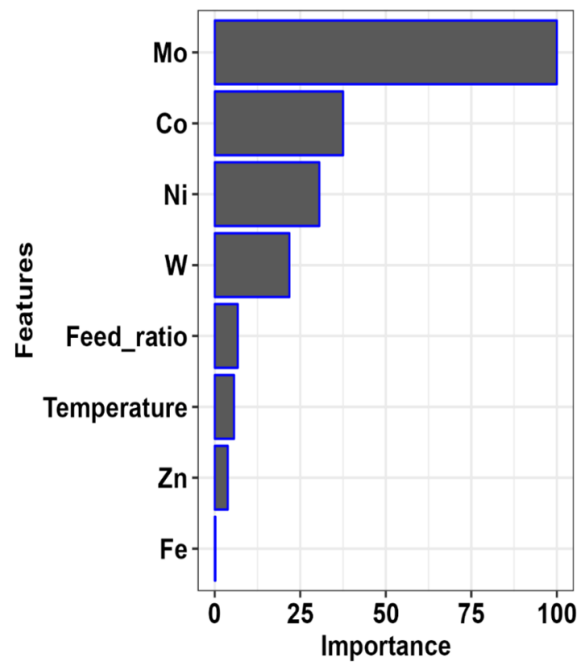
yield dataset. While some deviations from the equality line exist, the overall proximity suggests the ensemble method's reliable and consistent predictive performance.

The parity plot also reveals that individual models demonstrate a close alignment between predicted and observed DH yield values. Despite occasional deviations, the overall pattern along the equality line signifies reasonable predictive capabilities based on specified predictor variables. Collectively, parity plots visually confirm the predictive performance of BT, XGBL, SVMR models, and their ensemble method for DH yield. These visualizations validate the models' ability to approximate true values and offer insights into their accuracy and reliability in real-world scenarios. Particularly,



(a) BT Model

(b) XGBL Model



(c) SVMR Model

Figure 6.4: Feature importance analysis from the ML models

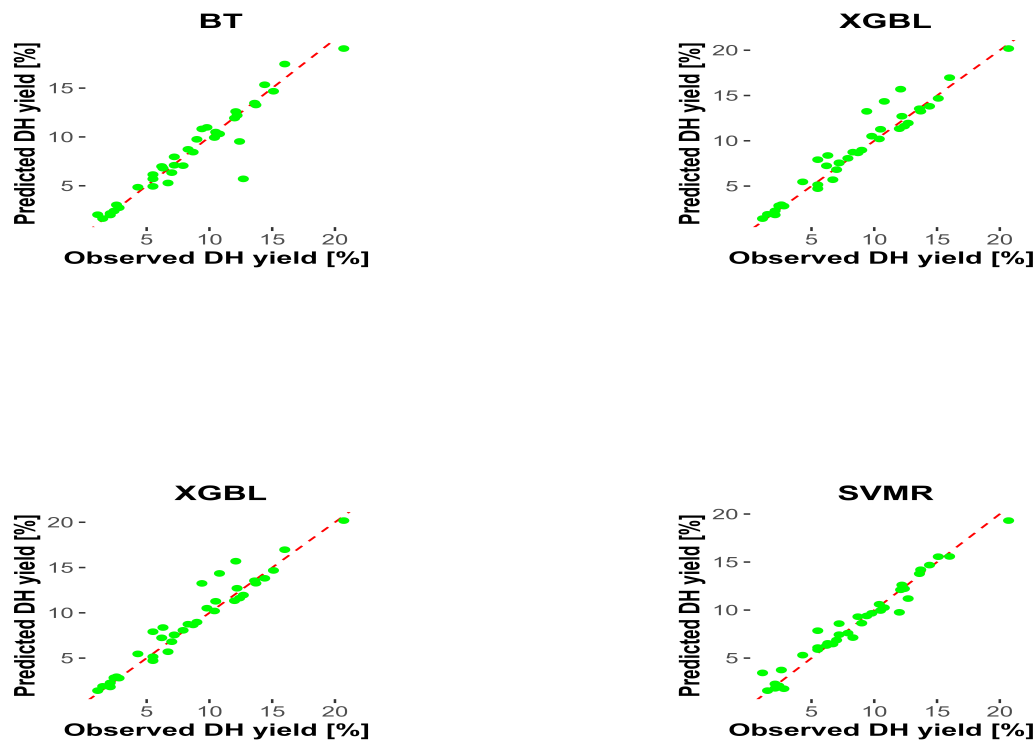


Figure 6.5: Parity plots for BT, XGBL, SVMR and Ensemble Models

the ensemble method emerges as the most effective choice for predicting DH yield, evident in its consistent alignment with the equality line and minimal prediction errors. The practical use of the ensemble method lies in its ability to accurately predict the DH yield, as demonstrated by the close alignment between the predicted and observed values in the parity plots. This robust predictive capability translates into real-world benefits, where process optimization can be achieved more confidently. Accurate predictions can guide researchers and practitioners in setting optimal process conditions and designing tailored catalyst compositions. Consequently, integrating ML models into catalysis research offers a pathway to accelerate the discovery and optimization of high-performance catalysts. The implications of this study extend beyond the laboratory setting to industrial applications. Companies in the chemical and petrochemical

sectors can enhance their catalytic processes by deploying an Ensemble method or similar ML approaches. An improved DH yield not only leads to higher product yields but also minimizes waste and reduces the environmental impact of these processes. Ultimately, this study will contribute to the development of sustainable and eco-friendly industrial practices. It is important to acknowledge the limitations of this study. Although the Ensemble method demonstrated a remarkable predictive performance, the choice of other ML models and tuning approaches may yield different results. Careful consideration of the specific applications and dataset characteristics is essential to ensure that the most suitable model is selected. In addition, the experimental data used in this study were obtained from a controlled fixed-bed reactor. Real-world industrial processes may involve more complex and dynamic conditions that could influence the predictive accuracy of the models. Further validation and refinement of the models using data from pilot- or industrial-scale reactors would enhance their applicability and reliability in practical scenarios.

6.6 Summary

The findings of this study underscore the practical significance of employing ML models, particularly their ensemble technique, in oxidative dehydrogenation (ODH) reactions. The successful application of the ensemble method to predict the DH yield holds immense promise for catalysis research and industrial applications. By accurately predicting the DH yield, this method offers a valuable tool for optimizing the ODH process and catalyst design, with real-world implications for enhancing the pro-

cess efficiency and sustainability. One of the key insights from feature importance analysis is the identification of critical predictor variables that influence DH yield. Notably, Mo, Ni, and W were highlighted as significant factors in determining the DH yield. Understanding the impact of these variables provides essential contextual meaning as it sheds light on the underlying mechanisms governing the ODH reaction. Such insights will enable researchers and process engineers to make informed decisions regarding the design and operation of catalytic systems for maximum DH yield. Moreover, this knowledge contributes to the exploration of novel catalyst materials and development of more efficient and sustainable DH processes.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

The general motivation of this thesis was to develop new random projection techniques for dimensionality reduction in big-data domains where traditional methods suffer from more expensive computing power or information loss. It also explores the use of machine learning algorithms for predicting catalysis performance, with the potential to advance our understanding and application of catalytic processes.

In Chapter 3 We introduce and validate a novel hybrid RP method, referred to as HRP, designed specifically to address the challenges of dimensionality reduction in high-dimensional data. This innovative approach synthesizes the principles of the normal and plus-minus projection methods, effectively balancing computational efficiency with the preservation of critical information.

Our method stands on a foundation of rigorous mathematical proofs, demonstrating

its robustness and reliability. We develop a suite of rigorous mathematical proofs that demonstrate its adherence to the stringent requirements of high-dimensional data processing. These proofs serve as the foundation of our methodology, confirming its ability to preserve the essential geometrical and statistical properties of the data during dimensionality reduction. They provide a mathematical basis for the effectiveness of the method, reinforcing its robustness in maintaining the integrity of post-projection data. These theoretical principles are not only essential for comprehending the mechanics of the HRP method but also play a crucial role in enhancing its practical applicability and reliability in high-dimensional data processing. By adhering to these mathematical proofs, this method can handle the complexities of high-dimensional data. This is of utmost importance in the field of data science, where preserving the pairwise distance of the data is of paramount importance.

Empirical evaluations of the HRP, incorporating both simulation and experimental studies, have further reinforced its effectiveness. Comparative analyses against established RP techniques—Normal and Plus-Minus projections—across four diverse real-world datasets have consistently demonstrated the superiority of the proposed method. This superiority was manifested in its notable ability to maintain the data structure and enhance prediction accuracy in a wide range of scenarios, including medical diagnoses in the lung cancer dataset and complex pattern recognition in the Darwin dataset. These findings underscore the consistent and superior performance of the HRP method across diverse datasets, positioning it as an effective and adaptable solution for high-dimensional data analyses. This method demonstrates a clear advantage over traditional individual methods, showing its potential as a versatile and

reliable solution for dimensionality reduction tasks. The implications of these findings are profound, particularly in fields where high-dimensional data are prevalent. The HRP method offers a promising avenue for researchers and practitioners to process large datasets efficiently while retaining essential information. Its adaptability and effectiveness make it a valuable tool in areas such as bioinformatics, image processing, and any domain in which data dimensionality poses a significant challenge.

In Chapter 4 we extended the HRP to several methods and the integration of the traditional methods proved promising in preserving the pairwise distances than the individual methods. Specifically, the NPLi method, which integrates Normal Random Projection (NRP) and Li methods, demonstrated superior performance in preserving pairwise distances in the original data after projection. These findings encourage further exploration and adoption of hybrid projection techniques in similar contexts, aiming to achieve a balance between efficient dimensionality reduction and high-accuracy data representation. Several promising avenues for further research are warranted. An area of significant interest is the potential applicability of these hybrid methods to other types of datasets and their integration with appropriate machine learning algorithms. Furthermore, the need for further refinement of the method to optimize computational efficiency and information preservation, particularly in rapidly evolving data-driven environments, cannot be overstated. The possibility of customizing the proposed methods to suit specific data characteristics and analytical requirements also presents an exciting opportunity for future development. In this thesis we used average distortions as performance metrics. Another metrics we could use is the median distortion

In 5, we developed unified machine learning models to predict 1,3 butadiene Selectivity and n-butane conversion using data generated from the field bed reactor. Based on the models' prediction analyses, it can be concluded that the Support Vector Machine with Radial base kernel function(SVMR) model is the most suitable for both n-butane conversion and 1,3-butadiene selectivity predictions. Hence, these models can be utilized to design highly active and selective oxidative dehydrogenation catalysts.

In addition to the developed model, this study introduced a digital tool that uses Machine Learning algorithms to provide instantaneous predictions of catalyst performance in the oxidative dehydrogenation of n-butane to 1,3-butadiene. The tool is designed to help researchers, scientists, and industry professionals determine the optimal conditions for catalyst operation, thus improving catalyst design and process efficiency. The application acts as a bridge between high-level ML insights and practical real-world applications, making a substantial contribution to the field of heterogeneous catalysis. To explore the transformative potential of this application, please find the link provided below. [Oxidative Dehydrogenation Catalyst Application](#). This fusion of catalyst informatics and explainable AI aids in simplifying complex heterogeneous catalytic processes, shedding light on the catalyst performance, and opening new horizons for catalyst optimization.

In Chapter 6, this study demonstrates the practical potential of ML models,

particularly the Ensemble method, for catalysis research and industrial applications. The ability to accurately predict DH yield and identify critical predictor variables provides a powerful tool for enhancing process efficiency and optimizing catalyst design, and we continue to explore the synergies between experimental data and ML techniques. The field of catalysis benefits from accelerated catalyst discovery, improved process efficiency, and a more sustainable approach to chemical production. Integrating ML approaches into catalysis research is not just a technological advancement, but a transformative shift towards a greener and more efficient future for industrial processes.

APPENDIX A

SUPPLEMENTARY MATERIAL

FOR HYBRID RANDOM

PROJECTION: INTEGRATING

DENSE AND SPARSE

TECHNIQUES FOR ENHANCED

REPRESENTATION IN

HIGH-DIMENSIONAL DATA

A.1 Supplementary Information for Results on

Simulation

errors. Specifically, Appendix 1 Table 1 represents the time taken for average distortions, while Appendix 1 Table 2 provides insights into the time required for standard errors. Analysis of the tables reveals that the computational time associated with the calculation of average distortions and standard errors by the Hybrid Random Projection (HRP) method surpasses that of Normal Random Projection (NRP) and Plus-Minus Random Projection (PMRP) counterparts. This discrepancy is anticipated, given that HRP integrates NRP and PMRP, introducing a balancing parameter α . Consequently, the time investment is a combination of that needed by NRP, PMRP, and the computation of the optimal α value.

Table A.1: The time taken to calculate each average distortion and the standard error for each method employed for $d = 500$

d	n	PM (mean T.)	PM (sd T.)	NP (mean T.)	NP (sd T.)	HRP (mean T.)	HRP (sd T.)
500	50	9.36×10^{-05}	5.39×10^{-05}	0.000127501	4.03×10^{-05}	0.091702312	0.031750046
500	50	9.36×10^{-05}	4.92×10^{-05}	0.000159845	0.000254208	0.08770341	0.025757058
500	50	8.63×10^{-05}	4.23×10^{-05}	0.000133345	3.78×10^{-05}	0.085617066	0.023471945
500	50	9.33×10^{-05}	5.35×10^{-05}	0.000129493	3.95×10^{-05}	0.086174489	0.02344815
500	50	8.85×10^{-05}	4.60×10^{-05}	0.000128682	3.50×10^{-05}	0.083429923	0.01446132
500	50	9.61×10^{-05}	0.000267297	9.95×10^{-05}	3.39×10^{-05}	0.080163526	0.012526594
500	50	6.61×10^{-05}	4.12×10^{-05}	9.79×10^{-05}	3.22×10^{-05}	0.084257363	0.023486515
500	50	6.10×10^{-05}	2.80×10^{-05}	9.23×10^{-05}	2.73×10^{-05}	0.083405461	0.022181937
500	50	8.76×10^{-05}	0.000263899	8.78×10^{-05}	1.56×10^{-05}	0.084597902	0.024201713
500	50	6.68×10^{-05}	4.75×10^{-05}	9.38×10^{-05}	3.68×10^{-05}	0.084340653	0.026775872
500	50	6.11×10^{-05}	4.77×10^{-05}	9.01×10^{-05}	3.19×10^{-05}	0.083772397	0.020028148
500	50	6.18×10^{-05}	4.07×10^{-05}	8.73×10^{-05}	2.63×10^{-05}	0.08236704	0.019196744
500	50	6.29×10^{-05}	4.52×10^{-05}	8.31×10^{-05}	1.80×10^{-05}	0.078909888	0.013393739
500	50	6.23×10^{-05}	4.20×10^{-05}	8.57×10^{-05}	2.95×10^{-05}	0.080731084	0.016085115
500	50	5.46×10^{-05}	3.39×10^{-05}	8.52×10^{-05}	2.74×10^{-05}	0.080189599	0.01601233
500	200	8.36×10^{-05}	4.49×10^{-05}	0.000125299	3.79×10^{-05}	0.177605233	0.132950319
500	200	8.45×10^{-05}	4.70×10^{-05}	0.000127925	4.03×10^{-05}	0.176983374	0.130262781
500	200	7.99×10^{-05}	4.49×10^{-05}	0.000124664	3.49×10^{-05}	0.180057311	0.125237331
500	200	7.86×10^{-05}	4.08×10^{-05}	0.00012361	4.27×10^{-05}	0.167453382	0.110415874
500	200	7.94×10^{-05}	4.13×10^{-05}	0.000120988	3.82×10^{-05}	0.194005911	0.13484981
500	200	8.36×10^{-05}	0.000247289	8.92×10^{-05}	2.88×10^{-05}	0.170089317	0.104791575
500	200	6.31×10^{-05}	4.65×10^{-05}	9.44×10^{-05}	7.81×10^{-05}	0.16747815	0.114552282
500	200	9.07×10^{-05}	0.000255777	9.42×10^{-05}	4.57×10^{-05}	0.166884219	0.118319962
500	200	6.64×10^{-05}	5.20×10^{-05}	9.22×10^{-05}	4.14×10^{-05}	0.154756369	0.088775257

Table A.2: The time taken to calculate each average distortion and the standard error for each method employed for $d = 10000$

d	n	PM (sd Time)	NP (mean Time)	NP (sd Time)	HRP (mean Time)	HRP (sd Time)
10000	50	5.39×10^{-05}	0.000127501	4.03×10^{-05}	0.091702312	0.0317500
10000	50	4.92×10^{-05}	0.000159845	0.000254208	0.08770341	0.0257570
10000	50	4.23×10^{-05}	0.000133345	3.78×10^{-05}	0.085617066	0.0234719
10000	50	5.35×10^{-05}	0.000129493	3.95×10^{-05}	0.086174489	0.0234483
10000	50	4.60×10^{-05}	0.000128682	3.50×10^{-05}	0.083429923	0.0144613
10000	50	0.000267297	9.95×10^{-05}	3.39×10^{-05}	0.080163526	0.0125265
10000	50	4.12×10^{-05}	9.79×10^{-05}	3.22×10^{-05}	0.084257363	0.0234865
10000	50	2.80×10^{-05}	9.23×10^{-05}	2.73×10^{-05}	0.083405461	0.0221819
10000	50	0.000263899	8.78×10^{-05}	1.56×10^{-05}	0.084597902	0.0242017
10000	50	4.75×10^{-05}	9.38×10^{-05}	3.68×10^{-05}	0.084340653	0.0267758
10000	50	4.77×10^{-05}	9.01×10^{-05}	3.19×10^{-05}	0.083772397	0.0200281
10000	50	4.07×10^{-05}	8.73×10^{-05}	2.63×10^{-05}	0.08236704	0.0191967
10000	50	4.52×10^{-05}	8.31×10^{-05}	1.80×10^{-05}	0.078909888	0.0133937
10000	50	4.20×10^{-05}	8.57×10^{-05}	2.95×10^{-05}	0.080731084	0.0160851
10000	50	3.39×10^{-05}	8.52×10^{-05}	2.74×10^{-05}	0.080189599	0.0160123

Appendix 1 Tables 3 to 6 provide the optimal alpha values and their corresponding standard deviations for all data scenarios considered in the simulation of the study. As mentioned earlier, the parameter α acts as the balancing factor between the two classical RP methods (NRP and PMRP) integrated to form our

Hybrid (HRP). Specifically, Tables S3, S4, S5, and S6 present the results for feature dimensions (d) of 500, 1000, 5000, and 10000, respectively. Across all scenarios, it is evident from the tables that the optimal (α) consistently centers around 0.5.

Table A.3: Optimal Alpha and Standard Deviation for Optimal Alpha for The HRP when $d = 500$

d	n	ϵ	δ	Optimal α (HRP)	SD for Optimal α
500	50	0.6	0.2	0.500008934	0.00460995
500	50	0.6	0.4	0.499672827	0.00488614
500	50	0.6	0.6	0.49968619	0.004408066
500	50	0.6	0.8	0.499944952	0.004757491
500	50	0.6	1	0.500118518	0.004798936
500	50	0.8	0.2	0.49927205	0.005481304
500	50	0.8	0.4	0.500096338	0.005207829
500	50	0.8	0.6	0.501169826	0.005706827
500	50	0.8	0.8	0.499906145	0.005132636
500	50	0.8	1	0.500096834	0.005699541
500	50	0.9	0.2	0.49950713	0.006167703
500	50	0.9	0.4	0.500690128	0.006313263
500	50	0.9	0.6	0.500041653	0.006084673
500	50	0.9	0.8	0.499540068	0.005565768
500	50	0.9	1	0.499680544	0.005527749
500	200	0.6	0.2	0.500112796	0.002640122
500	200	0.6	0.4	0.500134763	0.002453553
500	200	0.6	0.6	0.50025293	0.002535655
500	200	0.6	0.8	0.499395501	0.002690018
500	200	0.6	1	0.500094568	0.002818837
500	200	0.8	0.2	0.499222424	0.002990062
500	200	0.8	0.4	0.499774669	0.003060077
500	200	0.8	0.6	0.500404481	0.002801882

Table A.4: Optimal α and Standard Deviation for Optimal α for The HRP when $d = 1000$

d	n	ϵ	δ	Optimal α	SD for Optimal α
1000	50	0.6	0.2	0.499556462	0.004398651
1000	50	0.6	0.4	0.50085064	0.004537259
1000	50	0.6	0.6	0.499721326	0.004578293
1000	50	0.6	0.8	0.499292493	0.004373736
1000	50	0.6	1	0.499627005	0.003893676
1000	50	0.8	0.2	0.499772757	0.005085913
1000	50	0.8	0.4	0.500514105	0.005186971
1000	50	0.8	0.6	0.499546718	0.005399896
1000	50	0.8	0.8	0.500221578	0.005183111
1000	50	0.8	1	0.49976058	0.005471885
1000	50	0.9	0.2	0.500167525	0.005414758
1000	50	0.9	0.4	0.500328331	0.005194155
1000	50	0.9	0.6	0.499329585	0.005427859
1000	50	0.9	0.8	0.499794233	0.006225498
1000	50	0.9	1	0.49939472	0.005812735
1000	200	0.6	0.2	0.499991388	0.002167272
1000	200	0.6	0.4	0.499729549	0.00240404
1000	200	0.6	0.6	0.499733321	0.002152596
1000	200	0.6	0.8	0.499539565	0.00216842
1000	200	0.6	1	0.49972792	0.002504867
1000	200	0.8	0.2	0.499833229	0.002800192
1000	200	0.8	0.4	0.500335584	0.002628539

Table A.5: Optimal α and Standard Deviation for Optimal α The HRP when $d = 5000$

d	n	ϵ	δ	Optimal α	SD for Optimal α
5000	50	0.6	0.2	0.500025706	0.003982054
5000	50	0.6	0.4	0.499809176	0.004176599
5000	50	0.6	0.6	0.499787665	0.004310347
5000	50	0.6	0.8	0.499534731	0.003812847
5000	50	0.6	1	0.500163496	0.004117222
5000	50	0.8	0.2	0.499601184	0.004618396
5000	50	0.8	0.4	0.500274277	0.005072595
5000	50	0.8	0.6	0.500172655	0.004275443
5000	50	0.8	0.8	0.500927659	0.004688799
5000	50	0.8	1	0.50003265	0.004750703
5000	50	0.9	0.2	0.500605774	0.004897352
5000	50	0.9	0.4	0.499173793	0.00522993
5000	50	0.9	0.6	0.500382498	0.004488801
5000	50	0.9	0.8	0.499460206	0.005113604
5000	50	0.9	1	0.500021806	0.005248271
5000	200	0.6	0.2	0.500209811	0.001923964
5000	200	0.6	0.4	0.500087773	0.001747625
5000	200	0.6	0.6	0.499925209	0.001969511
5000	200	0.6	0.8	0.500079424	0.002092256
5000	200	0.6	1	0.500016689	0.002067915
5000	200	0.8	0.2	0.499887596	0.002456267
5000	200	0.8	0.4	0.499674248	0.002641183
5000	200	0.8	0.6	0.500232877	0.002670001

Table A.6: Optimal α and Standard Deviation for Optimal α for The HRP when $d = 10000$

d	n	ϵ	δ	Optimal α	SD for Optimal α
10000	50	0.6	0.2	0.500144936	0.003612293
10000	50	0.6	0.4	0.49913785	0.003629651
10000	50	0.6	0.6	0.500253003	0.004011213
10000	50	0.6	0.8	0.500493301	0.004010798
10000	50	0.6	1	0.500021279	0.003438563
10000	50	0.8	0.2	0.500303717	0.004119271
10000	50	0.8	0.4	0.500049923	0.004585036
10000	50	0.8	0.6	0.500503611	0.004715584
10000	50	0.8	0.8	0.500573893	0.004488758
10000	50	0.8	1	0.499926035	0.004242669
10000	50	0.9	0.2	0.500594114	0.004067884
10000	50	0.9	0.4	0.499658963	0.004906896
10000	50	0.9	0.6	0.499833456	0.00458627
10000	50	0.9	0.8	0.499491258	0.003985265
10000	50	0.9	1	0.499724676	0.004627034

Appendix 1 Figure 1 displays a box plot illustrating the average distortions across 100 repetitions for the three methods (PMRP, NRP, and HRP). This plot provides a clear visual representation of the distribution of average distortions, facilitating easier comparison between PMRP, NRP, and HRP.

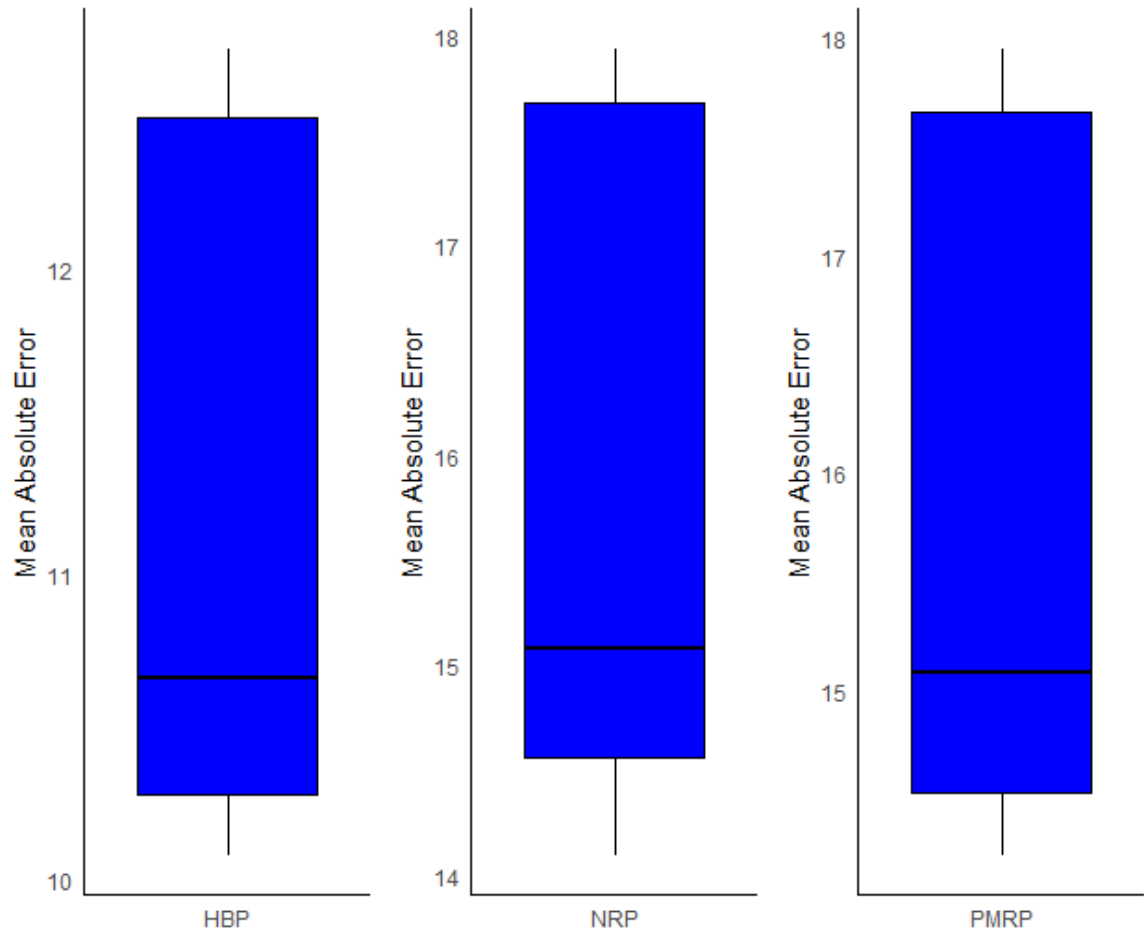


Figure A.1: average distortion Box Plot for PMRP,NRP and HRP methods

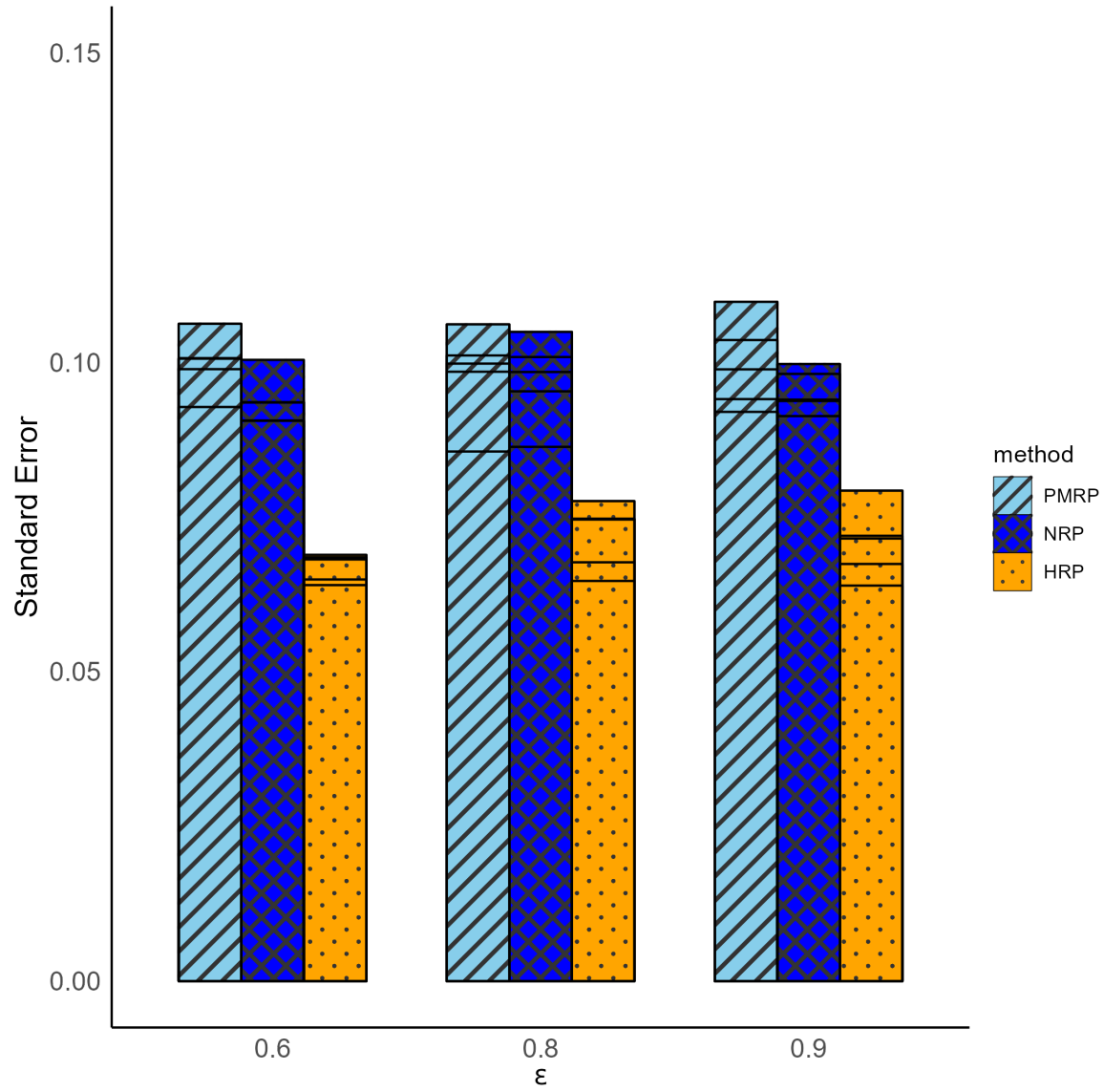


Figure A.2: Comparison of Standard Errors Across Different Methods of Random Projection for $d=5000$

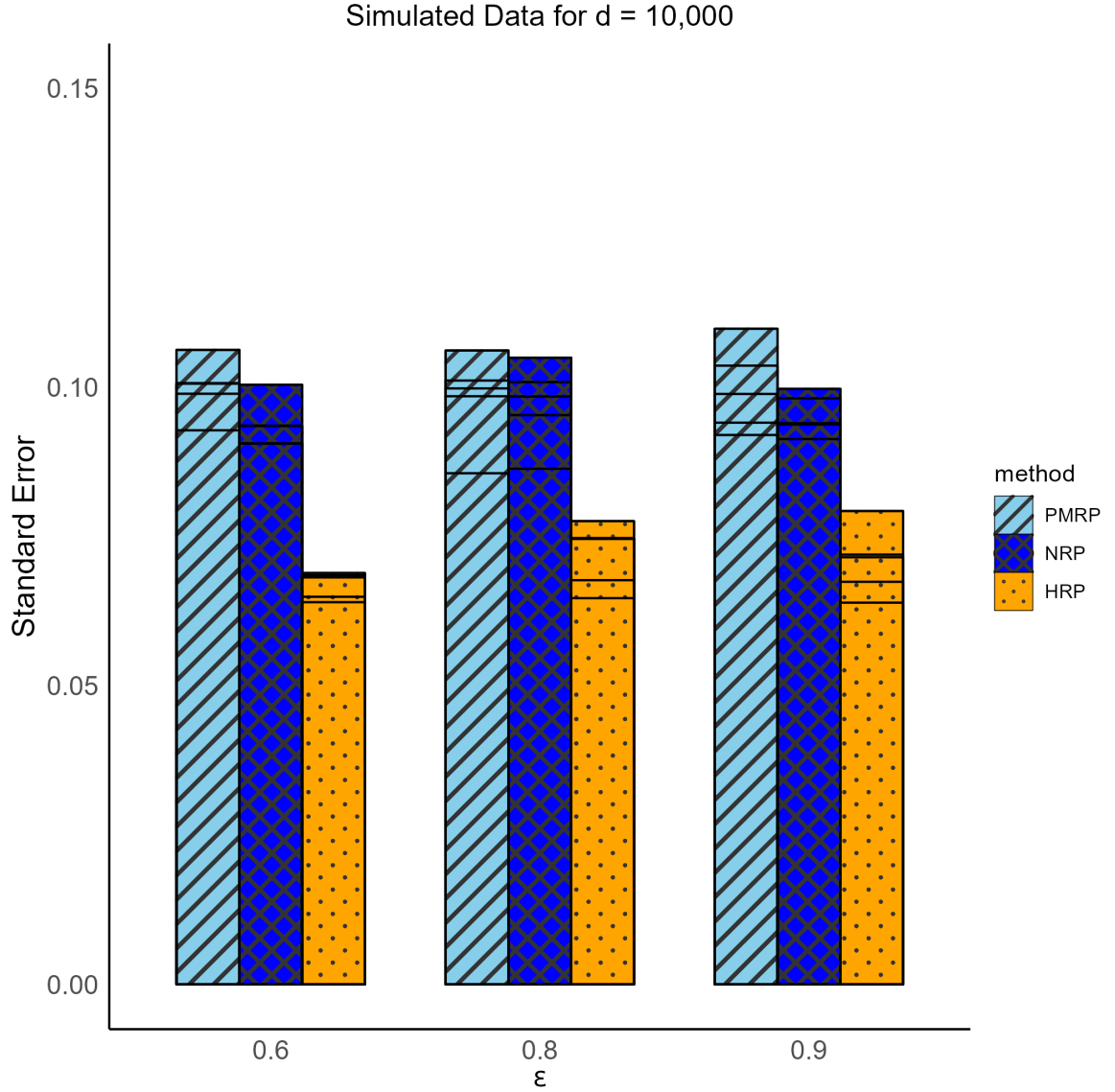


Figure A.3: Comparison of Standard Errors Across Different Methods of Random Projection for $d=10000$

Appendix 1 Figures 1 and 2 present the standard deviation around the mean across different random projection methods (PMRP, NRP, and HRP) at various epsilon parameter levels applied to our simulated data for feature dimensions $d = 5000$ and $d = 10000$, respectively. Each method is distinguished by distinct bar colors for clarity. In these figures, each bar corresponds to a specific epsilon level,

and its height reflects the standard deviation, indicating the variability around the average distortion.

A discernible trend in the standard error is apparent for all methods as the epsilon parameter increases. It is noteworthy that there exists a direct relationship between epsilon and standard error, signifying that an increase in the epsilon parameter corresponds to an increase in the standard error of the mean. HRP exhibits slightly higher values for standard errors compared to NRP and PMRP. Nevertheless, overall, all methods display minimal standard errors, suggesting limited variability and greater stability in average distortions across all methods.

REFERENCES

- [1] P. Li, T. J. Hastie, and K. W. Church, “Very sparse random projections,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 287–296.
- [2] D. Achlioptas, “Database-friendly random projections: Johnson-lindenstrauss with binary coins,” *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.
- [3] Global 1,3-butadiene market. Accessed on 5th August, 2022. [Online]. Available: <https://databridgemarketresearch.com/reports/global-1,3-butadiene-market>
- [4] K. McCullough, T. Williams, K. Mingle, P. Jamshidi, and J. Lauterbach, “High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery,” *Physical Chemistry Chemical Physics*, vol. 22, no. 20, pp. 11 174–11 196, 2020.
- [5] A. Medford, M. Kunz, S. Ewing, T. Borders, and R. Fushimi, “Extracting knowledge from data through catalysis informatics. *acs catal* 8 (8): 7403–7429,” 2018.

- [6] T. Mo, L. Wang, Y. Wu, J. Huang, W. Liu, R. Yang, and X. Zhen, “Classifier ensemble with evolutionary optimisation enforced random projections,” *Expert Systems with Applications*, vol. 222, p. 119845, 2023.
- [7] B. Denham, R. Pears, and M. A. Naeem, “Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining,” *Expert Systems with Applications*, vol. 152, p. 113380, 2020.
- [8] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 245–250.
- [9] W. B. Johnson, “Extensions of lipschitz mappings into hilbert space,” *Contemporary mathematics*, vol. 26, pp. 189–206, 1984.
- [10] P. Frankl and H. Maehara, “The johnson-lindenstrauss lemma and the sphericity of some graphs,” *Journal of Combinatorial Theory, Series B*, vol. 44, no. 3, pp. 355–362, 1988.
- [11] S. Dasgupta and A. Gupta, “An elementary proof of a theorem of johnson and lindenstrauss,” *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.
- [12] A. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.
- [13] T. Mitchell, *Machine Learning*. McGraw-Hill Education (ISE editions), 1997.

- [14] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, “Fast and accurate modeling of molecular atomization energies with machine learning,” *Physical Review Letters*, vol. 108, no. 5, p. 058301, 2012.
- [15] W. Yang, T. T. Fidelis, and W.-H. Sun, “Machine learning in catalysis, from proposal to practicing,” *ACS omega*, vol. 5, no. 1, pp. 83–88, 2019.
- [16] P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, and T. Bligaard, “Machine learning for computational heterogeneous catalysis,” *ChemCatChem*, vol. 11, no. 16, pp. 3581–3601, 2019.
- [17] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-i. Shimizu, “Machine learning for catalysis informatics: recent applications and prospects,” *Acs Catalysis*, vol. 10, no. 3, pp. 2260–2297, 2019.
- [18] M. Erdem Günay and R. Yıldırım, “Recent advances in knowledge discovery for heterogeneous catalysis using machine learning,” *Catalysis Reviews*, vol. 63, no. 1, pp. 120–164, 2021.
- [19] R. Monreal-Corona, A. Pla-Quintana, and A. Poater, “Predictive catalysis: A valuable step towards machine learning,” *Trends in Chemistry*, 2023.
- [20] H. Li, Z. Zhang, and Z. Liu, “Application of artificial neural networks for catalysis: a review,” *Catalysts*, vol. 7, no. 10, p. 306, 2017.
- [21] A. Callahan and N. H. Shah, “Machine learning in healthcare,” in *Key advances in clinical informatics*. Elsevier, 2017, pp. 279–291.

- [22] H. Habehh and S. Gohel, “Machine learning in healthcare,” *Current Genomics*, vol. 22, no. 4, p. 291, 2021.
- [23] K. Shailaja, B. Seetharamulu, and M. Jabbar, “Machine learning in healthcare: A review,” in *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2018, pp. 910–914.
- [24] A. Nayyar, L. Gadhavi, and N. Zaman, “Machine learning in healthcare: review, opportunities and challenges,” *Machine Learning and the Internet of Medical Things in Healthcare*, pp. 23–45, 2021.
- [25] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, 2022.
- [26] T. J. Saleem and M. A. Chishti, “Exploring the applications of machine learning in healthcare,” *International Journal of Sensors Wireless Communications and Control*, vol. 10, no. 4, pp. 458–472, 2020.
- [27] D. Sathya, V. Sudha, and D. Jagadeesan, “Application of machine learning techniques in healthcare,” in *Handbook of Research on Applications and Implementations of Machine Learning Techniques*. IGI Global, 2020, pp. 289–304.
- [28] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.
- [29] Y. Zhang and J. C. Rajapakse, *Machine learning in bioinformatics*. John Wiley & Sons, 2009.

- [30] H. Bhaskar, D. C. Hoyle, and S. Singh, “Machine learning in bioinformatics: A brief survey and recommendations for practitioners,” *Computers in biology and medicine*, vol. 36, no. 10, pp. 1104–1125, 2006.
- [31] K. A. Shastry and H. Sanjay, “Machine learning for bioinformatics,” *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*, pp. 25–39, 2020.
- [32] I. Inza, B. Calvo, R. Armananzas, E. Bengoetxea, P. Larranaga, and J. A. Lozano, “Machine learning: an indispensable tool in bioinformatics,” in *Bioinformatics methods in clinical research*. Springer, 2009, pp. 25–48.
- [33] Z. R. Yang, *Machine learning approaches to bioinformatics*. World scientific, 2010, vol. 4.
- [34] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” *Sensors*, vol. 18, no. 8, p. 2674, 2018.
- [35] G. Sambasivam and G. D. Opiyo, “A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks,” *Egyptian informatics journal*, vol. 22, no. 1, pp. 27–34, 2021.
- [36] V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. Ramkteke, “Machine learning in agriculture domain: A state-of-art survey,” *Artificial Intelligence in the Life Sciences*, vol. 1, p. 100010, 2021.

- [37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [38] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [39] S. Deegalla, K. Walgama, P. Papapetrou, and H. Boström, “Random subspace and random projection nearest neighbor ensembles for high dimensional data,” *Expert systems with applications*, vol. 191, p. 116078, 2022.
- [40] R. Siddharth and G. Aghila, “Randpro-a practical implementation of random projection-based feature extraction for high dimensional multivariate data analysis in r,” *SoftwareX*, vol. 12, p. 100629, 2020.
- [41] W. B. Johnson, “Extensions of lipshitz mapping into hilbert space,” in *Conference modern analysis and probability, 1984*, 1984, pp. 189–206.
- [42] C.-H. Yang, L.-Y. Chuang, and Y.-D. Lin, “Multiobjective multifactor dimensionality reduction to detect snp–snp interactions,” *Bioinformatics*, vol. 34, no. 13, pp. 2228–2236, 2018.
- [43] C. Grellmann, J. Neumann, S. Bitzer, P. Kovacs, A. Tönjes, L. T. Westlye, O. A. Andreassen, M. Stumvoll, A. Villringer, and A. Horstmann, “Random projection for fast and efficient multivariate correlation analysis of high-dimensional data: A new approach,” *Frontiers in Genetics*, vol. 7, p. 102, 2016.

- [44] N. Ailon, B. Chazelle, S. Comandur, and D. Liu, “Estimating the distance to a monotone function,” *Random Structures & Algorithms*, vol. 31, no. 3, pp. 371–383, 2007.
- [45] F. Galton, “Regression towards mediocrity in hereditary stature,” *Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246–263, 1886.
- [46] A. N. Tikhonov, “On the stability of inverse problems,” *Doklady Akademii Nauk SSSR*, 1943.
- [47] C. K. Williams and C. E. Rasmussen, “Gaussian processes for regression,” in *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, 1996, pp. 514–520.
- [48] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [49] C. Eyupoglu, “Breast cancer classification using k-nearest neighbors algorithm,” *The Online Journal of Science and Technology*, vol. 8, no. 3, pp. 29–34, 2018.
- [50] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [51] E. Fix and J. L. Hodges, “Discriminatory analysis, nonparametric discrimination: Consistency properties,” *USAF School of Aviation Medicine*, 1951.
- [52] J. Chen and A. K. Gupta, “Parametric statistical change point analysis: with applications to genetics, medicine, and finance,” 2012.

- [53] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92)*, 1992, pp. 144–152.
- [54] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [55] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
- [56] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, and Y. Xiang, “Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china,” *Energy conversion and management*, vol. 164, pp. 102–111, 2018.
- [57] B. Boehmke and B. M. Greenwell, *Hands-on machine learning with R*. CRC press, 2019.
- [58] D. H. Wolpert, “The lack of a priori distinctions between learning algorithms,” *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.
- [59] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2012, vol. 792.
- [60] I. Tougui, A. Jilbab, and J. El Mhamdi, “Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications,” *Healthcare informatics research*, vol. 27, no. 3, pp. 189–199, 2021.

- [61] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, “Prediction error estimation: a comparison of resampling methods,” *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.
- [62] D. M. Hawkins, S. C. Basak, and D. Mills, “Assessing model fit by cross-validation,” *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 579–586, 2003.
- [63] J.-H. Kim, “Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap,” *Computational statistics & data analysis*, vol. 53, no. 11, pp. 3735–3745, 2009.
- [64] S. Mine, Y. Jing, T. Mukaiyama, M. Takao, Z. Maeno, K.-i. Shimizu, I. Takigawa, and T. Toyao, “Machine learning analysis of literature data on the water gas shift reaction toward extrapolative prediction of novel catalysts,” *Chemistry Letters*, vol. 51, no. 3, pp. 269–273, 2022.
- [65] K. Suzuki, T. Toyao, Z. Maeno, S. Takakusagi, K.-i. Shimizu, and I. Takigawa, “Statistical analysis and discovery of heterogeneous catalysts based on machine learning from diverse published data,” *ChemCatChem*, vol. 11, no. 18, pp. 4537–4547, 2019.
- [66] S. Nishimura, J. Ohyama, T. Kinoshita, S. Dinh Le, and K. Takahashi, “Revisiting machine learning predictions for oxidative coupling of methane (ocm) based on literature data,” *ChemCatChem*, vol. 12, no. 23, pp. 5888–5892, 2020.

- [67] J. Ohyama, S. Nishimura, and K. Takahashi, “Data driven determination of reaction conditions in oxidative coupling of methane via machine learning,” *ChemCatChem*, vol. 11, no. 17, pp. 4307–4313, 2019.
- [68] K. Takahashi, I. Miyazato, S. Nishimura, and J. Ohyama, “Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data,” *ChemCatChem*, vol. 10, no. 15, pp. 3223–3228, 2018.
- [69] N. Madaan, N. R. Shiju, and G. Rothenberg, “Predicting the performance of oxidation catalysts using descriptor models,” *Catalysis Science & Technology*, vol. 6, no. 1, pp. 125–133, 2016.
- [70] R. I. Arriaga and S. Vempala, “An algorithmic theory of learning: Robust concepts and random projection,” *Machine learning*, vol. 63, pp. 161–182, 2006.
- [71] D. Achlioptas, “Database-friendly random projections,” in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001, pp. 274–281.
- [72] X. Z. Fern and C. E. Brodley, “Random projection for high dimensional data clustering: A cluster ensemble approach,” in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 186–193.
- [73] D. Fradkin and D. Madigan, “Experiments with random projections for machine learning,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 517–522.

- [74] N. Goel, G. Bebis, and A. Nefian, “Face recognition experiments with random projection,” in *Biometric technology for human identification II*, vol. 5779. SPIE, 2005, pp. 426–437.
- [75] R. O. Duda, P. E. Hart *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.
- [76] J. Buhler and M. Tompa, “Finding motifs using random projections,” in *Proceedings of the fifth annual international conference on Computational biology*, 2001, pp. 69–76.
- [77] H. C. Leung, F. Y. Chin, S.-M. Yiu, R. Rosenfeld, and W. W. Tsang, “Finding motifs with insufficient number of strong binding sites,” *Journal of Computational Biology*, vol. 12, no. 6, pp. 686–701, 2005.
- [78] M. Radermacher, “Three-dimensional reconstruction from random projections: orientational alignment via radon transforms,” *Ultramicroscopy*, vol. 53, no. 2, pp. 121–136, 1994. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0304399194900035>
- [79] M. Charikar, K. Chen, and M. Farach-Colton, “Finding frequent items in data streams,” in *International Colloquium on Automata, Languages, and Programming*. Springer, 2002, pp. 693–703.
- [80] . Gül and F. RAHIM, “Period Changer,” UCI Machine Learning Repository, 2022, DOI: <https://doi.org/10.24432/C5B31D>.

- [81] —, “Toxicity,” UCI Machine Learning Repository, 2022, DOI: <https://doi.org/10.24432/C59313>.
- [82] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz *et al.*, “A public domain dataset for human activity recognition using smartphones.” in *Esann*, vol. 3, 2013, p. 3.
- [83] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [84] J. D. Burrington and R. K. Grasselli, “Aspects of selective oxidation and ammoxidation mechanisms over bismuth molybdate catalysts,” *Journal of Catalysis*, vol. 59, no. 1, pp. 79–99, 1979.
- [85] M. Bender, “An overview of industrial processes for the production of olefins–c4 hydrocarbons,” *ChemBioEng Reviews*, vol. 1, no. 4, pp. 136–147, 2014.
- [86] B. Ajayi, B. R. Jermy, K. Ogunronbi, B. Abussaud, and S. Al-Khattaf, “n-butane dehydrogenation over mono and bimetallic mcm-41 catalysts under oxygen free atmosphere,” *Catalysis today*, vol. 204, pp. 189–196, 2013.
- [87] J. C. Jung, H. Kim, Y. S. Kim, Y.-M. Chung, T. J. Kim, S. J. Lee, S.-H. Oh, and I. K. Song, “Catalytic performance of bismuth molybdate catalysts in the oxidative dehydrogenation of c4 raffinate-3 to 1, 3-butadiene,” *Applied Catalysis A: General*, vol. 317, no. 2, pp. 244–249, 2007.

- [88] V. Murgia, E. M. F. Torres, J. C. Gottifredi, and E. L. Sham, "Sol-gel synthesis of $\text{V}_2\text{O}_5\text{-SiO}_2$ catalyst in the oxidative dehydrogenation of n-butane," *Applied Catalysis A: General*, vol. 312, pp. 134–143, 2006.
- [89] I. Rossetti, G. F. Mancini, P. Ghigna, M. Scavini, M. Piumetti, B. Bonelli, F. Cavani, and A. Comite, "Spectroscopic enlightening of the local structure of VOx active sites in catalysts for the odh of propane," *The Journal of Physical Chemistry C*, vol. 116, no. 42, pp. 22 386–22 398, 2012.
- [90] B. R. Jermy, B. Ajayi, B. Abussaud, S. Asaoka, and S. Al-Khattaf, "Oxidative dehydrogenation of n-butane to butadiene over $\text{Bi-Ni-O}/\gamma\text{-alumina}$ catalyst," *Journal of Molecular Catalysis A: Chemical*, vol. 400, pp. 121–131, 2015.
- [91] B. R. Jermy, S. Asaoka, and S. Al-Khattaf, "Influence of calcination on performance of $\text{Bi-Ni-O}/\gamma\text{-alumina}$ catalyst for n-butane oxidative dehydrogenation to butadiene," *Catalysis Science & Technology*, vol. 5, no. 9, pp. 4622–4635, 2015.
- [92] G. Tanimu, B. Jermy, S. Asaoka, and S. Al-Khattaf, "Composition effect of metal species in $(\text{Ni, Fe, Co})\text{-Bi-O}/\gamma\text{-Al}_2\text{O}_3$ catalyst on oxidative dehydrogenation of n-butane to butadiene," *Journal of Industrial and Engineering Chemistry*, vol. 45, pp. 111–120, 2017.
- [93] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai *et al.*, "The rise of catalyst informatics: towards catalyst genomics," *ChemCatChem*, vol. 11, no. 4, pp. 1146–1152, 2019.

- [94] L. M. Madeira and M. F. Portela, “Catalytic oxidative dehydrogenation of n-butane,” *Catalysis Reviews*, vol. 44, no. 2, pp. 247–286, 2002.
- [95] G. Tanimu, S. Asaoka, and S. Al-Khattaf, “Effect of support in ni-bi-o/support catalyst on oxidative dehydrogenation of n-butane to butadiene,” *Molecular Catalysis*, vol. 438, pp. 245–255, 2017.
- [96] J. F. MacGregor, “Using on-line process data to improve quality: challenges for statisticians,” *International Statistical Review*, vol. 65, no. 3, pp. 309–323, 1997.
- [97] G. Tanimu, J. O. Ajadi, Y. Yahaya, H. Alasiri, and N. A. Adegoke, “Developing machine learning models for catalysts in oxidative dehydrogenation of n-butane,” *ChemCatChem*, 2023.
- [98] Y. Gambo, S. Adamu, G. Tanimu, I. M. Abdullahi, R. A. Lucky, M. S. Ba-Shammakh, and M. M. Hossain, “Co₂-mediated oxidative dehydrogenation of light alkanes to olefins: Advances and perspectives in catalyst design and process improvement,” *Applied Catalysis A: General*, vol. 623, p. 118273, 2021.
- [99] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *Journal of animal ecology*, vol. 77, no. 4, pp. 802–813, 2008.
- [100] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [101] M. Kuhn, “Building predictive models in r using the caret package,” *Journal of statistical software*, vol. 28, pp. 1–26, 2008.

- [102] R. D. C. Team, “R: A language and environment for statistical computing,” (*No Title*), 2010.

VITAE

- Name: Yussif Yahaya
- Nationality: Ghanaian.
- Email: *yahayayussif47@gmail.com*
- Permenant Address: Street E Plot 88, Buobai-Kumasi, Ashanti 00233 Ghana

Work History

Graduate Assistant

King Fahd University of Petroleum and Minerals 08/2023 - Current

Location: Dhahran, Saudi Arabia

Responsibilities: Laboratory instructor for STAT 319, teaching undergraduate students R programming for data analysis.

Graduate Assistant

King Fahd University of Petroleum and Minerals 01/2023 - 05/2023

Location: Dhahran, Saudi Arabia

Responsibilities: Python lab instructor for 60 undergraduate students across three sections, teaching and guiding them on using Python programming for mathematical expressions in Calculus II.

Graduate Assistant

King Fahd University of Petroleum and Minerals 08/2022 - 12/2022

Location: Dhahran, Saudi Arabia

Responsibilities: MATLAB lab instructor for 69 undergraduate students across three sections, teaching and guiding them on using MATLAB for mathematical expressions in Calculus I.

Graduate Assistant

King Fahd University of Petroleum and Minerals 01/2022 - 05/2022

Location: Dhahran, Saudi Arabia

Responsibilities: Python lab instructor for 48 undergraduate students across two sections, teaching and guiding them on using Python for mathematical expressions in Calculus II.

Data Officer

Asokore Mampong Municipal Health Directorate 10/2018 - 09/2021

Location: Kumasi, Ghana

Responsibilities: Managed processes and information in a health information management software (DHIMS) that provided regional and national authorities with a clear view of the municipality's health situation, leading to quick decision-making.

National Service Person

Asokore Mampong Municipal Health Directorate 09/2017 - 07/2018

Location: Kumasi, Ghana

Responsibilities:

- Prepared and processed health insurance claims of health facilities within the municipality.
- Assisted Head of finance in budget and report preparation.
- Organized financial and operational data to help with yearly budgeting and planning.

Education

MSc. Applied Statistics

King Fahd University of Petroleum and Minerals 01/2022 - Current

Location: Dhahran, Saudi Arabia

Image Processing Training Workshop

King Fahd University of Petroleum and Minerals 03/2023

Location: Dhahran, Saudi Arabia

Bachelor of Science in Actuarial Science

University for Development Studies (UDS) 09/2013 - 05/2017

Location: Tamale, Ghana

West Africa Senior School Certificate

Adventist Senior High School 09/2007 - 05/2011

Network/Civil Society Engagement

Serve Humanity Foundation Ghana

06/01/18 - 09/01/21

Responsibilities: Served as youth organizer responsible for recruiting youth from various deprived communities to be trained in various sustainable skills and empowerment projects of the organization.

Actuarial Science Students Association of Ghana (ASSAG) 04/01/15
- 04/01/16

Responsibilities: Served as chairman of the electoral commissioner responsible for conducting all elections of the association.

Ghana Red Cross Society (UDS Chapter)

04/01/15 - 04/01/16

Responsibilities: Served as President of the society responsible for leading and directing activities of the society at the university level.

Research Experiences and Publications

Current Project Title: “Hybrid Random Projection: Integrating Dense and Sparse Techniques for Enhanced Representation in High-Dimensional Data.”
(Under Review)

Status: Submitted

Current Project Title: “Machine Learning Advancements in Catalysis: An Ensemble Model Approach to Predict Dehydrogenation Yield” (Under Review)

Status: Submitted

Publication Title: "Developing Machine Learning Models for Catalysts in Oxidative Dehydrogenation of n-butane"

Authors: Dr. Gazali Tanimu, Dr. Jimoh Olawale Ajadi, Yussif Yahaya, Dr. Hassan Alasiri, Dr. Nurudeen A. Adegoke.

Journal Name: ChemCatChem

Publication Date: July 2023

DOI or Link: <https://doi.org/10.1002/cctc.202300598>

Research Interests Area

Machine learning, Data Science, Experimental Statistics, Health Informatics, Bio-informatics

References

References available upon request