2021

Robust statistical learning for optimal classification of imbalanced data.

Juma, Samuel Wanyonyi
Strathmore Institute of Mathematical Sciences
Strathmore University

Robust Statistical Learning for Optimal Classification of Imbalanced Data

Samuel Wanyonyi Juma

A Research Thesis Submitted to the Graduate School in partial fulfillment of the requirements for the Award of Master of Science Degree in Statistical Sciences at Strathmore University

Strathmore Institute of Mathematical Sciences (SIMS)

STRATHMORE UNIVERSITY

SEPTEMBER, 2021

NAIROBI, KENYA

DECLARATION AND APPROVAL

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Samuel Wanyonyi Juma

Signature: 

Date: 14^{th} September, 2021

APPROVAL

The thesis of Samuel Wanyonyi Juma was reviewed and approved by the following:

Dr. Ojwang Collins Odhiambo,
Senior Lecturer, Strathmore Institute of Mathematical Sciences (SIMS),
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

ABSTRACT

Neurobiological disorders such as Learning Disabilities (LD) are increasing becoming a major concern in education and health sectors, hence, precise identification of these disorders is critical. While neuropsychological assessments play an important role in diagnosis, there is limited conventional methodologies for test administration, scoring and interpretation of results. Consequently, there is frequent misclassification of children due to imprecise distinction between children with learning *disabilities* and those with learning *difficulties*. This research sought to apply statistical and Machine Learning (ML) approaches to strengthen the LD diagnostic process. This research addresses the challenges of learning from imbalanced data, a characteristic often associated with LD data due to low prevalence of the disorder. Imbalanced data poses a challenge in designing efficient ML solutions since standard classification models assumes fairly distributed classes. The study used experimental design to identify a suitable base learner, and corrective technique to tackle the challenge of imbalanced data. Statistical experiments performed were based on secondary data obtained from a Baseline Survey on Learning Disabilities conducted by Kenya Institute of Special Education in 2019. It was found that Support Vector Machine (SVM) is the best base learner for imbalanced data with the highest classification efficiency compared to other classification models. For data with high dimensionality, it was found that the classification power of Artificial Neural Network (ANN) was better than that of SVM despite the need for significantly higher computational effort. When data dimensionality is reduced, it was observed that classification power of ANN reduces significantly. SVM was also found to be a more flexible model whose classification power is least affected by changes in data dimensionality. It was found that both Adaptive Boosting (AdaBoost) and Adaptive Synthetic Sampling (ADASYN) equally perform well in tackling the imbalanced data, with AdaBoost performing slightly better, although the difference was not statistically significant. The study concludes that SVM and ANN can be used to model highly imbalanced data to achieve the highest classification accuracy with respect to the minority class. ADASYN and AdaBoost methods can be used jointly to built a more robust corrective algorithm to tackle highly imbalanced data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AUC	Area Under the Curve
CBO	Cluster Based Oversampling
DT	Decision Trees
EDA	Exploratory Data Analysis
IR	Imbalance Ratio
KISE	Kenya Institute of Special Education
kNN	K-Nearest Neighbours
LD	Learning Disabilities
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MAE	Mean Absolute Error
ML	Machine Learning
NB	Naive Bayes
PAC	Probably Approximately Correct
PCA	Principal Component Analysis
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROS	Random Oversampling
RUS	Random Undersampling
SML	Supervised Machine Learning
SMOTE	Synthetic Minority Oversampling Technique
SRM	Structural Risk Minimization
SU-IERC	Strathmore University Institutional Ethics Review Committee
SVM	Support Vector Machines

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to one polished scholar, mentor and coach; my supervisor Dr Collins Odhiambo for his support, patience, and encouragement throughout the research process. Thank you sir for taking time to listen to all my problems, both mighty and puny, always pointing me to the right direction as you would see it, and never tiring to remind me to keep my focus on the goal. Thank you Dr Collins for allowing me stand on the shoulders of a giant!

Am deeply indebted to my classmates, particularly Christopher Maronga and Laban Bore for their candid critique of my initial research ideas, conceptual and analytical framework. Your insightful views, comments, and suggestions played a critical role in shaping my research work, thank you my dear friends. I would also like to thank the entire Strathmore Institute of Mathematical Sciences (SIMS) faculty for rigorous academic and research training. Special thanks to Prof Thomas Achia, Prof Samuel Mwalili, Dr Christopher Ouma, Dr David Khaoya and my special friend Ferdinand Othieno for leading the way in research.

Last, but certainly not the least, gratitude goes to my dear friends and colleagues at Kenya Institute of Special Education (KISE); Lydia Chege and Flora Malasi, without whom, financing my studies at Strathmore University would have been a total nightmare. Thank you Flora and Lydia for touching my life, may you find favour with the Lord.

DEDICATION

I dedicate this thesis to my family. My dear wife, Harriet Nasievanda, who has been my all time friend, cheerleader, and pillar, I offer you my utmost appreciation and love. Thank you for enduring many cold nights alone, taking care of the kids endless number of times while I sat at my desk reading, writing and coding.

My children, with whom I had almost no time to play with, I promise to do my best to have our fun outings, away from the city back again.

My dear mum, Janerose Nafula, I love you so much, and I can never ever be thankful enough to you. Your prayers, sacrifices, and love cannot be compared to anything or anyone, and there is no doubt they have brought me this far. Your handwriting is all over my physical, spiritual, emotional and intellectual being; you shaped me. You are fantastic, the most intelligent person I know. Most importantly mum, thank you for teaching me the power of silent prayer of faith, it works.

CHAPTER ONE

1 INTRODUCTION

1.1 Background of the study

Over the last few decades, there has been a rapid increase in the production, acquisition and storage of data in different fields such as business, communication, education among others, resulting into what is described as *the era of big data*. The rapid growth in the production of massive, voluminous, variety and high velocity data is attributable to advanced modern computer infrastructure to store and process the data. Availability of machine learning (ML) toolkits such as R, Python, TensorFlow and Hadoop provide opportunities to synthesize data into useful information. The advancement in computational and data processing infrastructure provides opportunities to apply these ML methodologies in different domains of life (Hyde et al., 2019). One such area of life that statistical and ML methodologies can be applied is diagnostic assessment of neurodevelopmental disorders such as cognitive or learning disabilities (LD).

Diagnostic assessment of cognitive or learning disabilities (LD) closely resembles supervised machine learning tasks. A typical conceptual model for a supervised machine learning problem is characterised by both input space of the attribute set X and a predefined outcome, Y. Thus, historical data points (x^i, y^i) are used to train a machine learning model. LD data has both attribute set variables (scores from the psychometric tool) and predefined binary outcome which shows whether the individual has LD or not, thus, with appropriate Mathematical definition of LD data, the diagnostic process can be modelled using supervised machine learning approaches.

In general, learning disability (LD) is a rare condition a low prevalence in a normal population. The prevalence of LD vary significantly across different parts of the world. For instance, in 2016, the Centres for Disease Control and Prevention (CDC) estimated that 1 in 53 children has a learning disability in the USA. In 2019, the Learning Disabilities Association of America (LDA) estimated that 1 in 45 children has

a learning disability. In the same year, 2019, the Kenya Institute of Special Education conducted a survey in Kenya (KISE, 2019) and found that at least 1 in 40 children in public primary schools have a learning disability.

Learning disability (LD) manifests itself in heterogeneous forms such as dyslexia, dyscalculia, dysgraphia, auditory processing disorder, language processing disorder, nonverbal learning disabilities and visual perceptual/visual motor deficit (Sternberg, 2018) among others. The heterogeneous nature of LD with respect to its manifestation, etiology, risk factors, severity and response to interventions has led some researchers to cast doubt on the traditional analysis methods used to diagnose LD (Miciak & Fletcher, 2020). This situation causes one to wonder whether statistical and machine learning approaches may contribute towards finding better solutions to diagnostic assessment of LD.

Supervised machine learning (SML) tasks involve estimation of a predictive mapping function $f : A \mapsto B$ such that;

$$\forall\, b \;\in\; B\, \exists \text{ some } a \;\in\; A \tag{1.1}$$

so that, for a sample n, the $f(a) = b$ for every i^{th} instance $(i \ldots n)$ with the minimum error. A typical SML task involves providing complete instances of input variables and the corresponding target variable at model training stage from which the model *'learns'*. A supervised learning model is thought to be successful when it can; (a) predict accurately the target outcomes of the training data to a reasonably high degree of accuracy (in-sample performance) and (b) be used to generalise new datasets out of the training sample (out of-sample performance). The training dataset is, thus, expected to be as much as possible a full representation of all dynamics of the problem under study. A binary bisector of data into one training and one testing datasets may not provide sufficient dynamics to represent the problem under study, and thus, additional statistical treatment of the data is needed during model training.

One such statistical treatment is cross-validation (CV) technique (Raschka, 2018). The CV technique is often used to improve the model's predictive ability by recursively separating sections of the data into training and testing datasets, until the model has been trained and tested on all the available data (Gujar & Vakharia, 2019). Depending on the size of data and the computational ability available, either K-fold cross-validation or leave-one-out cross validation (LOOCV) may be used (Magnusson et al., 2020). A K-fold cross-validation approach separates the data into K-folds (subsets) and trains the model on all but one of the subsets and tests on the remainder, and the average of scores on all the runs is computed. In LOOCV, all the data points except one are used to train the model and tested using the left-out point. This process is repeated for each of the data points, and average scores computed.

Classification and linear regression analysis are the two main approaches in the supervised machine learning (SML) framework (B. Hou et al., 2017). The main difference between the application of these two approaches lies in the nature of the outcome variable Y. Linear regression analysis is used in the cases where the target variable is continuous while classification is used where the target variable is categorical. Further, if the target variable has only two possible outcome, the problem is described as a binary classification. In cases where the target variable has more than two possible outcomes, the problem is described as multi-class classification. This study focuses on binary classification a SML methodology since the target variable (e.g., diagnosis) in the learning disabilities (e.g. absence=0 or presence=1 of LD).

The concept of classification in ML has previously been considered in a broader sense cross-cutting supervised learning (Zheng et al., 2018), unsupervised learning (L. Zhang et al., 2019) and semi-supervised learning (B. Hou et al., 2017) approaches. In the context of unsupervised learning, classification is used as pattern discovery technique. Classification is used to model differentiable clusters or groups in the data (e.g., segmentation tasks). In supervised learning, models are trained to classify objects based on existing data points (x^i, y^i) to model statistical rules to generalize classification criteria for unobserved data (i.e., prediction tasks).

The construction of a classification model is aimed at studying the underlying relationships between the input set of attributes X and the class label $y^i, i \in [0, 1]$, and identifying a

function that best maps the input space of the training data points to the target variable. It is hoped that the function so identified will correctly map a set of attribute to the correct class label in any previously unseen data. The overall objective in supervised machine learning is to build a function or data model with a reasonably good generalization ability. For the best generalization, the model should properly fit the training dataset (Nasteski, 2017). If the model poorly fits the training dataset, then both the training error and generalization error are high, this phenomenon is described as *model underfitting*. In such a case, a better learning approach should be sought to simultaneously reduce both the training error and the generalization error.

There arises situations where the model fits the training dataset too well, increasing performance on training dataset while the generalization error becoming worsen. This phenomenon where a model performs well on training dataset and poorly on new observation is described as *model overfitting*. According to Bilbao & Bilbao (2017) overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. To avoid overfitting, regularity conditions such as Minimum Description Length (MDL) may be introduced to limit exhaustive learning on the dataset. The MDL states that given a limited set of observed data, the best model is the one that permits the greatest compression of the data (Grünwald & Roos, 2019). That is,the more data is compressed, the more we learn about the underlying regularities that generated the data. Such a process generates an unavoidable *maximum-generality* bias favouring the discovery of more general rules (Datta et al., 2017). However, such an inductive bias of maximum-generality poses a serious challenges in learning from and classification of imbalanced data.

The problem of imbalanced data in binary classification tasks arises when there are significantly more instances of one class type than the other. According to Lin et al. (2020), most of the contemporary works in class imbalance concentrate on imbalance ratios ranging from 1:4 up to 1:100. In real-life applications such as fraud detection or cheminformatics we may deal with problems with imbalance ratio ranging from 1:1000 up to 1:5000 (Lin et al., 2020). Diagnosis of a rare disease for instance is a good illustration of a binary classification task, where the interest is to establish the

presence or absence of the disease. In such a situation, the number of positive cases will definitely be significantly lower compared to negative cases.

In practice, correct identification of individuals at risk (the positive cases) is crucial for the purpose of prescribing interventions. In this application, a favourable classification model would be one that provides a precise and higher identification rates of the infrequently occurring class. A classification model that describes the attributes of the rare classes with higher precision need to be more specialized rather than a conventional simplification by use of general rules emanating from data compression techniques as described by Grünwald & Roos (2019). Currently, classification rules that predict the small classes tend to be rare, ignored or undiscovered; consequently, test samples belonging to the small class are misclassified more often than those belonging to the most prevalent class. Therefore, in this study, we argue that *maximum-generality* bias works well for the large class but not for the small class.

Noisy data also contributes to difficulties experienced in learning from imbalanced data. In practice, most data obtained from psychometric assessment tools contains various types of errors. Stochastic errors, for instance, occur frequently in disciplines that rely substantially on domain expert judgement to compliment the conclusions arrived at through a rigorous scientific procedure, thus increasing the background noise in the data. If the noise in the data is sufficiently high, a standard classification model may fail to distinguish between genuine rare instances and the noise-induces instances. Noise-tolerant models, as described by Bootkrajang & Chaijaruwanich (2020) are aimed at minimizing the impact of noise in the training dataset also tend to treat the minority class as noise. This is because the attribute set associated with these rare class may appear to have an anomalous distribution, making it difficult to learn from imbalanced data (Lin et al., 2020).

The classification task in machine learning poses difficulties when the distribution of classes in the data is imbalanced. In most real world application, identification of members of the minority class is of prime importance (Grünwald & Roos, 2019). For example, diagnosis of neurobiological condition which is a rare event using machine learning is important yet difficult because of imbalanced distribution in the data because positive cases are the minority. Any dataset with an unequal class distribution is

technically imbalanced. However, a dataset is said to be imbalanced when there is a significant, or extreme, disproportion among the number of examples of each class of the problem (Lin et al., 2020).

A slight within-class imbalance is not a concern for standard classification tasks using well-known approaches such as ZeroR, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), K-Nearest Neighbours (KNN), Naive Bayes (NB), Logistic regression (LR), Support Vector Machines (SVM), Decision Trees (DT) and Random Forests (RF). This is because the problem can be assumed to be like a normal classification predictive modelling problem. However, an extreme class imbalance can be challenging to model, requiring the use of more specialized techniques. Several approaches have been proposed to tackle the problem of class imbalanced. These approaches can be put in three general categories namely: data level approaches, algorithm level approaches and cost-sensitive approaches.

Sampling methods primarily constitute data-level approaches that attempt to rebalance the distorted class distribution by either creating new instances of minority classes by duplication (oversampling) or deleting instances of major classes (undersampling). Both oversampling and undersampling uses the same logic of rebalancing the distribution of classes. There are many existing undersampling approaches such as Random Undersampling (Hasanin & Khoshgoftaar, 2018), One-sided Undersampling (Y. Hou et al., 2019), Condensed Nearest Neighbour Rule (Siddappa & Kampalappa, 2019), Cluster Based Undersampling (Rayhan et al., 2017), and NearMiss Undersampling (Ishwaran & O'Brien, 2020). Similarly, there are many oversampling approaches such as Random Oversampling (J. Zhang & Chen, 2019), Synthetic Minority Oversampling Technique (SMOTE) (Elreedy & Atiya, 2019), Borderline-SMOTE (Smiti & Soui, 2020), Adaptive Synthetic Sampling (ADASYN) (Hasmita et al., 2019), and Cluster Based Oversampling (CBO)(Gong et al., 2019).

Algorithm level approaches attempt to tackle the problem of imbalanced data by modifying the learning procedure itself through a recursive process. Statistical boosting is a popular class of algorithm level approaches to tackling imbalanced data problem. While statistical boosting methodology was not primarily designed for imbalanced data problem, adaptive boosting (AdaBoost) algorithm is has been widely used to tackle

the imbalanced data problem (Baig et al., 2017). The AdaBoost Algorithm pay more attention to examples of minor class or examples which are hard to classify by assigning them higher weights than the majority class on each boosting round. Cost-sensitive methods use a cost matrix with different misclassification costs for each class (Tao et al., 2019) and try to minimize misclassification costs during learning instead of maximizing accuracy. Typically, cost-sensitive approaches are a blend of data level and algorithm level approaches to tackling imbalanced data problem.

1.2 Statement of the problem

Despite significant development in the field of machine learning, classification of imbalanced data still poses significant challenges in the development of efficient learning algorithms. Available research evidence suggest that standard classification models perform poorly when applied to imbalanced data, yet we find imbalanced data in most real life applications in areas such as fraud detection, medical diagnosis, instruction detection among others. The field of cognitive neuroscience for instance, frequently encounter imbalanced data in diagnostic assessment of individuals with learning disabilities. Thus, robust statistical learning methods are required for successful application of machine learning in predictive modelling of any rare event. Literature reports a number of solutions to tackle the challenges posed by imbalanced data. These solutions can be either algorithm-based or data-based approaches. Algorithm-based approaches attempt to adapt a standard classifier to suit the case at hand, while data-based approaches attempt to tackle the problem by rebalancing the distribution between the majority and minority classes. To date, experts in different disciplines choose either algorithm-based or data-based solutions in addressing merely based on contextual convenience rather than conventional statistical principles. This may be the case due to limited research evidence that blend rigorous statistical theory and empirical data from the field of interest, thus lacking general guiding principles for choosing a particular solution. This study, therefore, sought to bridge this knowledge gap by providing a systematic statistical learning framework from imbalanced data, with applications in diagnostic assessment of learning disabilities.

1.3 Research Objectives

The **main objective** of this study is to analyse effectiveness of classification models in a supervised machine learning context, with the aim to document a framework for statistical learning from imbalanced data as applied to diagnostic assessment of learning disabilities data.This study answers the following **research questions;**

1. Which classification model is best suited as a base learner classifier for imbalanced data?
2. Which sampling technique is best suited to tackle the imbalance data problem?
3. How does the best sampling technique compare with statistical boosting in tackling the imbalanced data problem?

1.4 Justification of the study

The challenge of imbalanced data classification continues to be a difficult problem when designing and deploying machine learning solutions in modelling rare events. This challenge is even worse when minority class also exhibit character heterogeneity among individual members. The best illustration of this scenario is found in cognitive neuroscience field, particularly in assessment of children with learning disabilities (LD). LD manifests differently in different persons as dyslexia (*difficulties in reading*), dyscalculia (*difficulties in arithmetic*), dysgraphia (*difficulties in graphs interpretation*) among others. In classroom, one can easily mistake a child with LD to one with other academic difficulties or vice versa. When the cause of the child's difficulties is correctly identified, appropriate intervention measures can be taken to support the child gain compensatory skills to their inadequacies. In the era of artificial intelligence (AI), it may be desirable and indeed beneficial to develop learning algorithms to detect LD with the highest precision. To this end, literature suggest that application of machine learning in diagnosis of LD is still at infancy stages especially in Sub-Saharan Africa (SSA). It is against this backdrop that it was important to conduct this research, which is founded on a rigorous statistical theory and a substantial understanding of diagnostic assessment of LD.

1.5 Thesis organization

This thesis is organised into five chapters as follows: **Chapter one** provides contextual background information on imbalanced data classification. The chapter also highlights the statement of the problem, research objective and research question, and the motivation of using learning disabilities diagnostic assessment data under the justification section. **Chapter two** presents a critique of literature on the nature of imbalanced data problem, metrics of evaluating model performance, standard classification models and a review of available techniques to tackling imbalanced data problem. **Chapter three** presents the research methodology applied to address the research questions. The chapter presents a detailed description of the data used, data transformation techniques, model selection criteria, data analysis procedures and ethical considerations in research. **Chapter four** presents the results and discussions of the study. The discussions presented in chapter four are based on the research questions beginning with identification of the base learner classifier, proceeds to identification the best sampling methods for tacking imbalanced data and closes by presenting a comparative analysis of the performance of the best sampling technique and statistical boosting method (Adaptive boosting) in tackling imbalanced data problem. **Chapter five** presents conclusions and recommendation of the study based on the analysis of data, and synthesis of contemporary literature on imbalanced data classification techniques. The conclusions made herein are aimed at advancing statistical knowledge on imbalanced data classification methodologies and offer insightful recommendations on best practices in assessment of learning disabilities.

CHAPTER TWO

2 LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of literature on the context of imbalanced data in diagnosis of rare events, classification models, model evaluation metrics and approaches to tackling imbalanced data. The chapter begins by critiquing perspectives of imbalanced data in machine learning, then proceeds to detail the theoretical, analytical and domain applications of standard classification models and performance evaluation metrics. The chapter ends by analysing approaches of tackling the imbalanced data problem with more emphasis on data-based approaches.

2.2 Review of the imbalanced data problem

Classification and regression analyses falls within the context of supervised machine learning category. Supervised machine learning tasks are characterised by '*labelled*' datasets. The conceptual model of a labelled dataset suggest presence of both an input space variances X and known or well-defined response variable Y. Broadly, regression analysis is preferred when the response variable Y hold numerical values while classification is preferred when the response variable Y is categorical in nature. In constructing a classification model, a learning algorithm reveals the underlying relationship between the attribute set and class label, and identifies a model that best its the training data. The task of the constructed classifier is to predict the class labels for any unseen input objects.

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is disproportional. According to Rayhan et al. (2017), the disproportional distribution can vary from a slight imbalance to a severe or extreme imbalance, where there is one example on the minority class type in hundreds, thousands, or millions of examples on the majority class or classes.Imbalanced classifications pose a challenge for predictive modelling as most of the machine learning

algorithms used for classification were designed around the assumption of an equal number of examples for each class (Rayhan et al., 2017). This results in models that have poor classification performance, specifically for the minority class. This is a problem because in many practical cases, identification of the minority class is more important. Therefore the problem is more sensitive to classification errors for the minority class than the majority class.

2.3 Classification models

Classification is a predictive modelling problem that involves assigning a class label to each observation. When the classification task has a only two classes, this is referred to as binary classification while in cases where there are more than one classes, then we have a multi-class problem. There are many different classification models that are used in machine learning for the purpose of classifying observations. These models can be grouped into categories based on their computational frame such as covariance matrix where linear discriminant analysis and logistic regression belong, similarity function where k-nearest neighbour belong, tree based models where decision trees and random forest belong and distance based models. This section presents a review of literature on these models.

2.3.1 Naive Bayes

Naive Bayes (S. Chen et al., 2020) is a popular classification algorithm in in machine learning and data mining. A Naive Bayes classifier is a basic probabilistic based technique, which can foresee the class membership probabilities (Abellán & Castellano, 2017). As it computes the likelihood of membership for each class, it can handle the missing attribute values by omitting the corresponding probabilities for those attributes. In a Naive Bayes classifier, the effect of an attribute on a given class is also independent of those of other attributes, which is known as class conditional independence. This classifier computes the probability to classify or predict the class in a given dataset. Granik & Mesyura (2017) used naive bayes in their study to detect fake news and reported 74% accuracy.

Feng et al. (2018) used naive bayes classifier in predicting slope stability. Another study discusses the performance of naive bayes within the context of deep learning where the classifier was found to have a classification error of 13.85% (Shi et al., 2017). Literature seems to suggest that this model is popular in multi-dimensional contexts. In fact, some researchers argue that naive classifier can sometimes cause zero-probability problems. Other draw backs that have been associated with naive bayes include high computational cost when the explanatory variables are significantly high. Since there is no guidance on how many input variables are considered to be significantly high, we would like to see the performance of this model with disability assessment data will 88 input items which are collapsible to seven variables.

2.3.2 Decision tree

A very well-known and mostly discussed technique for classification and then used for prediction is decision trees (Kamiński et al., 2018). According to some scholars such as Triantafyllidis et al. (2020) and Cañete-Sifuentes et al. (2021), decision trees models are popular classification models that resemble human reasoning. The popularity of this model emanate from its simplicity in making stand-alone decision at every instance node. However, the model may become overly complicated and slow to implement in high dimensional data (Cañete-Sifuentes et al., 2021). While multivariate decision trees (MDTs) have recently been thought to mitigate this problem, Triantafyllidis et al. (2020) claims that the implementation have no statistical sustain and, hence, there is a lack of general understanding of the actual capabilities of existing MDT induction algorithms.

The core algorithm for building decision trees called ID3 proposed by Quinlan (Elaidi et al., 2018). ID3 algorithm constructs a decision tree by employing a top-down approach in which a greedy searching through the given training dataset is used to test each attribute or context at every node. It calculates the entropy and information gain which is a statistical property that is used to select which attribute to test at each node in the tree (Elaidi et al., 2018). C4.5 builds decision trees from a training dataset in the similar procedure as ID3. This algorithm generates contextual decision rules to predict mobile user activity. C5.0 is another modified decision tree algorithm used for predictive

modelling (Damanik et al., 2019). C5.0 decision tree is significantly faster and more memory efficient than C4.5.

Despite progressive improvement in algorithmic development of decision tree algorithms but fails to show how these algorithms performs on different data. The situation is even worse when it comes to classification of imbalanced data where limited research has been done to test the performance of different decision tree algorithms. It will be important therefore to see how this turns out to be when applied to an empirical data with extreme class imbalance.

2.3.3 Linear Discriminant Classifier

Linear Discriminant Analysis (LDA) is a simple classification rule, which presents a low degree of overfitting, and is therefore useful for application in small-sample cases. Provided that the underlying feature-label distribution is linearly separable, LDA produces very good classifiers. Linear Discriminants is a statistical method of dimensionality reduction that provides the highest possible discrimination among various classes, used in machine learning to find the linear combination of features, which can separate two or more classes of objects with best performance.

Linear Discriminant Analysis (LDA) has been widely used in many applications, such as pattern recognition, image retrieval, speech recognition, among others. The method is based on discriminant functions that are estimated based on a set of data called training set. These discriminant functions are linear with respect to the characteristic vector, and usually have the form;

$$f\left(t\right) = W^t x + b_o, \tag{2.1}$$

where W represents the weight vector, x the characteristic vector, and b_o a threshold.

The criteria adopted for the calculation of the vector of weights may change according to the model adopted. In de Jia et al. (2019), for example, the parameters were determined by maximum-likelihood calculated from the training data. In terms of standardization, LDAs are the classifiers used that most meet recommendations from various insurance firms in Australia, being often chosen by the authors for their simplicity and the emphasis that is wanted on the proposed characteristics (Rajaguru & Prabhakar, 2017).

One of the most recent classifiers proposed with LDA was able to classify and accurately differentiate normal and abnormal heartbeats through a simple, efficient, and rapid method that did not require the use of complex mathematics (Kolukisa et al., 2018). LDAs can easily overcome problems generated by the imbalance of the training set a difficulty presented by approaches based on SVM. Another advantage is that the LDA classifier, when compared to support vector machine, requires less training time, because it simply calculates training data statistics, and then the classification model is defined; so it is not iterative. The current study sought to evaluate performance of linear discriminant analysis on classification of an imbalanced dataset in whose the minority class exhibit additional heterogeneous characteristics.

2.3.4 Logistic regression

Logistic regression (De Caigny et al., 2018) is another probabilistic based classification model that can be used in machine learning. Logistic regression estimates the probabilities of class membership using a logistic function, which is also referred to as sigmoid function. The hypothesis of logistic regression tends it to limit the function between 0 and 1. This classifier measures the relationship between the categorical dependent variable and one or more independent variables for a given dataset. The dependent variable is the target class to be predicted. The independent variables are the attributes or contextual features to be used in the prediction.

In this study, teacher rating scores from learning disabilities evaluation scale will be used as independent variables and a binary label assigned to the learner by the teacher as either at risk or not will represent the dependent variable. Consider a typical linear regression model;

$$p = \alpha + \beta x, \tag{2.2}$$

where a continuous dependent variable y is replaced with a probabilistic quantity p restricted to taking values between 0 and 1. Model 2.2 can be made suitable by introducing regularity conditions to transform variable $-\infty < y < \infty$ into a quantity $0 \leq p \leq 1$.

Suppose we define;

$$p = e^{\alpha+\beta x}. \tag{2.3}$$

(De Caigny et al., 2018)

We would guarantee that p is positive since right hand side of model 2.3 cannot produce a negative real value, however, it can result into a value greater than 1. To achieve our final constraint of estimating a non-negative p equal to or below 1, we define;

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{2.4}$$

(Dedeturk & Akay, 2020)

The right hand side of model 2.4 is a logistic function, which can be used to define the probability of an event occurring. By definition, if an event occurs with probability p, the odds in favour of the event are $\frac{p}{1-p}$ to 1. By implication, the odds in favour of success of an event occurring with a probability defined by model 2.4 are;

$$\frac{p}{1-p} = \frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1}{e^{\alpha+\beta x}}}, \tag{2.5}$$

$$= e^{\alpha+\beta x}$$

.

Taking the natural logarithm of each side of this equation,

$$ln\left[\frac{p}{1-p}\right] = ln[e^{\alpha+\beta x}], \tag{2.6}$$

$$= \alpha + \beta x$$

.

Thus, modeling the probability p with a logistic function is equivalent to fitting a linear

regression model in which the continuous response y has been replaced by the logarithm of the odds of success for a dichotomous random variable. Instead of assuming that the relationship between p and x is linear, we assume that the relationship between $ln[p/(1-p)]$ and x is linear. The technique of fitting a model of this form is known as logistic regression (De Caigny et al., 2018).

Logistic regression has been used in different fields for classification tasks. Limited studies have used logistic regression method in classification. For instance, Dedeturk & Akay (2020) considered logistic regression in their work to predict spam filtering trained by bee colon algorithm. Major draw back of logistic regression classifier is its linearity assumption. The model assumes that the classes are linearly separable which is not always true in most practical application tasks such as diagnosis tasks. For this reason, other extensions such as lasso logistic regressions and inverse logistic regressions have been developed (Robles et al., 2019). It may be important therefore for other application domains such as diagnosis of learning disabilities to study the applications of logistic regression. This study might be an opportunity to provide insights on how to extend the functionalities of logistic regression into the domain of disability assessment.

2.3.5 K-nearest neighbours

K-nearest neighbours (Cunningham & Delany, 2020) is a simplest classification technique in machine learning. It is a type of instance-based learning, or lazy learning. In this classification technique, it takes into account local approximation and all the computation is deferred until classification. The classifier stores all the available cases in the given dataset and classifies new cases based on distance functions like Euclidean distance as a similarity measure.The Euclidean distance between two arbitrary points q and p is defined by;

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \tag{2.7}$$

After that the K most comparable occasions, called the neighbours, are controlled via seeking through the whole preparing set for another test information point. Finally, the

expectation result is made by abridging the yield variable for those K cases dependent on larger part casting a majority voting of the neighbours. A number of researchers use KNN classifier in different areas such as the study of fuzzy logic (Patel & Thakur, 2019), threat detection (Sarma et al., 2017) and detection of plant disease (Hossain et al., 2019) among many other applications. In all these instances, researchers tend to use k-NN with a boosting or sampling mechanism of some sort unless the event of interest is thought to have a higher prevalence rate. These approaches seem to suggest that k-NN tend to ignore the minority classes and thus a more robust solution need to be sought in order to improve the performance of k-NN. Additionally, k-NN does not seem to have had widespread application in the field of diagnosis such as disability assessment. It will be interesting therefore to see the extension of k-NN in classification tasks for disability assessment.

2.3.6 Artificial Neural Networks

According to an overview by Dick et al. (2019), as a sub-field of AI, machine learning provides indispensable tools for intelligent data analysis. Three major branches of machine learning have emerged since electronic computers came in to use during the 1950s and 1960s: statistical methods, symbolic learning and neural networks. ANN have been successfully used to solve highly complex problems within the physical sciences and as of late by scholars in organizational research as digital tools enabling faster processes of data collection and processing. As practical and flexible modelling tools, ANN have an ability to generalize pattern information to new data, tolerate noisy inputs, and produce reliable and reasonable estimates (El-Khatib et al., 2019).

Artificial neural networks (ANNs) were originally developed as mathematical theories of the information-processing activity of biological nerve cells, the structural elements used to describe an ANN are conceptually analogous to those used in neuroscience, despite it belonging to a class of statistical procedures The algorithm has three layers; the input, hidden and output layers. This model has wide application in machine learning, and could be used in medical diagnosis, including cognitive neuroscience field. The era of artificial neural network (ANN) began with a simplified application in many fields and remarkable success in pattern recognition (PR) even in manufacturing

industries.Statistical pattern approach has been the most commonly studied and utilizes in practice (Liu et al., 2018).

ANN belong to a wide class of flexible nonlinear regression and discriminant models, data reduction models, and nonlinear dynamical systems. ANN are similar to statistical techniques including generalized linear models, nonparametric regression and discriminant analysis, or cluster analysis. ANN can be used to perform nonlinear statistical modeling and provide new alternatives to logistic regression, the most commonly used method for developing predictive models for dichotomous outcomes in medicine (Liu et al., 2018). Although ANN do not require knowledge of data source, they require large training sets due to the numerous estimated weights involved in computation. They may require lengthy training times and the use of random weight initializations may lead to different solutions.

Despite successful applications, ANN remain problematic in that they offer us little or no insight into the process(es) by which they learn or the totality of the knowledge embedded in them. Several limitations of ANN are identified in the literature: they are limited in their ability to explicitly identify possible causal relationships (El-Khatib et al., 2019), they are challenging to use in the field, they are prone to over fitting, model development is empirical potentially requiring several attempts to develop an acceptable model (Dick et al., 2019), and there are methodological issues related to model development.

2.3.7 Support Vector Machines

Support Vector Machines (SVM) (Pisner & Schnyer, 2020), is a widely used classification technique for various predictive and classification analysis. SVM was developed in the 1990s based on the theoretical considerations of Vladimir Vapnik on the development of a statistical theory of learning: Vapnik-Chervonenkis theory (Prasad & Rao, 2017). SVMs were quickly adopted for their ability to work with large data, the small number of hyper parameters, their theoretical guarantees, and their good results in practice (Kim et al., 2017). Since the scope of this study focuses on application SVM as a binary classifier, consider a classification problem with labels Y, and k-dimensional features space X. Let $Y \in [0, 1]$ denote class label, and W and b are operationally defined

parameters

$$f(X) = W^T X + b. \tag{2.8}$$

For simplicity, W in equation 2.8 corresponds to a typical 'best line of fit' on a two-dimensional Cartesian plane, and b defines the function bias. The function $f(x)$ denotes the hyperplane that separates the two regions and facilitates in classification of the data set into distinct groups. Zendehboudi et al. (2018) describes SVM as a method in statistical classification which is based on the maximization of the margin between the instances and the separation hyperplane. This implies that the SVM algorithm solves for parameter W in equation 2.8 then maximizes the margin of separation between distinct classes.

Nalepa & Kawulok (2019) describes SVM as a non-probabilistic binary linear classifier, that has the ability to linearly separate the classes by a large margin, it become one of the most powerful classifier capable of handling infinite dimensional feature vectors. To do this, a hyperplane is chosen in the vector machine, which is a line that can take part into the variable space. Utilizing this hyperplane, the vector machine learning computations finds the coefficients that brings about the best detachment of the classes. The margin of the hyperplane $W^T X + b = 0$ is given by Prasad & Rao (2017) as:

$$\frac{W}{\|W\|} \cdot (x_+ - x_-) = \frac{W^T(x_+ - x_-)}{\|W\|}$$

$$= \frac{W^T\left((\frac{+1-b}{W^T}) - (\frac{-1-b}{W^T})\right)}{\|W\|}, \tag{2.9}$$

$$= \frac{2}{\|W\|}.$$

There are many such hyperplanes which can split the data into two regions. But SVM ensures that it selects the hyperplane that is at a maximum distance from the nearest data points in the two regions. There are only few hyperplanes that shall satisfy this criterion. By ensuring this condition SVM provides accurate classification results (Pisner & Schnyer, 2020). Owing to this flexibility in application, SVM has been widely applied in many binary classification tasks. For instance, Huang et al., (2018) used SVM in cancer genomics and concluded that it is a powerful classification model. Zendehboudi et al. (2018) used SVM in forecasting solar and wind energy resources. There are many examples in literature that suggest that the performance of SVM outperforms many classification algorithms. However, little attention has been given to study the performance of SVM in classification imbalanced data, hence this study.

2.3.8 Random forest

Random forest (RF), which was formally proposed in 2001 by Leo Breiman and Adele Cutler, is part of the automatic learning techniques. RF classifier is an ensemble machine learning approach that takes into account multiple learning algorithms together while generating a prediction result. Random forest combines bootstrap aggregation (bagging) (Lee et al., 2020) and random feature selection (Barni et al., 2020) to construct a collection of decision trees exhibiting controlled variation.

The random forest is part of the family set methods that take the decision tree as an individual predictor, they are based on the methods of Bagging, Randomizing Outputs and Random Subspace excusing boosting. The RF algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy (Lakshmanaprabu et al., 2019). It is an ensemble learning method for classification and regression that constructs a number of decision trees at training time and delivers the class that is the mode of the classes output by individual trees.

Random forest algorithm begins by drawing a bootstrap sample Z^* of size n from the training data, then grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following three steps for each terminal node of the tree, until the minimum node size n_{min} is reached; (1)Select m variables at random from the p variables, (2) Pick the best variable/split-point among the m, and (3) Split the node into two daughter

nodes. Once this recursive process is complete, an output ensemble $\{T_{b_1} B\}$ of trees is produced (Lakshmanaprabu et al., 2019). To make a prediction of the new data point x, a regression of the form;

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x). \tag{2.10}$$

(Petkovic et al., 2018) is used while for classification, we define a function $\hat{C}_{rf}^{B}(x)$ such that the b_{th} random forest tree is define by Edla et al. (2018) as;

$$\hat{C}_{rf}^{B}(x) = \text{majority vote } \{\hat{C}_{rf}^{B}(x)\}_{1}^{B}. \tag{2.11}$$

In random forest classification method, many classifiers are generated from smaller subsets of the input data and later their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set (Paul et al., 2018). This ensemble learning strategy has recently become very popular. Before RF, Boosting and Bagging were the only two ensemble learning methods used. RF has been extensively applied in various areas including modern drug discovery, network intrusion detection, land cover analysis, credit rating analysis, remote sensing and gene microarrays data analysis. There are two ways to evaluate the error rate. One is to split the dataset into training part and test part (Lakshmanaprabu et al., 2019). We can employ the training part to build the forest, and then use the test part to calculate the error rate. Another way is to use the Out of Bag (OOB) error estimate. Because random forests algorithm calculates the OOB error during the training phase, we do not need to split the training data (H. Zhang et al., 2017).

Overall, random forest generates a number of decision trees rather than a single decision tree, to predict the final output activity class of the mobile phone users utilizing their phone log data. By generating multiple decision trees for a given dataset, it reduces the over-fitting problem causes in the single decision tree discussed earlier. Despite the higher classification rates achieved by the above-mentioned algorithms, there is little evidence to suggest how these models classify imbalanced data.

2.4 Solutions to tackling imbalanced data problem

Data sampling has received significant attention among data mining researchers. Rayhan et al. (2017) compare over and under sampling using C4.5 decision trees. Their results show under sampling to be more effective than over sampling for improving the performance of models built using decision tree algorithms such as C4.5. Hasanin & Khoshgoftaar (2018) however, shows that under and over sampling result in roughly equivalent models when using C5.0 (C4.5's commercial successor) and Naive Bayes. In addition to random over and under sampling, several more 'intelligent' data sampling techniques have been proposed to sample data in such a way that benefits the classifier. Smiti & Soui (2020) and Fernández et al. (2018) examine the performance of some of these 'intelligent' data sampling techniques such as SMOTE, borderline-SMOTE, and Wilson's Editing. Hasmita et al. (2019), examine the performance of seven different sampling techniques including both random and 'intelligent' techniques.

Some literature suggests that undersampling methods have the advantage of decreasing the amount of data which is good for learning with large data sets, however by deleting many instances of major classes we lose information and classification performance may decrease. A second drawback of many undersampling methods is that they are designed for a binary case and a proper extension for the multi-class case is not always possible. Oversampling methods are in general easily extendable to a multi-class case since each minor class is oversampled separately, it is a beneficial to have all data for learning. The way of new instances generation in oversampling may influence learning performance. Random Oversampling clones instances of minor classes by duplication of existing data points.

SMOTE, Borderline-SMOTE, ADASYN and CBO new instances are created by random linear combinations of existing instances. There have also been proposed data cleaning methods which might be combined with oversampling or undersampling methods. Well-known data cleaning methods are Neighbourhood Cleaning Rule algorithm (Pereira et al., 2020) and Tomek links (Vuttipittayamongkol & Elyan, 2020). Data cleaning methods used to delete noise and border instances between classes may improve the performance of learning algorithms.

CHAPTER THREE

3 RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the nature of the data that were used in the study, experimental design, model selection,dimensionality techniques, data transformation, data analysis approaches and ethical considerations in research.

3.2 Data sources and description

This study was based on the analysis of secondary data from a baseline survey on *'Development of a Culturally Responsive Learning Disabilities Diagnostic Inventory' (LDDI)* conducted by Kenya Institute of Special Education (KISE) between 2018 and 2019. The purpose of the survey was to establish the role of teachers in identification and support of children with learning disabilities in primary schools in Kenya.Teachers were interviewed on their awareness of characteristics of children who may be at risk of learning disabilities (LD). After the interview, the teacher was asked to select some learners they suspected might be at risk of learning disabilities based on their classroom encounter with the learner.

Teachers were then given a *Learning disabilities evaluation scale (LDES-R2)* which is psychometric tool with 88 test items. The test items are behavioural characteristics established by cognitive neuroscience and psychology professionals to assess the risk factors associated with learning disabilities. Table 3.1 shows an extract of some of the test items as presented on the LDES-R2 psychometric tool. Each item is scored on a three-point Likert scale where; 1=Rarely, 2=Inconsistently and 3=Consistently.

According to the report (KISE, 2019), 4,473 learners were assessed using the LDES-R2 because their teachers had suspected them to be at risk of learning disabilities (LD). In addition, the learner's class performance in language and Mathematics was recorded on the assessment sheets, and collected by the research team. A multidisciplinary team

analysed and compared the scores on the assessment sheets and would label the child as either being at risk of LD or not at risk of LD. All the data was then typed into an excel file, including the diagnosis made by the multidisciplinary team. Based on this methodology, it was found that 2.41% of children were labelled as being at risk of learning disabilities.

Table 3.1: Sample test items from LDES-R2 psychometric tool

Test item	1	2	3
a). The child has difficulty classifying objects based on similarities			
b). The child has a limited speaking vocabulary			
c). The child uses inappropriate spacing between words or sentences			
d). The child does not remember math facts			
e). The child reverses letters and numbers when writing			

The data used in this study therefore, has 89 variables; 88 attribute variables and 1 binary response variable (*At risk* or *Not at risk*). The attribute set is made up of 7 sub-scales namely; cognition (17 test items), listening (7 test items), speaking (9 test items), reading (14 test items), writing (14 test items), spelling (7 test items), and Mathematics (20 test items). Finally, this data is imbalanced since there are two classes where one class dominates the other; Children at risk of LD (2.41%) form the minority class while the rest who are not at risk of LD, yet experience academic difficulties stemming from other heterogeneous reasons (97.59%) form the majority class.

3.3 Theoretical Foundation

This study was motivated by the Chaos theory by Edward Lorenz (1963). The theory posits that within apparent randomness of chaotic and complex systems, there exist underlying patterns, interconnectedness, self-similarity, fractals, and self-organization (Gardini et al., 2020). The chaos theory provides a theoretical framework for Statistical modelling of stochasticity in the behaviour of naturally occurring systems. The complex system, Ω, in this study is a binary classification of a set of scores of n children some of whom (*minority*) have neurobiological condition described as *'At risk'* of learning disabilities (LD), while the majority are *'Not at risk'* of LD, but may exhibit similar characteristics. As such, the scores from either groups overlap each other as presented in Figure 3.1.

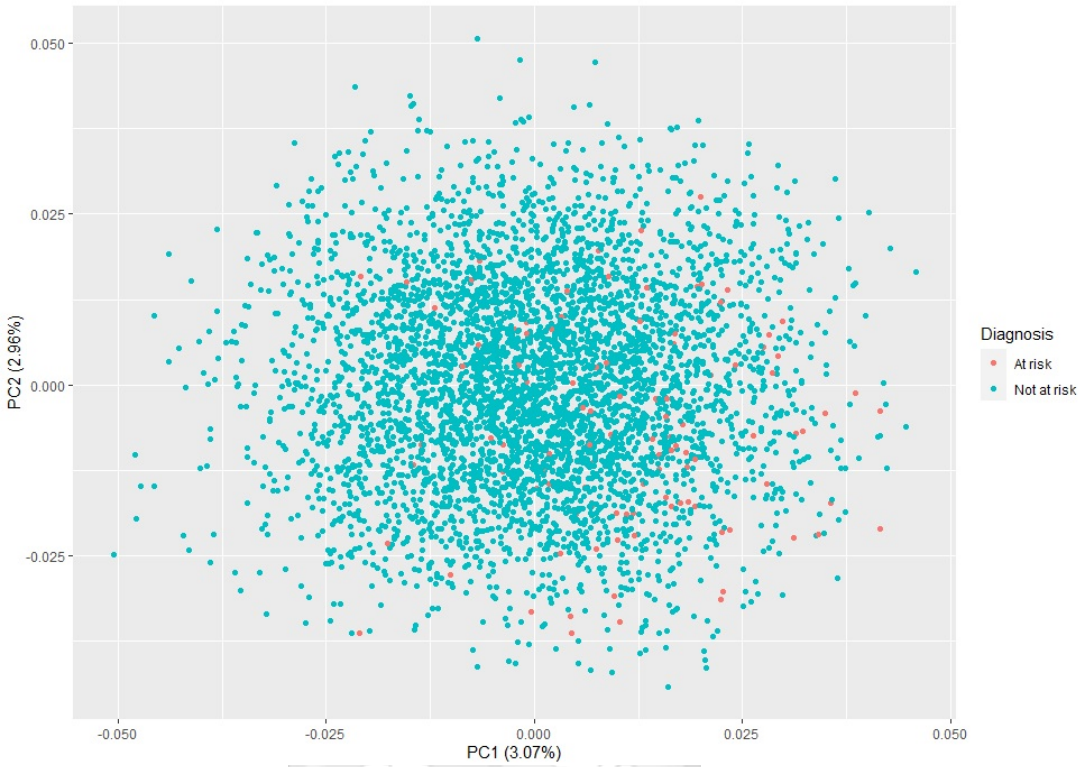


Figure 3.1: Distribution of Scores from Assessment: *A sample complex system*

The chaos theory guarantees that despite the seemingly unrelated distributed of scores of children *'At risk'* relative to those of children *'Not at risk'*, there exist a pattern that can be learned by training a robust statistical model. Clearly, the two groups are not linearly separable and the overlaps observed between them is one of classic examples of complex natural systems, chaos.

The concept of nearby trajectories diverging such as linear, geometric and exponential growth rate plays an important role in describing chaotic systems (Gardini et al., 2020). For binary classification tasks of this nature, trajectories can be described using sensitive dependencies to initial conditions (SDIC), such that the current state x_t is conditional on the previous state x_{t-1}. The non-linear distribution of scores presented in Figure 3.1 can be described in form of global Lyapunov exponents (Bruijn et al., 2012). SDIC is a stochastic behaviour of a system with sufficiently large number of variables. For instance, a construct measured using a psychometric tool with many variables has inherent SDIC. Additionally, an increasing number variables (dimensionality) is characterised by numerical complexity, hence, the need for machine computation.

To contextualize the application of global Lyapunov exponents in binary classification of non-linear system as applied in this study, we consider the first-order, ordinary differential equation (ODE) system $d\boldsymbol{x}/dt = \boldsymbol{F}(\boldsymbol{x})$ and suppose that $\boldsymbol{x}^*$ is an arbitrary initial state such that $\boldsymbol{F}(\boldsymbol{x}^*) = 0$. Further, we suppose that the presence or absence of a condition of interest (say *At risk of* LD or otherwise) can be described by a propagator $\boldsymbol{J}(\boldsymbol{x}(t))$, a function that evolves trajectories $\boldsymbol{x}(t)$. The behaviour of trajectories near $\boldsymbol{x}^*$ can be studied by considering $\boldsymbol{x}(t) = \boldsymbol{x}^* + \varepsilon(t)$, where $\varepsilon(t)$ is an infinitesimal perturbation to every component of $\boldsymbol{x}$. Substituting back into $\boldsymbol{F}$ and expanding to first order in $\varepsilon(t)$ (considering only the perturbations at $t = 0$ and dropping the explicit dependence on t from $\varepsilon(t)$) yields

$$\boldsymbol{F}(\boldsymbol{x}^* + \varepsilon) = \boldsymbol{F}(\boldsymbol{x}^*) + \boldsymbol{J}(\boldsymbol{x}^*)\varepsilon + O(\varepsilon^2), \tag{3.1}$$

(Hansen et al., 2018)

where $\boldsymbol{J}(\boldsymbol{x})$ is the $n \times n$ Jacobian matrix of partial derivatives of $\boldsymbol{F}$ evaluated at the point $\boldsymbol{x}$. The time for time dependence of the perturbation of $\boldsymbol{x}$ can be obtained by

$$\frac{d\varepsilon}{dt} = \boldsymbol{J}(\boldsymbol{x}^*)\varepsilon + O(\varepsilon^2). \tag{3.2}$$

(Ma & Xie, 2018)

Further, consider a linear stability analysis results if we neglect terms of ε^2 or higher powers in equation 3.2. It follows that, if ε is a real-valued vector and $\boldsymbol{J}$ a real-valued matrix, a solution of the form $\varepsilon = \lambda e^{st}$ can be assumed, and equation 3.2 reduces to the eigenvalue equation

$$\boldsymbol{J}\lambda = s\lambda. \tag{3.3}$$

(Abdeljabbar, 2021)

If we define a binary outcome $\boldsymbol{y^i}, i \in [0, 1]$, linear stability analysis results can be used to characterize Lyapunov exponents for non-linear systems. For the initial condition $\boldsymbol{x}(0)$ for ODE defined in equation 3.1 and an infinitesimal displacement from $\boldsymbol{x}(0)$ in the direction of some tangent vector, $\boldsymbol{y}(0)$, the evolution of $\boldsymbol{y}$ according to equation 3.2 is given by

$$\frac{d\boldsymbol{y}}{dt} = \boldsymbol{J}(\boldsymbol{x}).\boldsymbol{y}, \tag{3.4}$$

(Hansen et al., 2018)

valid for only an infinitesimal neighbourhood about $\boldsymbol{x}(0)$. In this regard, supervised machine learning techniques for classification can be used to train a model for diagnosis of LD.

3.4 Dimensionality reduction

Dimensionality describes the number of input features in a dataset. Thus, high dimensional data have more input variables while low dimensional data have few input variables. As earlier described, the dataset used in this study has 88 input variables and 1 response variable (diagnosis). According to Fusi et al. (2016), having a high dimensional feature space drastically reduces performance of machine learning models due to a problem described as permutationally invariant potential energy (Braams & Bowman, 2009), of high dimensional data. As a result, in cases of high dimensional data, it may be desirable to reduce the dimensionality as a means of minimizing model failure rate attributable to high dimensionality.

The established dimensionality reduction techniques include feature selection methods, matrix factorization methods (based on linear algebra) such as Principal Component Analysis (PCA),manifold learning (based on mapping functions), and autoencoder methods (based on neural networks). Fusi et al. (2016) observed that there is no single best way for dimensionality reduction, instead, the application domain and the model chosen dictates the best method based on best results within a given context. The linear algebra methods of dimensionality reduction assumes that the input features are from

the same statistical distribution. Given that the data is drawn from a strictly defined scale $[1, 2, 3]$, this study adopted matrix factorization method, PCA.

Further still, it was described earlier that the attribute set of the data used is made up of 7 sub-scales namely; cognition (17 test items), listening (7 test items), speaking (9 test items), reading (14 test items), writing (14 test items), spelling (7 test items), and Mathematics (20 test items). In this study, it was assumed that each test item is equally weighted (ω_{ij}) for every i^{th} item and j^{th} individual. The score S for each sub-scale (or domain) is obtained by equal weighted strategy defined by Malladi & Fabozzi (2017) as;

$$S = \sum_{i=1}^{n} \omega_i \gamma_i. \tag{3.5}$$

Clearly, the total scores across the seven domains vary significantly due to the fact that they have different number of test items. For instance, spelling has 7 test items while Mathematics has 20 items. In this case, the total score for Mathematics may be significantly higher compared to the total score in spelling, not as a result of better or worse score in any, but as a result of dimension difference. If comparability is desirable, data transformation may become inevitable.

3.5 Feature scaling

Feature scaling has been described as an important pre-processing task in machine learning as it is required to learn an accurate classifier (Bollegala, 2017). Numerous feature scaling (*data transformation*) techniques exist in literature, popular among them being data normalization, data standardization and logarithmic transformation. While each of these techniques have different Mathematical rigour underpinning their applications, there is consensus that they are meant to unify the units of measurements among different features. Without feature scaling, there is a tendency to give more weight to input variables that have higher values (due to nature of their measurement) and give less weight to variables that have less values. Feature scaling therefore plays a critical role in providing a *level ground* for all input variables in the model.

Bollegala (2017) argues that feature scaling can significantly improve the performance of some machine learning algorithms, but may not have the same effects in other algorithms. There is, however, limited evidence that feature scaling negatively affects the performance of a machine learning model provided plausible assumptions are in place. Gradient descent-based machine learning algorithms such as logistic regression, neural network and linear regression require data scaling for easier convergence of the parameter θ_j to the minima;

$$\theta_j \coloneqq \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)}. \tag{3.6}$$

The α value being the gradient, h_θ the learning rate, $x^{(i)}$ the training data points and $y^{(i)}$ class labels. When the input features x are updated at almost the same rate, the gradient descent moves more smoothly towards the minima. This is made possible by data transformation to stabilize the mean and variance across variables in the model. Further-still, distance-based algorithms such as support vector machines (SVM), k-nearest neighbour (k-NN), and $K-$means also require feature scaling because they rely on distances $(e.g., Euclidean)$ between points to establish their similarities. Tree-based algorithms such as decision trees are insensitive to feature scaling (Le et al., 2020).

Given a wide range of feature scaling techniques, there is limited general guidelines on which of the available techniques is the best. However, there seems to be a debate between choosing whether to standardize or normalize data. According to Huang et al. (2019), data normalization involves shifting the original data and rescaling it such that the new value $\acute{X}$ is between 0 and 1. Accordingly;

$$\acute{X} = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{3.7}$$

Huang et al. (2019) further states that data standardization involves rescaling the data such that the new value is centred around the mean with a unit standard deviation. Accordingly;

$$\acute{X} = \frac{X - \mu}{\sigma}. \tag{3.8}$$

As a general guide, data standardization is preferred when the data is assumed to follow a Gaussian distribution, while data normalization is preferred when no particular distribution is assumed. In this study therefore, there was no reason to assume a particular distribution, and hence, data normalization was chosen as a feature scaling technique of choice.

3.6 Model selection

The process of model selection is aimed at choosing the most suitable models from among a pool of candidate models given data. Available literature suggest many different approaches to model selection, however, the context of the specific study or the application domain takes prevalence in most studies. Sarker et al. (2019) used benchmark model where the performance of each candidate model was compared against the performance of ZeroR model. All models whose performance was poorer than the ZeroR model were dropped while those whose performance was better than the ZeroR models were selected and included in their study. A more recent study by Portet (2020) used Akaike Information Criterion (AIC) as their in model selection. Other practitioners prefer that an analysts should choose their favourite model and only tune it by introducing appropriate hyperparameters that suit the situation.

In literature each classification model has a unique Mathematical grounding underneath, however, classification models can be categorised into groups as follows; (1) Gradient Descent Based Algorithms which include logistic regression and artificial neural network, (2) Distance-Based Algorithms in which k-nearest neighbour, k-means and support vector machines belong and (3) Tree-Based Algorithms in which decision trees and random forest belong. In this study, eight models were selected to based on their Mathematical foundations so that even the basic single ruled models such as Naive Bayes and the complex machine learning models such as artificial neural network are

represented. In testing the performance of the model, previous researchers have had it as a practice to nominate a benchmark model. However, in this study, all models were tested and their classification efficiency tested using a set of seven evaluation metrics in order to identify a model that would be considered a base learner for classification of imbalanced data.

3.7 Data analysis

This thesis research sought to analyse the performance of standard classification models on imbalanced data and identify the best solution to tackle the imbalanced data problem in machine learning with application to secondary data on learning disabilities. To achieve this, data analysis was conducted in a systematic way as follows. First, different models were compared and the best one identified as the base learner. Different sampling methods were applied to the data and their performance on rebalancing the data measured. This was used to select the best sampling technique that tackles the problem of imbalanced data. The base learner model was then run on the same data twice, in the first instance after applying the best sampling technique identified and the second time after applying adaptive boosting. The classification accuracy of the model was compared in each case.

3.7.1 Testing classification models

To test classification performance of different models, we compared their accuracy, detection rate, sensitivity, specificity, F-measure and G-measure recall, F-measure and the Kappa values of each model. All these evaluation metrics were deduced from the confusion matrix as follows;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

$$\text{Detection rate} = \frac{TP}{TP + TN + FP + FN} \tag{3.10}$$

$$\text{Failure rate} = 1 - \text{Detection rate} \tag{3.11}$$

$$\text{Sensivity} = \frac{TP}{TP + FN} \tag{3.12}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.13}$$

$$\text{F-measure} = 2 \times \frac{\text{Specificity} \times \text{Sensitivity}}{\text{Specificity} + \text{Sensitivity}} \tag{3.14}$$

$$\text{G-measure} = \sqrt{\text{Specificity} \times \text{Sensitivity}} \tag{3.15}$$

Accuracy answers the question,overall, how often is the classifier correct?. Model sensitivity of recall is, when it's actually positive LD case, how often does it predict positive? Model specificity describes, when it's actually child with no LD, how often does it predict so? Detection rate describes, how often does the positive cases occur in the sample under study. F-measure is harmonic mean of specificity and sensitivity while G-measure is there geometric mean.

Classification task in this thesis research started by splitting the data into training and testing dataset using caTools library package in R software. The training dataset was 75% of the data while the remaining 25% of the data was used for testing. To ensure uniform samples for each model and promote reproducibility of the process, 1 was used as a random seed $set.seed(1)$ for each model. Implementation of machine learning models requires specific R packages. Necessary R packages were installed and loaded into the working environment.

For exploratory data analysis, visualization packages such as $ggplot2$ were used. Additionally *PerformanceAnalytics* and *tidyquant* packages were used in computation of skewness and kurtosis which are not included as default functions in base R

software. Machine learning packages that were used in this study include *caret*, *e1071*, *randomForest*,and *neuralnet* among others.

3.7.2 Testing sampling techniques and boosting algorithm

To test the ability of sampling methods in tackling the problem of imbalanced data, Receiver Operating Characteristic's (ROC's) area under the curve (AUC) was used. In this study, it is assumed that a base learner model that has good classification performance on imbalanced data will in fact do much better when the data is fairly balanced. The base learner was identified through experimentation on eight classification models.

The data used in this study had 4,473 observations out of which only 108 were identified as positive cases for LD (the minority class), giving an imbalance ratio of 1:42. It is hoped that the sampling techniques considered in this thesis research such as random undersampling (RUS), random oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) will reduce this imbalance ration in the data. Logically, a sampling technique that reduces the imbalance ratio significantly enable the base learner to classify the two classes more effectively.

The ROC curve is created by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis at various threshold settings. Since the emphasis in the case of imbalanced classification problems is placed particularly on the predictive accuracy of minority class while retaining accuracy for the majority class, AUC on the ROC was used as the indicator of such performance. In the previous sub-section 3.6.1, it was emphasized that a random seed *set.seed(1)* was used to draw the sample that was used to compare performance of different classification models. However, for testing efficiency sampling techniques in tackling imbalanced data problem, three different seeds were used for sampling. The purpose of varying the sampling seeds in R was to draw independent samples from the data for comparison.

Accordingly, three independent samples were drawn from the data as follows; First, the sampling ratio for model training and testing was retained as 75% and 25% respectively for all the samples, hence equal in size, all drawn from full dimensional data. **sample 1**

was drawn by setting *set.seed (2)*, **sample 2** by setting *set.seed (3)*, and **sample 3** by setting *set.seed (4)*. The sampling methods were only applied to the training sets of each of the three independent samples. The model was then run using testing dataset for each of the three randomly and independently drawn sample, and corresponding AUC computed. The results were tabulated. Chi-square test was used to test whether the difference in performance between the best and second best sampling method was statistically different at $(1 - \alpha)$ confidence level, $(\alpha = 0.05)$.

3.7.3 The boosting algorithm

Probably Approximately Correct (PAC) learning framework is a core theoretical foundation to the development of statistical boosting methodologies (Cullina et al., 2018), whose emergence can be traced in 1990s. A breakthrough in boosting methods is the invention of adaptive learning algorithm 'AdaBoost' that works on finite training data and has controllable computational complexity because of its nature as a iterative algorithm (Y.-F. Chen et al., 2016). Given the wide application of statistical boosting, different types of boosting may be most applicable in some specific contexts than others. For instance, extreme gradient boosting (XGBoost) perform extremely well in cases where there is either too much noise in the data and/or missing values. CatBoost on the other hand performs best when input feature space is made of categorical variables.

Further, it should be noted that the initial ideas in machine learning for boosting was to improve classification of already existing weak classifiers by attempting simultaneously reduce model bias and variance. However, statistical developments have led to the extension of application of some types of boosting such as AdaBoost to correcting skewed class imbalance. The paradigm shift from machine learning to statistical modelling and application of boosting is motivated by the acknowledgement that boosting is not algorithmically constrained (Lee et al., 2020).

Therefore, this study applied adaptive boosting (AdaBoost) algorithm, which is used to tackle the problem of imbalanced data by introducing weighting recessively. Adaboost is the first practical boosting algorithm, developed by Freund and Schapire (1996), and achieves huge success in a wide range of applications according to Cullina et al. (2018). This thesis research describes and apply Adaboost algorithm for the binary

classification problem due to Freund and Schapire (1997). Suppose that $X \in \prod$ and $Y \in \Omega$ are predictor space and response space respectively. In binary classification, the response variable $Y \in \{-1, 1\}$ and $X \in \mathbb{R}^d$ is the predictor vector. A classifier is a function $G : \mathbb{R}^d \to \{-1, 1\}$. And a base classifier satisfies $\mathbb{E}[\mathbb{I}(G(X) \neq Y)] < 1/2$ where (X, Y). The error rate of a classifier $G_m(x)$ is

$$err_m = \frac{1}{M} \sum_{i=1}^{N} \mathbb{E}[\mathbb{I}(G(X) \neq Y)], \tag{3.16}$$

for $m = 1, 2 \cdots, M$. For Adaboost, we only consider the classifiers

$$G(x) = sgn(F(x)), \tag{3.17}$$

where $F(x)$ is a real function standing for the classier boundary.
In Adaboost, the base learning algorithm trains M classifiers using the filtered sample data in each iteration and then the weighted sum of all M classifiers is the strong classier we want. The ultimate classifier boundary takes form

$$F(x) = \sum_{m=1}^{M} \alpha_m F_m(x), \tag{3.18}$$

where $F(x)$ is the corresponding classifier boundary curve of $G_m(x)$ and α_m is the classifier weight. The classifier boundary is computed from the training data.

3.7.4 Comparing the best sampling technique boosting algorithms

To compare the application of sampling technique and boosting approaches in tackling the problem of imbalanced data, the best classification model termed an base learner was then re-run on the same datasets when the best sampling technique has been applied and in a different instant when the AdaBoost algorithm has been applied. The model classification accuracy, detection rate, failure rate, sensitivity,specificity, F-measure and G-measure will be computed. To this end, it would be possible to compare model classification accuracy when no class rebalancing has been applied and when class rebalancing has been applied using the best sampling technique and the AdaBoost algorithm.

3.8 Ethical Considerations

The LDDI-2019 data was collected by Kenya Institute of Special Education (KISE) as part of collating empirical evidence to developing a culturally responsive learning disabilities diagnostic inventory for Kenya. The candidate sought permission from the institute to access and use LDDI-2019 data for academic purposes only, which was granted on 6^{th} July 2020, see (*see Appendix 2*). Further, the candidate conducted research in compliance with national regulations on ethical considerations in research in line with Strathmore University's code of responsible conduct in research. Accordingly, this research work was reviewed and approved by Strathmore University's Institutional Ethics Review Committee (SU-IERC), reference number SU-IERC1039/21, see (*see Appendix 4*) and (*see Appendix 5*)

CHAPTER FOUR

4 RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter presents the results of the study, and statistical discussions with implication to the best practices in diagnosis of learning disabilities. The chapter begins with exploratory analysis which presents two general data preprocessing result; dimensionality reduction and feature scaling. The chapter then proceeds to present the findings of the study guided by the objectives; comparing performance of standard classification models, comparing performance of sampling methods in correcting imbalanced data and Adaptive Boosting (AdaBoost) solution for imbalanced data. The chapter ends by analysing the effectiveness of the best sampling methods and the AdaBoost solutions for imbalanced data.

4.2 Exploratory Analysis

Descriptive analysis forms a foundation of inductive research which is concerned with discovery, exploration, and empirically detecting phenomena in data (Jebb et al., 2017). Exploratory analysis adopted in this study blends descriptive summaries and visual displays of the scores obtained by children. Table 4.1 presents numeric summaries of the data such as mean, median, standard deviation, skewness and kurtosis. The mean and median shows the measures of central tendency of the data. Some domains such as Maths and Cognition have higher mean scores compared to others such as spelling and speaking. Standard deviation measures dispersion in scores in each skill area, skewness is measure of symmetry while kurtosis is a measure of peakedness. For a normally distributed data, skewness and kurtosis values ought to be zero, with negative values indicating fat tails while positive values showing high peakedness in the data. As shown in Table 4.1, these values vary slightly in the neighbourhood of zero, suggesting normality. Shapiro-Wilk test for normality was used and the p-values $(p < 0.05)$ confirm that the scores were not normally distributed.

Table 4.1: Summary statistics of actual scores per domain area

Domain	Distribution measures					Shapiro-Wilk test	
	Mean	Median	Std Dev	Skewness	Kurtosis	W	P-value
Cognition	34.00	34.00	3.41	0.05	0.11	0.992	0.001
Listening	13.97	14.00	2.17	-0.03	-0.31	0.981	0.001
Speaking	18.00	18.00	2.51	0.02	-0.16	0.986	0.001
Reading	28.07	28.00	3.14	0.13	-0.03	0.990	0.001
Writing	28.14	27.97	3.13	-0.01	0.02	0.991	0.001
Spelling	14.05	14.00	2.15	-0.04	-0.15	0.981	0.001
Maths	40.12	40.00	3.81	0.12	0.06	0.993	0.001

To further understand the score distribution among the two groups of children *(at risk or not at risk)*, a kernel density graph was plotted and results presented in Figure 4.1.

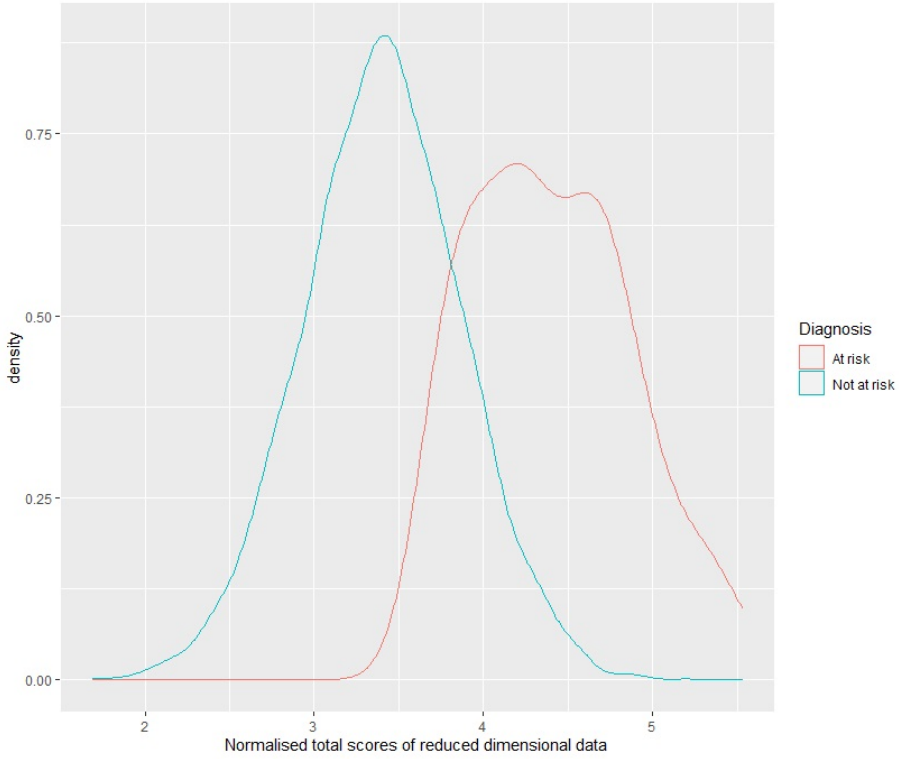


Figure 4.1: Kernel density estimates of normalised scores per group of children

The results shows that the scores of the class of children who are not at risk has a greater kurtosis than normal distribution. The scores of children who are at risk shows fat tails on the right. While the two classes may seem separable on the basis on total scores, there is a significant overlap between the two classes.

Figure 4.2 presents visual display of scores in seven skill areas; cognition, listening, speaking, reading, writing, spelling and maths. As earlier described in Table 4.1, the

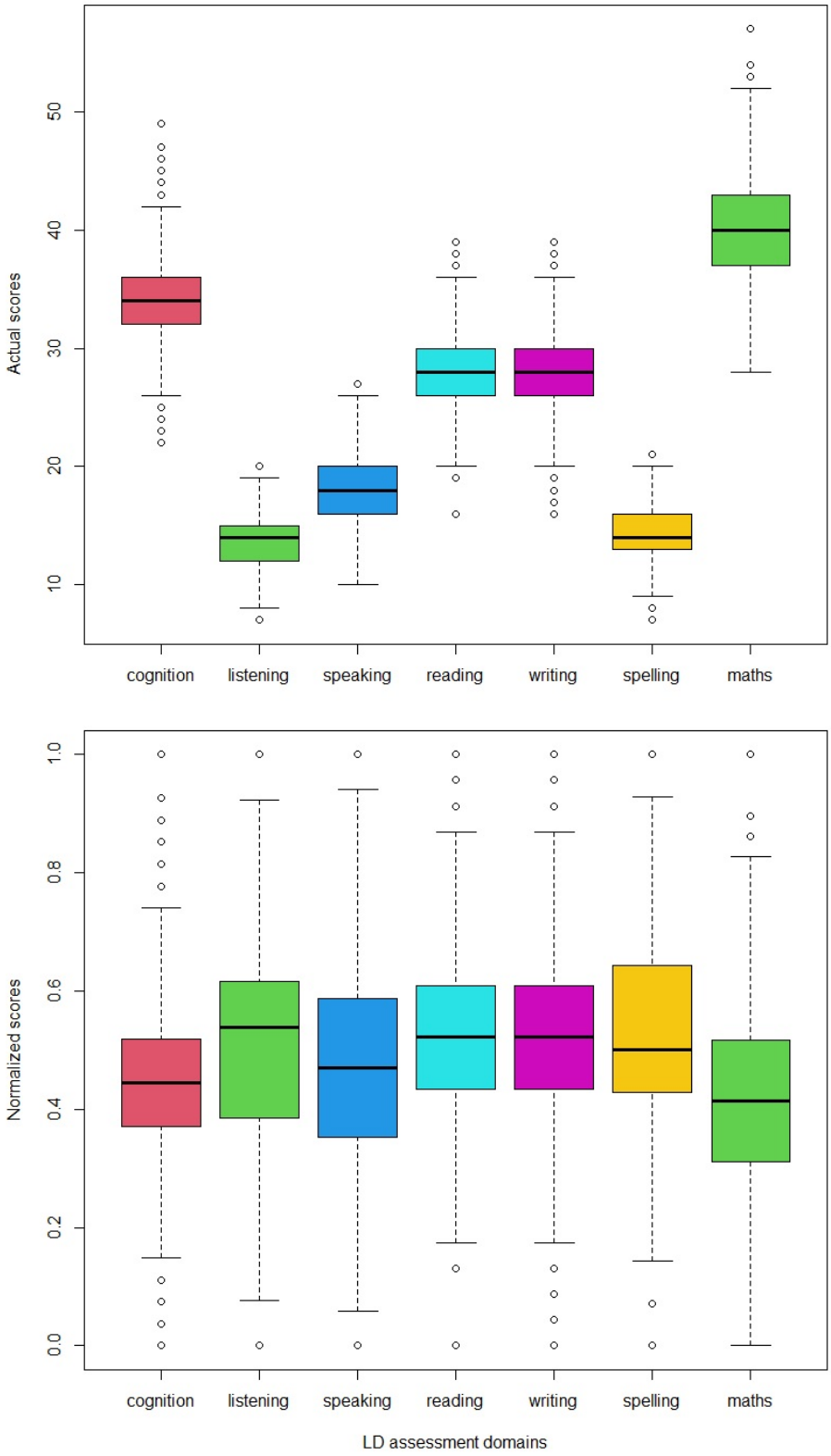


Figure 4.2: Boxplot for actual total scores per domain

average scores of each skill area varies significantly due to the total number of test items used to measure the skill. For instance, 20 items are used to measure maths skills, hence the average score is expected to be higher than spelling which has only 7 test items. As shown by the '*Actual score*' axis in Figure 4.2, the average scores for each skill area varies, highest being maths followed by cognition, while spelling and listening being the lowest.

According to Bollegala (2017), when variables vary significantly in magnitude, it is a good idea to scale the variables as part of preprocessing for building machine learning models. While our exploratory analysis suggest that variables, are approximately normally distributed, data normalization discussed in section 3.4 under feature scaling was conducted with the aim of minimizing the variation among the attribute set, and improve comparability. The boxplot shown by the '*Normalized score*' axis presents a summary of scaled scores. Clearly, the variation among the scores in different skill area was significantly minimized. However, general properties of the data is retained scaling. For instance, cognition has heavy tails in both the actual and scaled data space.

4.3 Model Classification Performance on Imbalanced Data

In binary classification problem, the model can either correctly or incorrectly classify a new observation. For this reason, a confusion matrix can be used to gain insights on the performance of a model for binary classification task. A confusion matrix is useful in computing model recall or sensitivity, specificity, accuracy and precision. In addition, other performance indicators such as F-measure, G-measure, model detection rate and misclassification error can be derived. In this study, model accuracy, detection rate, failure rate, sensitivity, specificity, F-measure and G-measure were computed for eight classification models; Naive Bayes (NB), Decision Trees (DT), Linear Discriminant Analysis (LDA), Logistic Regression (LR), k-nearest neighbour (k-NN), Artificial Neural Network (ANN), Support Vector Machines (SVM) and Random Forest (RF).

4.3.1 Model Performance Based on Full Dimensional Data

This section presents results of classification model performance when full dimensional data is used. As earlier described in Section 3.4 of this thesis, a full dimensional data refers to the initial dataset with 88 input features from across all the seven domains, without any statistical treatment. Table 4.2 shows performance analysis based on training dataset while Table 4.3 shows performance analysis based on testing dataset. Table 4.4 shows marginal changes in model performance. The marginal changes describes the increase or decrease in individual evaluation metric score of each classification model on testing dataset with reference to scores obtained from training dataset as a baseline. It can deduced from the theory of generalization (Kawaguchi et al., 2018), that a robust learning model should have minimal error in the out of sample performance.

Table 4.2: Model performance based on full dimensional training dataset

Evaluation metric	**NB**	**DT**	**LDA**	**LR**	**kNN**	**ANN**	**SVM**	**RF**
Accuracy	0.98	0.98	**0.99**	0.98	0.98	0.98	**0.99**	0.99
Detection rate	0.39	0.39	0.63	0.53	0.39	**0.80**	0.78	0.39
Failure rate	0.61	0.61	0.37	0.47	0.61	**0.20**	0.22	0.61
Sensitivity	0.40	0.34	0.62	0.52	0.39	0.36	**0.85**	0.41
Specificity	**1.00**	0.99	**1.00**	0.99	0.99	0.89	0.89	**1.00**
F-measure	0.57	0.51	0.76	0.68	0.56	0.52	**0.87**	0.58
G-measure	0.63	0.58	0.78	0.72	0.62	0.57	**0.87**	0.64

Table 4.3: Model performance based on full dimensional testing dataset

Evaluation metric	**NB**	**DT**	**LDA**	**LR**	**kNN**	**ANN**	**SVM**	**RF**
Accuracy	0.97	0.97	**0.99**	0.98	0.96	0.97	**0.99**	0.97
Detection rate	0.00	0.35	0.56	0.47	0.00	**0.71**	0.69	0.03
Failure rate	1.00	0.65	0.44	0.53	1.00	**0.29**	0.31	0.97
Sensitivity	0.00	0.31	0.56	0.47	0.00	0.33	**0.77**	0.03
Specificity	**1.00**	0.99	**1.00**	0.99	0.99	0.96	0.98	**1.00**
F-measure	0.00	0.47	0.72	0.64	0.00	0.49	**0.86**	0.06
G-measure	0.00	0.55	0.75	0.68	0.00	0.56	**0.87**	0.18

As presented in Table 4.4 below, there were minimal changes in the accuracy and specificity of classification model for both training and testing dataset. This implies that there were negligible changes in the said evaluation metrics when applied on either training or testing dataset. On the other hand, significant changes were observed in the model detection rate, failure rate and sensitivity.

Table 4.4: Marginal changes based on full dimensional dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	-0.01	0.01	0.00	0.00	-0.02	-0.01	0.00	-0.01
Detection rate	-1.00	-0.09	-0.12	-0.12	-1.00	-0.12	-0.12	-0.92
Failure rate	0.63	0.06	0.20	0.13	0.63	0.47	0.41	0.58
Sensitivity	-1.00	-0.09	-0.09	-0.09	-1.00	-0.09	-0.09	-0.93
Specificity	0.00	0.00	0.00	0.00	0.00	0.08	0.10	0.00
F-measure	-1.00	-0.07	-0.06	-0.06	-1.00	-0.05	-0.01	-0.90
G-measure	-1.00	-0.05	-0.04	-0.05	-1.00	-0.01	0.00	-0.72

These results shows that accuracy and specificity of a classification model are least responsive to the minority class while detection rate and sensitivity or model recall are most responsive metrics for minority class when the data is highly imbalanced. On specific models, the results shows that k-nearest neighbour (kNN) and Naive Bayes (NB) showed the highest reduction in both detection rate and sensitivity followed by the random forest (RF) model. On the other hand, decision trees (DT) model shows the least change in detection rate and sensitivity between training and testing dataset. This was closely followed by linear discriminant analysis (LDA) model, logistic regression (LR) model, artificial neural network (ANN) and support vector machine (SVM) models.

4.3.2 Model Performance Based on Reduced Dimensional Data

This section presents results of classification model performance based reduced dimensional data. As earlier described in Section 3.4 of this thesis, reduced dimensional dataset refers to a condensed dataset with 7 input feature space, each corresponding to a sum scores for a particular sub-scale (or domain areas) namely cognition (17 test items), listening (7 test items), speaking 9 (test items), reading (14 test items), spelling (7 test items), writing (14 test items) and Mathematics (20 test items). Data reduction from 88 test items into 7 raw scores was based on the conventional practice that the test items are equally weighted (ω_{ij}) for every i^{th} instance and j^{th} individual. The score S for each sub-scale (or domain) was obtained by equal weighted strategy defined by Malladi & Fabozzi (2017), refer to Equation 3.5 in Section 3.4. Table 4.5 shows performance analysis based on training dataset while Table 4.6 shows performance analysis based on testing dataset. For a robust learning model, it is expected that the error in the out of sample performance relative to in-sample performance be as minimal as possible

(Kawaguchi et al., 2018).

Table 4.5: Model performance on reduced dimensional data training dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Detection rate	**0.79**	0.45	0.50	0.53	0.45	0.45	0.78	0.45
Failure rate	**0.21**	0.55	0.50	0.47	0.55	0.55	0.22	0.55
Sensitivity	0.69	0.39	0.44	0.47	0.39	0.39	**0.83**	0.39
Specificity	0.89	0.99	**1.00**	0.99	0.99	0.89	0.89	**1.00**
F-measure	0.78	0.56	0.61	0.64	0.56	0.54	**0.86**	0.56
G-measure	0.78	0.62	0.66	0.68	0.62	0.59	**0.86**	0.62

Table 4.6: Model performance on reduced dimensional data testing dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	0.97	**0.98**	**0.98**	**0.98**	0.97	**0.98**	**0.98**	**0.98**
Detection rate	**0.70**	0.35	0.44	0.47	0.00	0.24	0.69	0.37
Failure rate	**0.30**	0.65	0.56	0.53	1.00	0.76	0.31	0.63
Sensitivity	0.69	0.28	0.44	0.47	0.00	0.36	**0.83**	0.38
Specificity	0.98	0.99	**1.00**	0.99	0.99	0.97	0.98	**1.00**
F-measure	0.81	0.44	0.61	0.64	0.00	0.53	**0.90**	0.55
G-measure	0.82	0.53	0.66	0.68	0.00	0.59	**0.90**	0.61

The results shows that model accuracy and specificity are invariant for training and testing datasets for highly imbalanced data. On the other hand, model detection rate, failure rate and sensitivity are highly variant evaluation metrics between training and testing datasets. For instance, it was observed that model detection rate and sensitivity reduced for testing dataset compared to training dataset and model failure rate increased. Table 4.7 shows marginal changes in model performance. The marginal changes describes the increase or decrease in individual evaluation metric score of each classification model on testing dataset with reference to scores obtained from training dataset as a baseline. Comparing the variation in model evaluation metrics conducted on full and reduced dimensional data, it was observed that there were significantly lower variations in reduced dimensional data. Such observations may be used to conclude that higher dimensionality in the data feature space increases chances of model failure.

On specific model performance, it was found that k-nearest neighbour (kNN) recorded the highest reduction in model detection rate followed by artificial neural network (ANN) and decision tree (DT) models respectively. A plausible reason why ANN and DT exhibit this behaviour is their reliance on instance-based while kNN used

Table 4.7: Marginal changes based on reduced dimensional dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
Detection rate	-0.12	-0.22	-0.12	-0.12	-1.00	-0.46	-0.12	-0.17
Failure rate	0.44	0.18	0.11	0.13	0.81	0.38	0.41	0.14
Sensitivity	0.00	-0.28	0.00	0.00	-1.00	-0.07	0.00	-0.02
Specificity	0.10	0.00	0.00	0.00	0.00	0.09	0.10	0.00
F-measure	0.04	-0.21	0.00	0.00	-1.00	-0.02	0.05	-0.02
G-measure	0.05	-0.14	-0.01	0.00	-1.00	0.00	0.05	-0.02

distance-based approaches to model classification rules. Such classification algorithms tend to perform better when the feature space provides flexible degrees of freedom. Despite high deterioration in detection rate by ANN, the model still had a significantly minimal reduction in model sensitivity or model recall compared to kNN and DT. Further, kNN recorded the highest failure rate between training and testing dataset results for reduced dimensional data. A plausible reason for this observation could be that data reduction by equal weighted summation masks off essential relationship within the structure of the original data.

4.3.3 Model Performance Based on Normalized Scores

This section presents results of classification model performance based normalized dataset with 7 input feature space each corresponding to normalized scores of each sub-scale (or domain area). As earlier described in Section 3.5 of this thesis, feature scaling is a technique to standardize the independent features present in the data in a fixed range. While there are many approaches to feature scaling, this study was based on data normalization of domain scores, preferred over data standardization owing to its non-parametric assumptions about data distribution, refer to Equation 3.7 and Equation 3.8 in Section 3.5. Table 4.8 shows performance analysis based on training dataset while Table 4.9 shows performance analysis based on testing dataset.

Table 4.8: Model performance on normalized scores of training dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Detection rate	**0.79**	0.47	0.50	0.53	0.47	0.47	0.78	0.50
Failure rate	**0.21**	0.53	0.50	0.47	0.53	0.53	0.22	0.50
Sensitivity	0.69	0.40	0.40	0.47	0.40	0.40	**0.83**	0.40
Specificity	0.89	0.99	**1.00**	0.99	0.99	0.89	0.89	**1.00**
F-measure	0.78	0.57	0.58	0.64	0.57	0.56	**0.86**	0.58
G-measure	0.78	0.63	0.64	0.68	0.63	0.60	**0.86**	0.64

Table 4.9: Model performance on normalized scores of testing dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	0.97	0.98	**0.98**	**0.98**	0.97	**0.98**	**0.98**	**0.98**
Detection rate	**0.70**	0.35	0.44	0.47	0.00	0.35	0.69	0.44
Failure rate	**0.30**	0.65	0.56	0.53	1.00	0.65	0.31	0.56
Sensitivity	0.69	0.31	0.44	0.47	0.00	0.41	**0.83**	0.44
Specificity	0.98	0.99	**1.00**	0.99	0.99	0.97	0.98	**1.00**
F-measure	0.81	0.44	0.61	0.64	0.00	0.58	**0.90**	0.61
G-measure	0.82	0.53	0.66	0.68	0.00	0.63	**0.90**	0.66

The structure of variation in model performance between in Sub-Section 4.3.2 for reduced dimensional data and Sub-Sections 4.3.3 for normalized data is similar, yet different from the full dimensional data structure described in the earlier Sub-Section 4.3.1. Full dimensional data was characterised by large variations between training and testing datasets compared to reduced dimensional data with actual raw scores and normalized scores. Nonetheless, there are minor variations in the structure of reduced dimensional actual scores and normalized scores as presented earlier in Table 4.7 and in Table 4.10 below.

Table 4.10: Marginal changes based on normalized scores dataset

Evaluation metric	NB	DT	LDA	LR	kNN	ANN	SVM	RF
Accuracy	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
Detection rate	-0.12	-0.26	-0.12	-0.12	-1.00	-0.26	-0.12	-0.12
Failure rate	0.44	0.23	0.11	0.13	0.90	0.23	0.41	0.11
Sensitivity	0.00	-0.23	0.09	0.00	-1.00	0.02	0.00	0.09
Specificity	0.10	0.00	0.00	0.00	0.00	0.09	0.10	0.00
F-measure	0.04	-0.23	0.06	0.00	-1.00	0.04	0.05	0.06
G-measure	0.05	-0.16	0.04	0.00	-1.00	0.05	0.05	0.04

The results shows that most of the evaluation metrics such as model accuracy and specificity remain the same for both actual scores data and normalized scores data.

However, it was observed that model detection rate become worse by an average margin 2% for normalized data compared to actual data scores. Similarly, model sensitivity also become worse by an average of 3% for normalized data compared to to actual data scores. While these variations may be considered negligible in magnitude, the consistency with which they appear across many classification models may be an indication that data transformation is not necessarily a great idea in modelling imbalanced data problems, unless context specific justification is unquestionable.

Naive Bayes (NB) Classifier

The results show that naives Bayes model performs poorly when the classification task has a large number of input feature variables as shown in Table 4.3. When applied to full dimensional data will 88 input variables, Naive Bayes models fails to learn or recognize patterns between the two groups in the data. A plausible reason to this observation is a possible violation of Naive Bayes classifier assumption of independence of input variables. It is possible that among the 88 variables, there would be higher correlation among some of them, leading to redundancy in the learning process. A research conducted by Abellán & Castellano (2017) found that classification efficiency of Naive Bayes model could be improved via a quick variable selection method using maximum of entropy.

There is an improvement in the model's detection rate from 0% to 70% as shown in Table 4.6 and Table 4.9 after reducing the input variables from 88 to 7. With reduced dimensionality, Naive Bayes ' classification power increases due to its enhanced ability to select the most informative variables without setting a threshold (Abellán & Castellano, 2017). This study also established that classification efficiency of Naive Bayes model is least affected by data transformation by normalization as shown by similar results in Table 4.6 and Table 4.9. Despite the Naive Bayes model being a simple classifier, it has applications in artificial intelligence especially in areas with limited predictor variables. For instance, Granik & Mesyura (2017) used Naive Bayes to detect fake news on social media with 74% accuracy. The current study, uses several model performance evaluation metrics to ascertain a holistic view of Naive Bayes performance since accuracy as a measure of model efficiency is misleading especially when the data

is highly imbalanced. Further, Naive Bayes model may not guarantee high performance when the attribute in a classification problem is relatively high. Assessment of learning disabilities for instance is associated with a large number of test items, hence, Naive Bayes may not be the best model when considering options to automate classification of assessment process.

Decision Trees (DT) Classifier

The decision tree classifier starts at a random single point node in a dataset and recursively splitting data at every instance until a classification decision is made at the end of the tree (See Figure 4.3). The results in Table 4.3,Table 4.6 and Table 4.9 shows that the detection rate of decision trees model is 35%. This shows that decision trees model's classification power on imbalanced data is low. However, the model accuracy increases from slightly after the number of input variables is reduced from 88 to 7. Similarly, the model sensitivity reduces from 31% as shown in Table 4.3 to 28% in Table 4.6 and 4.9. Consequently, the F and G-measures reduces from 47% and 55% to 44% and 53% respectively.

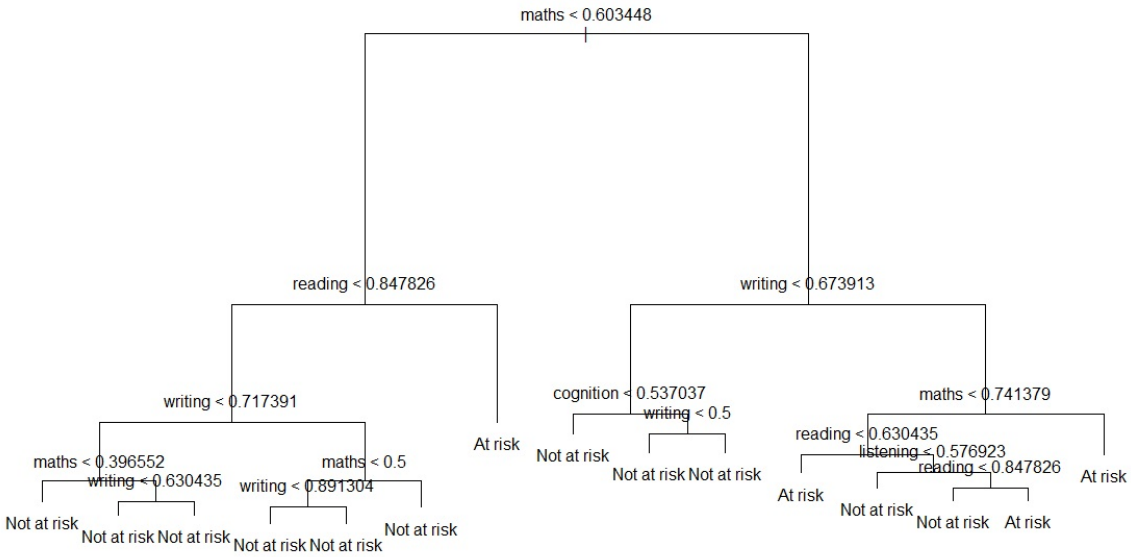


Figure 4.3: Sample decision tree model from diagnostic assessment of LD

The results imply that while dimensionality reduction increases classification accuracy, reduce model sensitivity while model detection rate remains the same on imbalanced data. Previous experiments with decision trees by Kamiński et al. (2018) indicates that decision tree models are sensitive to dimensionality. This is in fact true based on optimization of decision trees based on c4.5 algorithm (Elaidi et al., 2018). According to the c4.5 algorithm, a dataset with n attributes would have 2^{2^n} recursions. On this premise, it can be seen that increase in data dimensionality can increase the number of recursions performed by a decision tree model.

High number of recursions in the decision tree model increases computational complexity, however, in this study it was found that training and testing time for the model was relatively small compared to other models such as ANN. In this study, when decision tree model was applied to a full dimension data with 88 attributes, the model sensitivity was 31% and 28% was applied to a reduced dimension data with 7 attributes, it would be feasible (theoretically) to apply a decision trees uncompressed dataset. However, there is immediate drawback on computational complexity with this approach, since the user has to follow through a longer loop. With this dilemma therefore, decision trees may not be the best option when considering options to automate assessment process. after all, the failure rate of decision tree model for imbalanced data is 65% which is undesirable.

Linear Discriminant Classifier

Linear discriminant analysis is a popular dimension reduction technique. Additionally, LDA is a covariance matrix-based classifier that can be trained as a supervised learning algorithm for classification tasks by maximizing the mean difference between distinct groups in a dataset. The results presented in Table 4.3 shows that model detection rate is 56%, while Table 4.6 and Table 4.9 shows that detection rate at 44%. Further, the results show that LDA classifier has significantly higher model specificity relative to its sensitivity. This shows that while this model has a considerably high detection rate, its classification is biased towards the majority class, hence, the model may not be suitable for classification imbalanced data.

The F- and G- measures remain approximately the same regardless of data dimensionality or transformation used. This is expected because the general properties of LD data is not affected by reducing the data dimensions since all the initial variables are retained. However, the arithmetic combination of variables in a psychometric tool could potentially mask more informative indicators.Gyamfi et al. (2017) contends that under normality and homoscedasticity assumptions, LDA is known to be an optimal classifier in terms of minimising the Bayes error for binary classification. As earlier described in Figure 4.1, the total scores for the two groups in the dataset can be described to be approximately normally distributed, although the scores of the group *At risk* has a slightly lower kurtosis compared to their counterparts *Not at risk*.

Logistic Regression (LR) Classifier

Logistic regression is also a covariance matrix-based classifier just like LDA classifier. Unlike LDA classifier that can be extended into multiclass classification tasks, logistic regression works only for binary classification tasks. The results presented in Table 4.3, Table 4.6 and Table 4.9 indicate that the model's detection rate is 47% regardless of data dimensionality or transformation. In fact, all the evaluation metrics do not change for a logistic regression classifier with different data structures. Plausible reason for this observation is that logistic regression model is based on odds ratio,

$$\text{logit}\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{4.1}$$

which is least affected by data dimensionality and feature scaling. The results show that logistic regression a considerably high failure rate of 53% despite having a high specificity of 99%. The high failure rate of logistic regression model in this study can be described in the context of linearly separable decision boundary between the majority and minority class. In exploratory analysis section particularly Figure 4.1, there is a significant overlap in the scores between the two groups of children. As a result of absence of linear boundary to separate the classes, this model is biased towards the majority class while ignoring the minority class. This is no wonder the model has high model specificity whose computation is highly influenced by the true negatives, which

in this case are the majority. Despite the draw backs of logistic regression classifier, some researchers such as Christodoulou et al. (2019) established that advanced machine learning methodologies may not perform better than logistic regression when limited number of predictors are considered. They claim that these advanced ML approaches are only better at handling complex datasets with voluminous data.

k-Nearest Neighbour (kNN) Classifier

Results show that kNN classification model performs poorly based on detection rate and sensitivity despite having high accuracy and specificity of over 96%. This shows that the kNN model is highly biased towards the majority class and treats the minority class as noise or outliers. Gao & Li (2020) contends that kNN is a distance-based model that determines similarity between objects by measuring distance between them. Given two arbitrary points i and j in a two-dimensional space, the k-NN classifier holds that the rule;

$$R_i = \{x : d(x, x_i) < d(x, x_j), i \neq j\} \tag{4.2}$$

is a primary distance metric undermining Euclidean, Manhattan, Minkowski or Hamming distances in establishing similarities or dissimilarities between points of interest (Gao & Li, 2020). In most cases, kNN is based on Euclidean distance between points, however, Chebyshev, Manhattan and Minkowski distances are also commonly used. The results presented in Table 4.3, Table 4.6 and Table 4.9 show that k-NN classification efficiency is lowest among the eight classification models considered in this study. The model detection rate, sensitivity, F-measure and G-measure are at 0%. The high accuracy and specificity of kNN further advances the argument that the model is biased towards the majority class in an imbalanced data problem.

Amra & Maghari (2017) reported that Naive Bayes model perform better than kNN in prediction of student performance. Some of the reasons cited as being drawbacks of kNN include its failure to handle large sample sizes, high number of dimensions and also being sensitive to outliers in the data. The learning disabilities data used in this study had a large sample and had many input variables. Additionally, most of the total scores of children from the two classes would overlap.Since KNN is a distance-based algorithm, the cost of calculating distance between a new point and each

existing point is very high which in turn degrades the performance of the algorithm (Amra & Maghari, 2017).Again, in higher dimensional space, the cost to calculate distance becomes expensive and hence impacts the performance.

Artificial Neural Network (ANN) Classifier

The results presented in Table 4.3, Table 4.6 and Table 4.9 shows that classification efficiency of artificial neural network is highest in full dimensional data with detection rate at 71%. When the data's dimension is reduced from 88 items to 7, the performance of ANN model drops drastically, with detection rate 24%. This shows that when the data has many input variables, the ANN model is able to learn the subtle underlying structures leading to better classification. However, when the data dimension is reduced by arithmetic summation of each domain area, the model detection rate is reduced. This may be due to masking of important features of within the data.

The finding on classification efficiency in relation to the number of input variable was further evident in pictorial form of the ANN model that was built using full dimensional data and scaled data. Figure 4.4 presents two ANN models and their associated statistics. The ANN models were built using one hidden layer with three nodes each, and a binary classification shown in the output layer. In the full dimensional data with 88 input variables, the model classifies at a small error of 33.01, while in the reduced dimensional data with 7 variables, the model classifies with an error of 725.91, which is relatively big. Given that the ANN configuration is the same, we conclude based on this findings that ANN model performs better when the input variables are many. However, it should be noted that minimizing an error in an ANN model requires heavy computational strength. As shown in Figure 4.4, a small error of 33.01 is associated with very many computational steps of about 99,338. The big error of 725.91 is associated with fewer computational steps of about 46.

Further, the study established that data transformation through normalization improves classification performance of ANN model as shown in Table 4.9. The model detection rate improves from 24% to 35% after data normalization. This implies that data normalization reveals the actual data structure, hence can improve the performance of a machine learning algorithm. Similarly, sensitivity of an ANN model was found to

increase after data scaling from 36% to 41%.

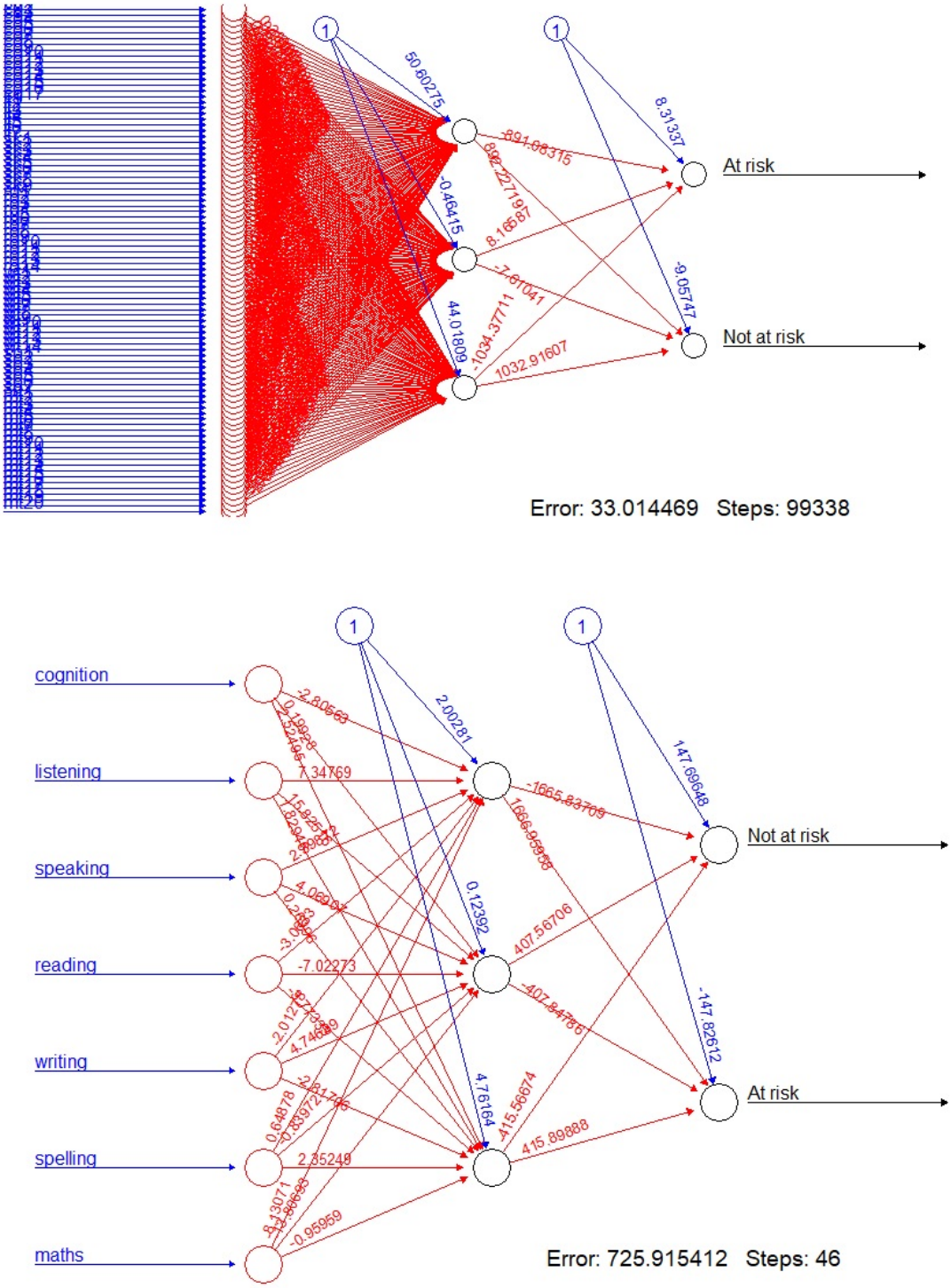


Figure 4.4: Artificial neural network model for diagnosis of learning disabilities

Support Vector Machine (SVM) Classier

The results in this study shows that support vector machines performs classification of imbalanced data better than all the eight models evaluated on 4 out of 7 evaluation metrics used in the study. While model accuracy is a biased measure of classification efficiency in imbalanced data, SVM shows high accuracy of over 98%. The detection rate of SVM was also found to be fairly high and stable as data structure changes. When the models were run on full dimensional data, ANN had the highest detection rate at 71% followed closed with SVM at 69% as shown in Table 4.3. When data dimension was reduced, detection rate of ANN was negatively affected by SVM remained the same at 69% but second to Naive Bayes. By implication, failure rate of SVM fell consistently below 32% across all the dataset.

The results show that SVM had low failure rate consistently compared to all other classification models. For instance, in full dimensional data, ANN had the lowest failure rate at 29% followed by SVM at 31%. Other models had relatively higher failure rate of upto particularly the random forest model and Naive Bayes. On reducing the data dimensionality, ANN's failure rate increased but SVM's failure rate remained the same at 31%, being the best model based on failure rate after Naive Bayes' 30%. The same failure rate is retained even after scaling the data.This shows that SVM is a stable learning algorithm whose performance is not easily affected by structural changes in the data.

Recent research studies have also confirmed that SVM are resilient models, for instance, Zendehboudi et al. (2018) contends that SVM has been shown to perform well in many real learning problems with a variety of settings and is often considered one of the best 'out-of-the-box' classifiers. Another study by Tao et al. (2019) found that SVM models are effective in high-dimensional spaces, even when the number of dimensions is greater than the number of samples. They are memory-efficient and versatile. However, if the number of features is much greater than the number of samples, the method is likely to result in poor performance (Tao et al., 2019). This means that SVM could have been problematic if we had a sample size of less than 88 making the attribute set to be higher than the sample size. In the premise of dimensionality, therefore, SVM can be

used to model imbalanced data with high dimension such as the problem of diagnostic assessment of learning disabilities.

Further still, the results show that SVM had the highest sensitivity for all datasets, outperforming the rest of the classification models considered in this research thesis. By definition, model sensitivity measures how often a model predicts a positive case, when it is actually true. Thus, in the case of imbalanced data, model sensitivity measures how responsive the model is in correctly classifying the minority class. By logic we conclude that a model with high sensitivity is desirable in assessing the efficiency of classification model in imbalanced data problems. Therefore, consistently higher sensitivity values associated with SVM suggest that the model is less biased towards the majority class.

SVM also had considerably high values of model specificity, which is of course expected due to high class imbalance ratio. However, the advantage of SVM is that it produces high values for both model sensitivity and specificity, which are then included in subsequent computation of F-and G-measures. Since both model specificity and sensitivity are close to each other, the gap between F-and G-measures. Other classification models that are biased towards majority class have significantly varying values of F-and G-measures.

Random Forest (RF) Classier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (Lee et al., 2020). This shows that decision trees form the fundamental building blocks of random forest classifier, hence, this study attempts to compare the classification performance of both of these closely related classification models.

The results show that for imbalanced datasets, the accuracy of random forest is the same as that of decision tree, they are all biased towards the majority class. Based on model sensitivity and detection rate, random forest perform poorly in high dimensional data compared to decision trees. Further, based on F-and G-measures, classification performance of random forest classifier is among the poorest, only better than Naive

Bayes and kNN model. However, when data dimension is reduced, it was observed that detection rate and sensitivity of random forest improves beyond that of decision trees. For imbalanced data problem where identification of the minority class is of interest, model sensitivity and detection rate are superior indicators of model classification efficiency. On this premise therefore, random forest classier can be considered a better algorithm for imbalanced data com pared to decision trees, the building block.

The results presented in Table 4.3, Table 4.6 and Table 4.9 and the subsequent discussions on each of the classification models leads us to conclude that support vector machine (SVM) is the most suitable base learner for imbalanced data. SVM showed consistently a fairly good sensitivity to the minority class compared to other classification models. Therefore, SVM is the most suitable base learner for imbalanced data regardless of data dimensionality. The next section of the experiment focusses on random undersamplin, Synthetic Minority Oversampling

4.4 Performance of Sampling Methods for Imbalanced Data

Three independent samples were drawn from the data as follows; First, the sampling ratio for model training and testing was retained as 75% and 25% respectively for all the samples, all drawn from full dimensional data. **sample 1** was drawn by setting *set.seed (2)*, **sample 2** by setting *set.seed (3)*, and **sample 3** by setting *set.seed (4)* in R software. The descriptive summaries of the three samples is presented in Table 4.11.

Table 4.11: Random samples drawn from full dimensional LDDI data

Random sample—	Training dataset			Testing dataset		
	Instances	Positive	IR	Instances	Positive	IR
Sample 1	3354	81	1:40	1119	27	1:40
Sample 2	3354	80	1:41	1119	28	1:40
Sample 3	3354	72	1:46	1119	36	1:30

The results presented in Table 4.11 shows that the three random samples were of equal size partitioned into 3,354 observations for training and 1,119 for testing. Sample 1 and sample 2 present an approximately congruent structure as occasioned by their imbalance ratios (IR) in both training and testing datasets, while sample three presents a high imbalance ratio in the training dataset compared to its testing dataset.

Receiver Operating Characteristics (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis at various threshold settings. By definition;

$$TPR = \frac{TP}{TP + FN} \tag{4.3}$$

which is model's sensitivity, plotted along the y-axis

$$FPR = \frac{FP}{FP + FN} \tag{4.4}$$

which is model's $(1 - \text{Specificity})$, plotted along the x-axis

In the case of imbalanced classification problems, emphasis is placed particularly on the predictive accuracy of minority class while retaining accuracy for the majority class. This would correspond to high TPR and low FPR, and thus reflected by a high AUC. The perfect classifier would have an area under the curve (AUC) of 1, where $TPR = 1$ and $FPR = 0$. Hence this study used AUC to compare performance of different sampling techniques in tackling the imbalanced data.

Table 4.12: Average performance of ROS, RUS, SMOTE and ADASYN

Random sample	Base Learner	ROS	RUS	SMOTE	ADASYN
Sample 1	0.58	0.71	0.61	0.78	**0.79**
Sample 2	0.59	0.72	0.63	0.79	**0.81**
Sample 3	0.64	0.74	0.69	**0.86**	**0.86**
Mean AUC	0.60	0.72	0.64	**0.81**	**0.82**

The results presented in Table 4.12 shows the AUC computed after running SVM model under different conditions as follows; The base learner refers to the original SVM model, whose average AUC on original imbalanced data was 0.60 or 60%. When random oversampling (ROS) is used in training the model, the AUC increases from 0.60 to 0.72, 19.9% improvement on overall classification efficiency. ROS rebalances the imbalanced data by duplicating examples from the minority class in the training set. This approach enhances the learning process by providing more instances of the under-represented. Since the training set ought to represent as much as possible the

scenario being modelled, one can argue that duplicating example could inherently mislead. A study by Kaur & Gosain (2018) on the behaviour of oversampling and undersampling approaches revealed that duplicating examples from the data can result into model overfitting, which is undesirable.

Applying random undersampling (RUS) improves the classification efficiency based on AUC of the base learner from 0.60 to 0.64, a 6.6% improvement. RUS rebalances the imbalanced data by deleting examples from the majority class such that the gap between the counts in the minority and majority class is reduced. In this case, when the imbalance ratio is very high, it would mean that a lot of instances in the majority class would be dropped. Additionally, while this thesis research focussed on binary classification, random undersampling might be a catastrophic attempt in a multi-class setting due to enormous loss of information. As observed by Kaur & Gosain (2018), random undersampling can result in losing information invaluable to a model during training.

Synthetic Minority Oversampling Technique (SMOTE) improves the AUC of the base SVM model from 0.60 to 0.81, an improvement of 34.3%. Compared to ROS and RUS, using SMOTE to rebalance class sizes can significantly improve model classification efficiency. The basic principle difference between ROS and SMOTE is that while ROS duplicates already existing examples, SMOTE synthesizes new instances, thus SMOTE adds new information in the data unlike ROS.SMOTE generates more minority class samples so that they are positioned randomly between a minority sample and its nearest neighbour (Fernández et al., 2018). SMOTE prevent over-fitting unlike ROS as it populates empty data space instead of simply replicating existing samples (Kaur & Gosain, 2018).

Adaptive Synthetic Sampling (ADASYN) improves the AUC of the base SVM model from 0.60 to 0.82 accounting for 35.9% increase in classification efficiency. ADASYN is similar to SMOTE, only featuring one important difference, it biases the sample space towards points which are located not in homogeneous neighbourhoods. As a result of this difference, some researchers suggest that ADASYN is more flexible than SMOTE and can be extended to utilizes the density distribution of a minority sample as criteria for how much synthetic data should be created (Alhudhaif, 2021). This is done so by

finding the differing density distribution for each minority sample and supplementing minority samples as much as each sample requires to become balanced with the majority class. This approach helps focus attention on minority examples depending on how difficult they are to learn.

As earlier presented in Figure 4.1, it is more difficult to distinguish the two classes of children based on total score than it would to base on distribution of the class score. (Kaur & Gosain, 2018) contends that class imbalance itself is not a problem but there are certain other data distribution complexities, which when combined with the class imbalance degrade the performance of classifier. One such complexities as brought out in this study is failure to recognize difference in the underlying statistical properties in the data from distinct groups, and over reliance on basic descriptives such as the quartile measures.

Based on the results presented in Table 4.12, random undersampling (RUS) and random oversampling (ROS) can tackle imbalance data problem, however, the margin of improvement is small and the two approaches are prone to misleading results due to risks of over fitting or loss of information. SMOTE and ADASYN are two robust sampling techniques that can tackle the problem of imbalanced data to a satisfactory level. However, ADASYN technique is 2.7% better than SMOTE at tackling imbalanced data problem. A chi-square test was used to test whether the difference is statistically significant. Under the null hypothesis that results from SMOTE and ADASYN are not statistically different, the results $\chi^2_{0.025,2} = 1.00 < 7.38$ do not provide sufficient evidence to reject the null hypothesis at 95% confidence level. Thus, we conclude that the two methods, SMOTE and ADASYN perform equally. However, for the purpose of this thesis research ADASYN was considered better than SMOTE due to some marginal positive difference it has over SMOTE.

4.5 Comparison of Sampling and AdaBoost for Imbalanced Data

This section presents comparative results of classification of support vector machine (SVM) model on imbalanced data under simple random sampling (SRS), Adaptive Synthetic Sampling (ADASYN) and Adaptive Boosting (AdaBoost).Under random sampling, a random sample is drawn from the population assuming homogeneity. Model training for classification task is done assuming each class is equally distributed. ADASYN and AdaBoost attempt to rebalance distribution among or between classes in data. ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. AdaBoost on the other hand rebalances the imbalanced data by assigning more weights to the minority class and less weight to majority class during model training. Table 4.13 shows the classification accuracy, detection rate, Failure rate, sensitivity, specificity, F-measure and G-measure of SVM on the same dataset.

Table 4.13: Comparative performance of ADASYN and AdaBoost Techniques

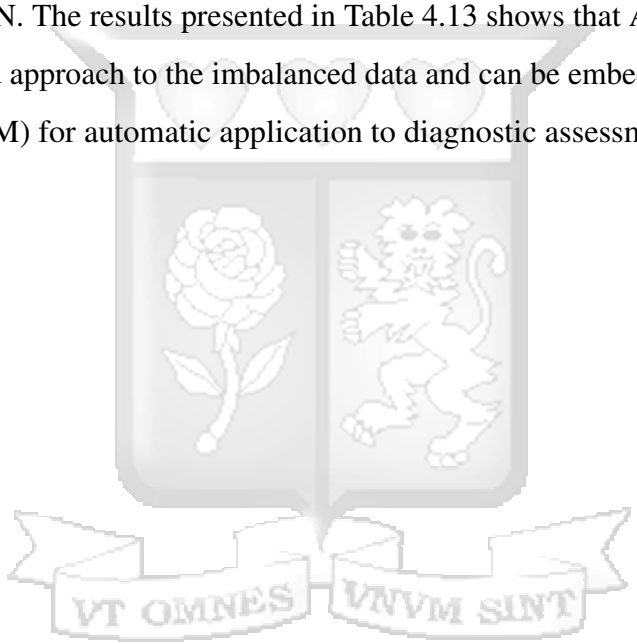
Evaluation metric	SVM under SRS	SVM under ADASYN	SVM under AdaBoost
Accuracy	0.99	0.99	0.99
Detection rate	0.69	0.88	0.91
Failure rate	0.31	0.12	0.09
Sensitivity	0.77	0.88	0.91
Specificity	0.98	0.99	0.99
F-measure	0.86	0.93	0.95
G-measure	0.87	0.93	0.95

Accuracy and Specificity of SVM

Both model classification and specificity are functions of true positives, hence, for imbalanced data where the minority class constitute the positive while the majority class constitute the negative, these evaluation metrics will be associated with high values. For this reason, SVM has 99% accuracy and over 98% specificity under SRS, ADASYN and AdaBoost as shown in Table 4.13. As such, both accuracy and specificity may not be ideal in measuring classification performance for imbalanced data.

Sensitivity and Minority Detection Rate of SVM

Model sensitivity and minority detection rate are the ideal measures of classification efficiency since they are both functions of true positives. Hence, they all contain useful information on how often the model detect a true positive given new dataset. As shown in Table 4.13, SVM was fairly sensitive to the minority class at 77% under SRM with a lower detection rate of 69%. ADASYN improve model detection rate by 27.5% while AdaBoost improves model detection rate by 31.9%. ADASYN improves model sensitivity by 14.3% while AdaBoost improves model sensitivity by 18.2%. Overall, the results shows that AdaBoost improves model classification efficiency by 5.7% better than ADASYN. The results presented in Table 4.13 shows that AdaBoost technique is the best suited approach to the imbalanced data and can be embedded in support vector machine (SVM) for automatic application to diagnostic assessment of neurobilogical conditions.



CHAPTER FIVE

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the logical conclusions drawn from the experimental results of this study and recommendations for action by practitioners and future research work by scientific community in computational science and disability assessment.

5.2 Conclusion

5.2.1 Base learner models for classification of imbalanced data

One of the objectives of this thesis research was to evaluate the performance of standard classification models on imbalanced data. As part of a robust statistical learning system, a feasible classification model termed as a base learner ought to be identified for a particular task, then modifications made to the base learner to adapt its functionality to the specific situations. This thesis research evaluated performance of eight classification models with the aim of identifying the most suitable base learner for a highly imbalanced data collected from a baseline survey on learning disabilities. The following conclusions are made as far as selection of a base learner is concerned;

(i) Naive Bayes, decision trees and logistic regression can be trained relatively fast and their prediction speed is high. By implication, application of these models may need small computer memory. These classification models may still have a role in modern computing, however, their application could be limited. For instance, logistic regression is suitable for small problems with linear decision boundary while decision trees will be suitable for low dimensional data because its prone to overfitting.

(ii) In general, k-nearest neighbour classifier (kNN) has the lowest classification performance when the task has a considerably high number of input variables.

The model also performs poorly when the sample size is high, and much worse when the classes in the data are imbalanced. However, it should be noted that for low dimensional data, with a small sample, kNN can be a robust classification model of choice with easy implementation and interpretation.

(iii) Training an artificial neural network (ANN) can be extremely low especially when the input variables are many, and by implication, building and using ANN may require bigger computer memory. However, prediction speed is fair, not too slow as training, but not as fast as Naive Bayes. Convergence of an ANN model is based on a random algorithm, and thus, running the ANN model several times can result into different outputs. However, the model error and training steps will remain stable, with negligible variations provided the model converges.

(iv) Training support vector machine (SVM) takes relatively shorter time compared to artificial neural network (ANN), however, SVM prediction speed takes longer than ANN.

(v) Support Vector machine(SVM) and artificial neural network (ANN) are by far the most flexible classification models that can be used for imbalanced data. ANN performs slightly better than SVM when input variables are many, although the difference is not statistically significant. However, ANN performs well below expectation compared to SVM when the dataset has fewer variables or scaled. This study therefore considers that while the two models are equally flexible, SVM has an edge over ANN when the data structure changes drastically.

(vi) While decision tree are basic building blocks for random forest models, random forest have higher sensitivity to minority class, and a reasonably higher detection rate in high dimensional data. Therefore, in considering between the two models, random forest would be a better option especially in cases where there is a sufficiently large sample size to train the RF model

5.2.2 Choosing a solution to tackle imbalanced data problem

(i) The challenge of imbalanced data is still a problem in designing machine learning algorithms in assessment, thus, practitioners in the field will have to deal with the problem in some way.

(ii) For developers with sufficient domain science knowledge, data-based solutions would serve them better while for developers with different background other than the domain science area, algorithm-based level solutions would be ideal.

(iii) Researchers and practitioners have statistical tools to treat the imbalanced data problem. These methods could be at data level and algorithm level. However, a blend of the two methods produces cost sensitive approaches for solving the problem of imbalanced data. This study concludes that when deciding how to deal with the problem of imbalanced data in practice, it would be a good idea to often blend between data level and algorithm level. This recommendation stems from full realization that either of these approaches have their own biases that need to be complimented by a similar methodology.

(iv) For highly imbalanced data, model sensitivity and detection rate are distinct evaluation metrics giving different values, however, for least imbalanced data, the two metrics overlap, and one can easily be mistaken for another.

5.2.3 Model performance evaluation metrics

(i) There are many different model evaluation metrics to choose from when analysing model performance. While some metrics may have received significant higher criticism than others, this study concludes that reliance on a single evaluation metric may be misleading. For instance, model specificity is sensitive to changes in true negative value while model sensitivity is to false negative. None of these two measures may provide sufficient information on model classification power as a stand alone metric.The same could be said on model detection rate, prevalence and precision.

(ii) F-measure and G-measure are composite functions of model sensitivity and

specificity. This study concludes that a large disparity between model sensitivity and specificity negatively affects the F and G measures, but F-measure is severely affected. Plausible reason for this effect is the computation structure underpinning the two measures; F-measure is computed by means of harmonic mean while the G-measure is computed by means of geometric mean. Accordingly, simultaneously application of the two measures would be a more robust indicator of model classification performance, rather than any one of them.

(iii) Model accuracy and specificity are biased model evaluation metrics for imbalanced data as they are skewed by the majority class. This is evidenced by consistently high accuracy and specificity for most classification models even when their detection rate was poorest. For learning disabilities, training an ANN model with full dimensional data gives better results compared to reduced dimensional data. This is possibly due to potential loss of patterns and subtle data structure when the scores of every domain are obtained by arithmetic summation

5.3 Recommendations

5.3.1 Recommendations for action

(i) Assessment of learning disabilities for educational purposes is often done by a team of professionals including teachers and psychologists. For medical reasons, LD assessment could be done cognitive neuroscientist alone with the help of specialized technologies such as EEG machines. In both cases of assessment, data is generated which is then analysed for conclusions on final diagnosis is made. This study recommends that computational scientists should be part of the assessment process for proper conceptualization of measurements, and provide support to avoid data contamination. This will go along way in generating data that is more useful for machine learning, that would eventually result into development of artificially intelligent assessment systems.

(ii) The study recommends that signal processing-based methodologies would be more ideal is distinguishing children at risk of learning disabilities from those with general academic difficulties. This is evidenced from the difficulties associated

with training classification models based on total scores per domain compared to training the model based on the score distribution between the two classes of children.

5.3.2 Recommendations for further research

(i) This study recommends future research on analysis of algebraic and topological structures underlying model detection rate and model sensitivity. This arises from an unusual overlap between these two evaluation metrics observed in this study.

(ii) This study recommends further research to focus on understanding ensemble learning approaches with a the aim of gaining a deeper understanding of how to deal with imbalanced data with heterogeneity in both minority and majority classes.

(iii) Since learning disabilities manifest in different ways such as dyslexia, discalculia, and disagraphia among other, this study recommends that further research on classification is needed focus on learning disability as a multi-class problem rather than binary classification problem. It would be interesting to see how the decision boundary that classify these subjects shifts, and whether the same learning framework could be replicated.

REFERENCES

Abdeljabbar, A. (2021). New double wronskian exact solutions for a generalized (2+ 1)-dimensional nonlinear system with variable coefficients. *Partial Differential Equations in Applied Mathematics*, *3*, 100022.

Abellán, J., & Castellano, J. G. (2017). Improving the naive bayes classifier via a quick variable selection method using maximum of entropy. *Entropy*, *19*(6), 247.

Alhudhaif, A. (2021). A novel multi-class imbalanced eeg signals classification based on the adaptive synthetic sampling (adasyn) approach. *PeerJ Computer Science*, *7*, e523.

Amra, I. A. A., & Maghari, A. Y. (2017). Students performance prediction using knn and naïve bayesian. In *2017 8th international conference on information technology (icit)* (pp. 909–913).

Baig, M. M., Awais, M. M., & El-Alfy, E.-S. M. (2017). Adaboost-based artificial neural network learning. *Neurocomputing*, *248*, 120–126.

Barni, M., Nowroozi, E., Tondi, B., & Zhang, B. (2020). Effectiveness of random deep feature selection for securing image manipulation detectors against adversarial examples. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2977–2981).

Bilbao, I., & Bilbao, J. (2017). Overfitting problem and the over-training in the era of data: Particularly for artificial neural networks. In *2017 eighth international conference on intelligent computing and information systems (icicis)* (pp. 173–177).

Bollegala, D. (2017). Dynamic feature scaling for online learning of binary classifiers. *Knowledge-Based Systems*, *129*, 97–105.

Bootkrajang, J., & Chaijaruwanich, J. (2020). Towards instance-dependent label noise-tolerant classification: a probabilistic approach. *Pattern Analysis and Applications*, *23*(1), 95–111.

Braams, B. J., & Bowman, J. M. (2009). Permutationally invariant potential energy surfaces in high dimensionality. *International Reviews in Physical Chemistry*, *28*(4), 577–606.

Bruijn, S. M., Bregman, D. J., Meijer, O. G., Beek, P. J., & van Dieën, J. H. (2012). Maximum lyapunov exponents as predictors of global gait stability: a modelling approach. *Medical engineering & physics*, *34*(4), 428–436.

Cañete-Sifuentes, L., Monroy, R., & Medina-Pérez, M. A. (2021). A review and experimental comparison of multivariate decision trees. *IEEE Access*.

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve bayes algorithm. *Knowledge-Based Systems*, *192*, 105361.

Chen, Y.-F., Hsieh, C., Lengál, O., Lii, T.-J., Tsai, M.-H., Wang, B.-Y., & Wang, F. (2016). Pac learning-based verification and model synthesis. In *2016 ieee/acm 38th international conference on software engineering (icse)* (pp. 714–724).

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, *110*, 12–22.

Cullina, D., Bhagoji, A. N., & Mittal, P. (2018). Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*.

Cunningham, P., & Delany, S. J. (2020). k-nearest neighbour classifiers–. *arXiv preprint arXiv:2004.04523*.

Damanik, I. S., Windarto, A. P., Wanto, A., Andani, S. R., Saputra, W., et al. (2019). Decision tree optimization in c4. 5 algorithm using genetic algorithm. In *Journal of physics: Conference series* (Vol. 1255, p. 012012).

Datta, S., Nag, S., Mullick, S. S., & Das, S. (2017). Diversifying support vector machines for boosting using kernel perturbation: Applications to class imbalance and small disjuncts. *arXiv preprint arXiv:1712.08493*.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772.

Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing*, *91*, 106229.

Dick, V., Sinz, C., Mittlböck, M., Kittler, H., & Tschandl, P. (2019). Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. *JAMA dermatology*, *155*(11), 1291–1299.

Edla, D. R., Mangalorekar, K., Dhavalikar, G., & Dodia, S. (2018). Classification of eeg data for human mental state analysis using random forest classifier. *Procedia computer science*, *132*, 1523–1532.

Elaidi, H., Benabbou, Z., & Abbar, H. (2018). A comparative study of algorithms constructing decision trees: Id3 and c4. 5. In *Proceedings of the international conference on learning and optimization algorithms: Theory and applications* (pp. 1–5).

El-Khatib, M. J., Abu-Nasser, B. S., & Abu-Naser, S. S. (2019). Glass classification using artificial neural network.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, *505*, 32–64.

Feng, X., Li, S., Yuan, C., Zeng, P., & Sun, Y. (2018). Prediction of slope stability using naive bayes classifier. *KSCE Journal of Civil Engineering*, *22*(3), 941–950.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, *37*, 66–74.

Gao, X., & Li, G. (2020). A knn model based on manhattan distance to identify the snare proteins. *IEEE Access*, *8*, 112922–112931.

Gardini, L., Grebogi, C., & Lenci, S. (2020). Chaos theory and applications: a retrospective on lessons learned and missed or new opportunities. *Nonlinear Dynamics*.

Gong, L., Jiang, S., & Jiang, L. (2019). Tackling class imbalance problem in software defect prediction through cluster-based over-sampling with filtering. *IEEE Access*, *7*, 145725–145737.

Granik, M., & Mesyura, V. (2017). Fake news detection using naive bayes classifier. In *2017 ieee first ukraine conference on electrical and computer engineering (ukrcon)* (pp. 900–903).

Grünwald, P., & Roos, T. (2019). Minimum description length revisited. *International journal of mathematics for industry*, *11*(01), 1930001.

Gujar, R., & Vakharia, V. (2019). Prediction and validation of alternative fillers used in micro surfacing mix-design using machine learning techniques. *Construction and Building Materials*, *207*, 519–527.

Gyamfi, K. S., Brusey, J., Hunt, A., & Gaura, E. (2017). Linear classifier design under heteroscedasticity in linear discriminant analysis. *Expert Systems with Applications*, *79*, 44–52.

Hansen, M. A., Armstrong, E., & Sutherland, J. C. (2018). State space parameterization of explosive eigenvalues during autoignition. *Combustion and Flame*, *196*, 182–196.

Hasanin, T., & Khoshgoftaar, T. (2018). The effects of random undersampling with simulated class imbalance for big data. In *2018 ieee international conference on information reuse and integration (iri)* (pp. 70–79).

Hasmita, S., Nhita, F., Saepudin, D., & Aditsania, A. (2019). Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling (adasyn) and k-nearest neighbor (knn) algorithms. In *2019 international conference on information and communications technology (icoiact)* (pp. 434–438).

Hossain, E., Hossain, M. F., & Rahaman, M. A. (2019). A color and texture based approach for the detection and classification of plant leaf disease using knn classifier. In *2019 international conference on electrical, computer and communication engineering (ecce)* (pp. 1–6).

Hou, B., Wu, Q., Wen, Z., & Jiao, L. (2017). Robust semisupervised classification for polsar image with noisy labels. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(11), 6440–6455.

Hou, Y., Li, B., Li, L., & Liu, J. (2019). A density-based under-sampling algorithm for imbalance classification. In *Journal of physics: Conference series* (Vol. 1302, p. 022064).

Huang, L., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2019). Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4874–4883).

Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, *6*(2), 128–146.

Ishwaran, H., & O'Brien, R. (2020). Commentary: the problem of class imbalance in biomedical data. *J Thorac Cardiovasc Surg*, *1*, 2.

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265–276.

Jia, W., Deng, Y., Xin, C., Liu, X., & Pedrycz, W. (2019). A classification algorithm with linear discriminant analysis and axiomatic fuzzy sets. *Mathematical Foundations of Computing*, *2*(1), 73.

Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European journal of operations research*, *26*(1), 135–159.

Kaur, P., & Gosain, A. (2018). Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In *Ict based innovations* (pp. 23–30). Springer.

Kawaguchi, K., Bengio, Y., Verma, V., & Kaelbling, L. P. (2018). Generalization in machine learning via analytical learning theory. *arXiv preprint arXiv:1802.07426*.

Kim, S., Jhong, J.-H., Lee, J., & Koo, J.-Y. (2017). Meta-analytic support vector machine for integrating multiple omics data. *BioData mining*, *10*(1), 1–14.

Kolukisa, B., Hacilar, H., Goy, G., Kus, M., Bakir-Gungor, B., Aral, A., & Gungor, V. C. (2018). Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. In *2018 ieee international conference on big data (big data)* (pp. 2232–2238).

Lakshmanaprabu, S., Shankar, K., Ilayaraja, M., Nasir, A. W., Vijayakumar, V., & Chilamkurti, N. (2019). Random forest for big data classification in the internet of things using optimal features. *International journal of machine learning and cybernetics*, *10*(10), 2609–2618.

Le, T. T., Fu, W., & Moore, J. H. (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, *36*(1), 250–256.

Lee, T.-H., Ullah, A., & Wang, R. (2020). Bootstrap aggregating and random forest. In *Macroeconomic forecasting in the era of big data* (pp. 389–429). Springer.

Lin, E., Chen, Q., & Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence*, *50*(8), 2488–2502.

Liu, Z., Zuo, Q., Wu, G., & Li, Y. (2018). An artificial neural network developed for predicting of performance and emissions of a spark ignition engine fueled with butanol–gasoline blends. *Advances in Mechanical Engineering*, *10*(1), 1687814017748438.

Ma, J., & Xie, L. (2018). The impact of loss sensitivity on a mobile phone supply chain system stability based on the chaos theory. *Communications in Nonlinear Science and Numerical Simulation*, *55*, 194–205.

Magnusson, M., Vehtari, A., Jonasson, J., & Andersen, M. (2020). Leave-one-out cross-validation for bayesian model comparison in large data. In *International conference on artificial intelligence and statistics* (pp. 341–351).

Malladi, R., & Fabozzi, F. J. (2017). Equal-weighted strategy: Why it outperforms value-weighted strategies? theory and evidence. *Journal of Asset Management*, *18*(3), 188–208.

Miciak, J., & Fletcher, J. M. (2020). The critical role of instructional response for identifying dyslexia and other learning disabilities. *Journal of Learning Disabilities*, *53*(5), 343–353.

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*, *52*(2), 857–900.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51–62.

Patel, H., & Thakur, G. (2019). An improved fuzzy k-nearest neighbor algorithm for imbalanced data using adaptive approach. *IETE Journal of Research*, *65*(6), 780–789.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012–4024.

Pereira, R. M., Costa, Y. M., & Silla Jr, C. N. (2020). Mltl: A multi-label approach for the tomek link undersampling algorithm. *Neurocomputing*, *383*, 95–105.

Petkovic, D., Altman, R., Wong, M., & Vigil, A. (2018). Improving the explainability of random forest classifier–user centered approach. In *Pacific symposium on biocomputing 2018: Proceedings of the pacific symposium* (pp. 204–215).

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101–121). Elsevier.

Portet, S. (2020). A primer on model selection using the akaike information criterion. *Infectious Disease Modelling*, *5*, 111–128.

Prasad, V., & Rao, T. S. (2017). Permissible thyroid datasets assessment through kernel pc algorithm and vapnik chervonenkis theory for categorization and classification. *Data-Enabled Discovery and Applications*, *1*(1), 5.

Rajaguru, H., & Prabhakar, S. K. (2017). Bayesian linear discriminant analysis for breast cancer classification. In *2017 2nd international conference on communication and electronics systems (icces)* (pp. 266–269).

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Rayhan, F., Ahmed, S., Mahbub, A., Jani, R., Shatabda, S., & Farid, D. M. (2017). Cusboost: Cluster-based under-sampling with boosting for imbalanced classification. In *2017 2nd international conference on computational systems and information technology for sustainable solution (csitss)* (pp. 1–5).

Sarker, I. H., Kayes, A., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, *6*(1), 1–28.

Sarma, M. S., Srinivas, Y., Abhiram, M., Ullala, L., Prasanthi, M. S., & Rao, J. R. (2017). Insider threat detection with face recognition and knn user classification. In *2017 ieee international conference on cloud computing in emerging markets (ccem)* (pp. 39–44).

Shi, Y., Sagduyu, Y., & Grushin, A. (2017). How to steal a machine learning classifier with deep learning. In *2017 ieee international symposium on technologies for homeland security (hst)* (pp. 1–5).

Siddappa, N. G., & Kampalappa, T. (2019). Adaptive condensed nearest neighbor for imbalance data classification. *International Journal of Intelligent Engineering and Systems*, *12*(2), 104–113.

Smiti, S., & Soui, M. (2020). Bankruptcy prediction using deep learning approach based on borderline smote. *Information Systems Frontiers*, *22*(5), 1067–1083.

Sternberg, R. J. (2018). Epilogue: Toward an emerging consensus about learning disabilities. In *Perspectives on learning disabilities* (pp. 277–282). Routledge.

Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Information Sciences*, *487*, 31–56.

Triantafyllidis, A., Polychronidou, E., Alexiadis, A., Rocha, C. L., Oliveira, D. N., da Silva, A. S., ... others (2020). Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature. *Artificial Intelligence in Medicine*, *104*, 101844.

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences*, *509*, 47–70.

Zendehboudi, A., Baseer, M., & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of cleaner production*, *199*, 272–285.

Zhang, H., Li, Q., Liu, J., Shang, J., Du, X., McNairn, H., ... Liu, M. (2017). Image classification using rapideye data: Integration of spectral and textual features in a random forest classifier. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(12), 5334–5349.

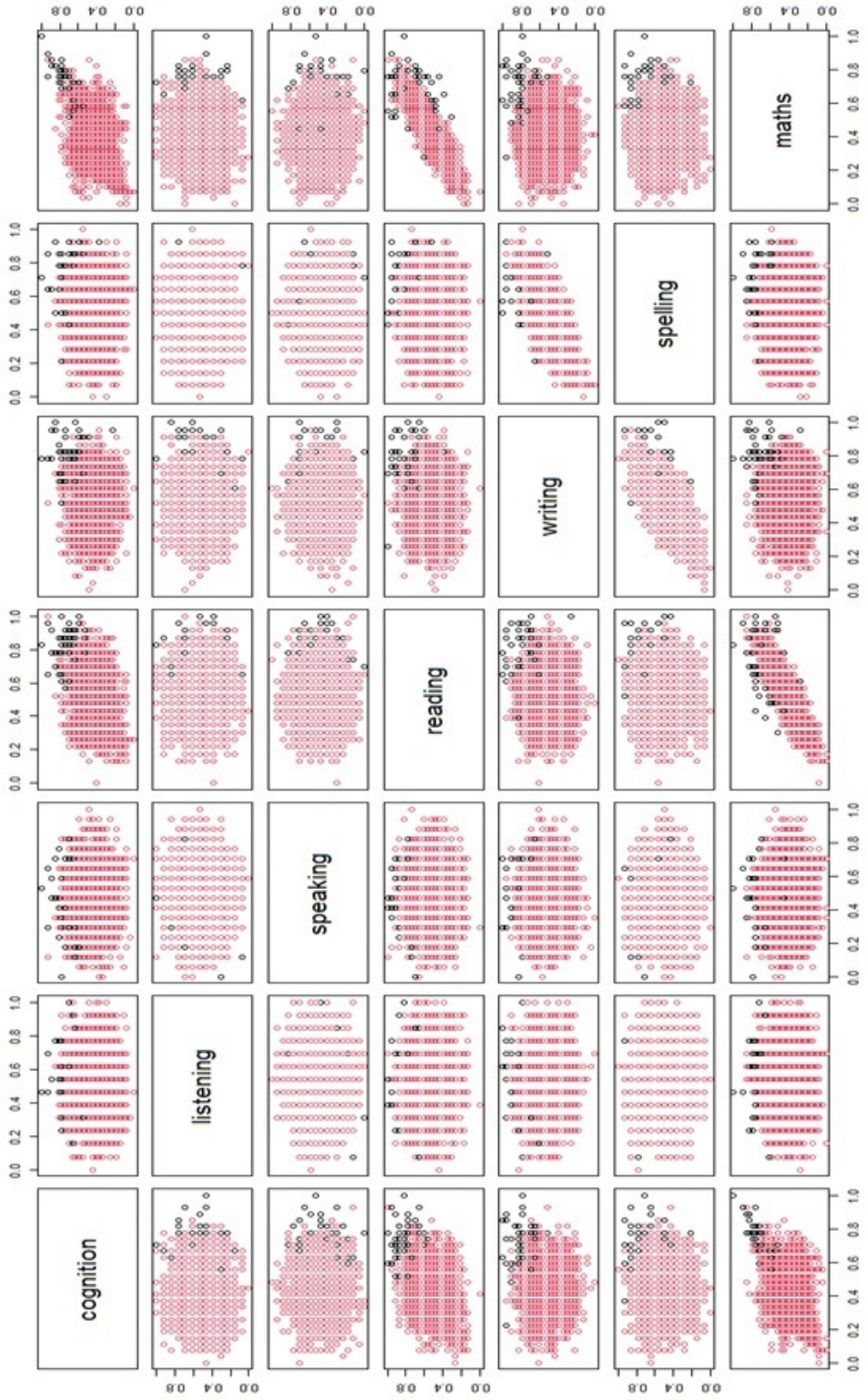
Zhang, J., & Chen, L. (2019). Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Computer Assisted Surgery*, *24*(sup2), 62–72.

Zhang, L., Zhang, L., Du, B., You, J., & Tao, D. (2019). Hyperspectral image unsupervised classification by robust manifold matrix factorization. *Information Sciences*, *485*, 154–169.

Zheng, X., Wang, Y., Wang, G., & Liu, J. (2018). Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, *107*, 55–71.

APPENDICES

Appendix 1: Cognitive profile of children with learning difficulties



Appendix 2: Permission from KISE to Access LDDI-19 Data

KENYA INSTITUTE OF SPECIAL EDUCATION

Tel: 020- 8007977,
Cell: 0734-801 - 860
Website: www.kise.ac.ke
Email: info@kise.ac.ke.

Kasarani, Thika Superhighway Exit 8
Off Kasarani-Mwiki Rd
P. O. Box 48413 - 00100
NAIROBI, KENYA

Our Ref: KISE/365/16

Your Ref: .....

Date: 6th July 2020

Samuel Wanyonyi Juma
KISE/365
P O Box 48413 – 00100
Nairobi

RE: PERMISSION TO USE LDDI-19 DATA FOR ACADEMIC PURPOSES

Reference is made to your letter dated 13/5/2020 requesting to access and use the institute's learning disabilities data for your thesis in MSc Statistical Science at Strathmore University.

This is to inform you that your request has been considered by the institute's research committee and permission is hereby granted.

Note that you shall be bounded by the Kenya Institute of Special Needs Education (KISE)'s guidelines on "Responsible Conduct of Research and Ethics" in the use of this data.

MUTISO T. WAMBUA HSC
DIRECTOR

KENYA INSTITUTE OF SPECIAL EDUCATION
Tel. 020-8007977; Email: info@kise.ac.ke
P. O. BOX 48413 – 00100
NAIROBI – KENYA

An inclusive Kenyan society in which every child with special needs and disability accesses quality education and achieves their fullest human potential
ISO 9001:2015 CERTIFIED.

Appendix 3: Similarity Report

Curiginal

Document Information

Analyzed document	Robust Statistical Learning for Optimal Classification of Imbalanced Data.pdf (D113509475)
Submitted	2021-09-27 04:11:00
Submitted by	
Submitter email	Samuel.Juma@strathmore.edu
Similarity	3%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

W	URL: https://digital.library.ryerson.ca/islandora/object/RULA%3A9363/datastream/OBJ/download/Athlete_Health_Prediction_Using_Machine_Learning_Methods.pdf Fetched: 2021-06-05 08:33:06	1
W	URL: http://vuir.vu.edu.au/29496/1/Jaree%20Thongkam.pdf Fetched: 2021-05-02 13:59:07	1
W	URL: https://www.researchgate.net/publication/263913891_Classification_of_imbalanced_data_a_review Fetched: 2021-09-27 04:12:00	6
W	URL: https://machinelearningmastery.com/what-is-imbalanced-classification/ Fetched: 2021-09-27 04:12:00	7
W	URL: https://www.researchgate.net/publication/334785564_Predicting_credit_default_probabilities_using_machine_learning_techniques_in_the_face_of_unequal_class_distributions Fetched: 2021-05-29 14:41:53	2
W	URL: https://eprints.utas.edu.au/32903/1/136017%20-%20Machine%20learning%20for%20mining%20imbalanced%20data.pdf Fetched: 2021-09-27 04:12:00	2
W	URL: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ Fetched: 2021-09-27 04:12:00	1
W	URL: https://www.researchgate.net/profile/Ebenezer-Esenogho/project/5G-COGNITIVE-RADIO-WIRELESS-NETWORKS/attachment/6058ad340f39c700013b4515/AS:1004170215710721%401616424244563/download/Improved%2BMachine%2BLearning%2BMethods%2Bfor%2BClassification%2Bof%2BImbalanced%2BData.pdf%3Fcontext%3DProjectUpdatesLog Fetched: 2021-06-26 07:24:16	2
W	URL: http://dspace.uiu.ac.bd/bitstream/handle/52243/159/Md.%20Yasir%20Arafat%20-%20012162024.pdf?sequence=1&isAllowed=y Fetched: 2021-09-27 04:12:00	1
W	URL: https://www.mdpi.com/2076-3417/11/8/3450/htm Fetched: 2021-07-25 13:05:37	2

Appendix 4: Ethics review approval Letter Ref. SU-IERC1039/21

Strathmore UNIVERSITY

4th June 2021

Samuel Juma,
samuel.juma@strathmore.edu

Dear Mr Juma,

RE: Robust Statistical Learning for Optimal Classification of Imbalanced Data

This is to inform you that SU-IERC has reviewed and **approved** your above **SU-master's** research proposal. Your application reference number is **SU-IERC1039/21.** The approval period is **4th June 2021 to 3rd June 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and also obtain other clearances needed

Yours sincerely,

for: Dr Virginia Gichuru,
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
04 Jun 2021
TEL: +254 (0) 70303400
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix 5: Ethics review Final Decision certificate Ref. SU-IERC1039/21

Final Decision Certificate

This document certifies that the study:

"Robust Statistical Learning for Optimal Classification of Imbalanced Data"

Principal Investigator: Juma, Samuel Wanyonyi
Reference number: SU-IERC1039/21

Was reviewed and received the following status:
"approved"
Additional Comments:

Reviewer #1:
'Good work and very relevant for students with special needs. '