# HIERARCHICAL MULTI-STAGE DIMENSIONAL REDUCTION BASED ON FEATURE HASHING AND BI-FILTERING STRATEGY FOR LARGE-SCALE TEXT CLASSIFICATION

ABUBAKAR ADO

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

JULY 2023

This thesis is dedicated to my beloved family, Late Alh. Ado Rogo’s family. Without their inspiration, support, and, most notably, their prayers, this study would not have been accomplished.

# ACKNOWLEDGEMENT

Firstly, I would like to thank my academic supervisor, Assoc. Prof. Dr. Noor Azah Samsudin for her valuable time, countless advice, and unforgettable support throughout the research period, and also a special appreciation goes to my former supervisor, Prof. Dr. Mustafa Mat Deris. Advice given by them has been of great important in solving the faced challenging issues.

Secondly, I would also like to convey my special appreciation to the Faculty of Computer Science and Information Technology (UTHM) for providing me with the conducive study environment towards achieving the research goals, and also the management of Yusuf Maitama Sule University (YUMSUK) for giving me the opportunity to embark on this journey.

Lastly, I will not end this without expressing my endless gratitude to my beloved parents, sisters, brothers, relatives, friends and my lovely wife for their pronounced encouragement and support.

## ABSTRACT

The advancement in technology has resulted in large size of data, which then introduce challenges to labelling or classification tasks with high dimensional features. Specifically, in the case of text labelling problem, the existing classification models are challenged with a huge number of instances, millions number of features, and large number of categories. Such challenge requires a well-defined hierarchy structure and automated classification models to label the instances within the hierarchy, which can be referred to as Large-Scale Hierarchical Text Classification (LSHTC). Even with a well-defined hierarchy, the LSHTC problem is still facing a scalability issue. Therefore, this requires improvements in the dimensional reduction phase of the LSHTC framework that aim at constructing a subset of informative features. However, using the existing dimensionality reduction methods in LSHTC problem has the consequence of introducing bad collisions or results discrepancy limitations. Therefore, in this thesis, a Multi-stage Dimensional Reduction Method (MDRM) based on feature hashing and bi-strategy filter method is proposed for the LSHTC problem. In view of solving the aforementioned problems, a Modified Feature Hashing (MFH) based on term weight to minimize bad collisions rate is presented, whereas for dealing with results discrepancy, a new Bi-strategy Filtering Approach (BFA) is presented. Experimental results show that the proposed MFH outperformed the conventional features hashing approximately by 3%. BFA has achieved the highest average micro-f1 score of 53.38% and 55.58%, and the highest average macro-f1 score of 45.83% and 49.23% compare to the single strategy filtering methods. It also achieves highest hierarchical-f1 of 79.99%, 67.83%, and 67.95% compare to existing multi-strategy filtering approaches. Lastly, the MDRM has achieved the best performance in terms of average micro-f1 (58.47% and 54.77%) and average macro-f1 (51.14% and 48.70%), respectively. In the case of running time, the MDRM has achieved 11% faster than the single stage reduction method and about 37% faster than baseline method.

## ABSTRAK

Kemajuan teknologi telah menghasilkan data bersaiz besar, lalu menyebabkan cabaran dalam melabel atau mengklasifikasi tugas yang mempunyai sifat dimensional yang tinggi. Secara lebih spesifik, dalam masalah "text labelling", model klasifikasi yang wujud sedang mendepani cabaran jumlah instances yang besar, jutaan features, dan jumlah kategori yang banyak. Cabaran ini memerlukan struktur hierarki yang jelas dan model klasifikasi automatik untuk melabel "instances" dalam hierarki, dan perkara ini dikenali sebagai masalah Large-Scale Hierarchical Text Classification (LSHTC). Walaupun hierarki adalah jelas, masalah LSHTC masih juga menghadapi isu skalabiliti. Masalah LSHTC ini memerlukan penambahbaikan dalam fasa pengurangan dimensional, yang bertujuan untuk membina sebuah subset yang mengandungi ciri-ciri bermaklumat. Oleh itu, dalam tesis ini, sebuah Multi-stage Dimensional Reduction Method (MDRM) berdasarkan ciri-ciri hashing dan kaedah penapisan dwi-strategi telah dicadangkan untuk menyelesaikan masalah LSHTC ini. Bagi menyelesaikan masalah yang telah dinyatakan, suatu Modified Feature Hashing (MFH) berdasarkan term weight telah diutarakan untuk meminimumkan kadar bad collisions. Selain itu, untuk menangani percanggahan results, Bi-strategy Filtering Approach (BFA) yang baharu telah dicadangkan. Hasil kajian menunjukkan bahawa MFH mempamerkan prestasi lebih baik berbanding features hashing konvensional sebanyak tiga peratus.BFA telah mencapai purata skor micro-f1 tertinggi iaitu sebanyak 53.38 and 55.58%, dan mencapai purata skor macro-f1 tertinggi sebanyak 45.83% and 49.23%, berbanding dengan kaedah penapisan strategi yang sedia ada. Serta mencapai hierarki-f1 tertinggi iaitu sebanyak 79.99%, 67.83%, dan 67.95% berbanding pendekatan penapisan pelbagai strategi sedia ada MDRM telah mempamerkan prestasi paling memberangsangkan, dari segi purata micro-f1 (58.47% and 54.77%), dan purata macro-f1 (51.14% and 48.70%). Dari aspek running time, MDRM mencapai 11% lebih kelajuan berbanding kaedah pengurangan single stage, dan lebih kurang 37% lebih laju berbanding kaedah baseline.

## CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| AF | - | All Features |
| BFA | - | Bi-strategy Filtering Approach |
| BOW | - | Bag-Of-Words |
| CFH | - | Conventional Feature Hashing |
| Chi-2 | - | Chi-square |
| CFS | - | Correlation Based Feature Selection |
| DR | - | Dimensional Reduction |
| DMOZ | - | Directory Mozilla |
| FiS | - | Fisher Score |
| FS | - | Feature Selection |
| FH | - | Feature Hashing |
| FE | - | Feature Extraction |
| FA | - | Feature Abstraction |
| FC | - | Flat Classification |
| FS-P | - | Feature Selection-Perceptron |
| GDA | - | General Discriminant Analysis |
| HC | - | Hierarchical Classification |
| GINI | - | Gini Index |
| GC | - | Global Classifier |
| HFH | - | Hierarchical Feature Hashing |
| HSA | - | Heuristic Search Algorithm |
| *hF1* | - | Hierarchical-F1 |
| *hEr* | - | Hierarchical-Error |
| *hP* | - | Hierarchical Precision |
| *hR* | - | Hierarchical Recall |
| IG | - | Information Gain |
| IPC | - | International Patent Classification |

IDF - Inverse Document Frequency
ISOMAP - Isometric Mapping
LSHC - Large Scale Hierarchical Classification
LCN - Local Classifier per Node
LCPN - Local Classifier per Parent Node
LCL - Local Classifier per Level
LDA - Linear Discriminant Analysis
LLE - Local Linear Embedding
MI - Mutual Information
MDRM - Multi-stage Dimensional Reduction Method
MFH - Modified Feature Hashing
MTL - Multi-Task Learning
NG - NewsGroups
ODP - Open Directory Project
PCA - Principle Component Analysis
TC - Text Classification
TF - Term Frequency
TD-LR - Top-Down Logistic Regression
TD-SVM -Top-Down Support Vector Machine
T-test - Student Statistical Test
PCA - Principal Component Analysis
$f_i$ - $i^{th}$ Feature
$\mathcal{H}$ - Original Hierarchy
$\mathcal{H}_m$ - Modified Hierarchy
$\aleph$ - Set of all nodes in $\mathcal{H}$
$\ell$ - Set of leaf nodes (categories) in $\mathcal{H}$ ; $\ell \subseteq \aleph$
$\aleph - \ell$ - Set of internal nodes; $\aleph - \ell \subseteq \aleph$
$\mathcal{Q}$ - Root node in $\mathcal{H}$
$\mathcal{C}(n)$ - Set of all children of node *n*
$\mathcal{P}(n)$ - Parent of node *n*
$sib(n)$ - Siblings of node *n*
$SL$ - Subset of relevant features selected; $SL \subseteq \phi(x,u)$
$T_n(x)$ - Total number of training instances at node *n* $T_n(x) \subseteq m$

$\{(x_i\ , y_i)\}_{i=1}^{m}$ - Dataset of *m* training instances, where $x_i \subseteq \mathcal{X}\ and\ y_i \subseteq \ell$

$R^d$ - Original feature space

$\phi(x, u)$ - Hash feature space; $\phi(x, u) \in R^d$ and $u \in \aleph$

# LIST OF APPENDICES

# LIST OF PUBLICATIONS

**Journal(s):**

1. A new feature filtering approach by integrating Information Gain with t-test evaluation metrics for text classification, *International Journal of Advanced Computer Science and Applications*, ***Scopus (Q3) and web of Science index*** **[IF: 1.092]**

**Conference(s):**

1. Comparative analysis of integrating multiple filter-Based Feature Selection methods using vector magnitude score on text classification, *in proceedings - International Conference of Industrial Engineering and Operations Management Society (IEOM)*, 2021, ***Scopus index.***
2. A new feature hashing approach based on term weight for dimensional reduction, in *IEEE proceedings - International Congress of Advance Technology and Engineering (ICOTEN)*, 2021, ***Scopus index***
3. Adaptive and global approaches based feature selection for large-scale hierarchical text classification, *in springer proceedings - 6th International Conference of Reliable Information and Communication Technology (Springer-IRICT), 2021,* ***Scopus and web of Science index***

# CHAPTER 1

# INTRODUCTION

## 1.1 Background of study

Currently, data is growing more extensive not only in terms of size but also in the dimension of features and the number of classes, which are growing in the order of millions and thousands, respectively [1]. These kinds of data are often referred to as "large-scale datasets." Nowadays, several applications need to classify a data with an extremely large number of instances, features, and classes [2]. To effectively analyze and extract valuable information from such types of data, a structured taxonomy of the data must first be defined [2]. As the name implies, this taxonomy or hierarchy is a well-known approach for dealing with large-scale datasets in numerous real-world application domains [3][4]. Various large-scale categorization problems termed "Large-Scale Hierarchical Classification (LSHC)" spine around the HC problem, such as webpage classification [5], image classification [6], music genre classification [7], gene sequence classification [8], and more importantly, document classification [9]. But this thesis focuses on applications that deal with text classification.

Indeed, studies have shown that taxonomy is continually becoming more popular for structuring large-scale text documents. Large-Scale Hierarchical Text Classification (LSHTC) is one fruitful and essential area of research that has to do with the taxonomical classification of large-scale textual data. Moreover, LSHTC has been widely employed in PubMed document classification, international patent records, web document classification, and web directories. The massiveness of the text data not only causes complexity and heterogeneity in such domains but also results in diverse dimensionalities and classes [1][10][11]. Therefore, accompanying the text data and concept space growth results in billions of parameter vectors. This is why the hierarchical classification of an instance among a large number of label classes has achieved significance predominantly in the perspective of large-scale classification [12][13][14][15][16][17]. Even though large-scale textual data with clearly defined inter-category dependency information is advantageous for improving Hierarchical Classification (HC) [3], the scalability problem severely affects the approach. This problem arises due to the high dimensionality produced by text datasets. Several research studies handle or improve the scalability issue by integrating suitable dimensional reduction approaches into the framework of LSHTC [4][18]. As the adoption of dimensionality reduction will enhance the scalability of the LSHTC problem, however, not all of the reduction techniques are efficient. Some drawbacks exist, especially with feature hashing and multi-strategy filtering approaches that inspired this research.

The "Curse of dimensionality" is one of the research challenges and a common problem associated with LSHTC problems, especially when they involve a considerable number of features [19][20][21]. HC models face severe computational issues when dealing with such LSHTC tasks. As said earlier, the concept of “dimensional reduction” is a well-established approach usually used to overcome such a problem (by reducing the storage and processing time requirements) [22]. This technique scales up or improves the performance of HC models by reducing the dimensionality of features set generated in each node of the LSHTC taxonomy [20][23]. The technique, which is established based on machine learning, statistics, and applied fields [24], is used to eliminate those features that are noisy, irrelevant, and redundant. The existing dimensionality reduction techniques comprise different methods that take the original high-dimensional feature space and produce a lower-dimensional feature space that preserves most of the necessary information [23]. These methods that generally keep important information are critical tools utilized as a pre-processing step in various LSHTC problems. However, the existing dimensional reduction approaches (Feature Hashing (FH) and multi-strategy filtering methods) integrated into the LSHTC framework have drawbacks. For FH methods, collisions occur as multiple features are mapped into a single bucket while projecting the original elements into a lower index despite an unused number of buckets exist. In the case of multi-strategy filtering methods, result discrepancies occur when assigning a rank to each feature by the integrated filtering approaches. Both the problems mentioned above result in a critical information loss, consequently sacrificing the performance of

HC models, whose performance seriously depends on the original input dimension [25].

Therefore, this thesis proposed an approach that reduces the size of the features in an LSHTC task to a much lower dimension by solving multiple issues in two stages, thus improving scalability and running time. In the first stage, an approach that enhances the hashing scheme of the existing feature hashing, [26][27] is proposed. It uses term weight to minimize the rates of destructive collisions associated with the existing methods. In the second stage, an improved ranking approach was proposed to address the issue of ranking mismatches associated with existing multi-filtering methods, [28][29][30][31][32].

In the following sections of this chapter, a brief description of the thesis, research problems, aim and objectives, the significance of the study, and the organization of the thesis will be presented.

## 1.2 Problem statement

Large-scale text data is considered the most critical and challenging issue in many real-world application domains [33]. There has been a lot of interest in constructing LSHTC for large-scale text datasets comprising thousands of classes and millions of instances with high-dimensional features [34]. However, for HC models, due to a high number of generated features, the task is tedious, complicated, and takes a more prolonged processing time [35][6]. Data restructuring, feature representation, dimensional reduction, classification, and prediction have been highlighted as the phases of LSHTC research challenges. The dimensional reduction phase plays a vital role in improving the scalability and performance of the LSHTC framework. But the existing dimensional reduction approaches, FH and multi-strategy filtering methods integrated into the framework are unreliable due to bad collisions and result discrepancy.

Nevertheless, bad collisions are an inherent problem present in current FH methods, these collisions occur in the process of hashing features into a lower hash space. This could lead to substantial information loss, mainly when collisions occur between features with different class distributions. Moreover, a single collision can significantly degrade the performance of the HC models. On the other hand, LSHTC has made filter methods complicated as they deal with many features during the feature filtering process. While current approaches that integrate multiple filter methods, known as “multi-filtering approaches,” suffer from result discrepancy problem. The problem occurs due to the different rankings assigned to a single feature by the integrated filter methods [3]. This issue miss-lead the multi-strategy approaches to filter out highly contributory features in the process of features filtering. Nevertheless, using either of the approaches with domain applications that deal with LSHTC tasks, such as international patent records and web directories, could increase error rates in classifying documents.

Therefore, this study focused on reducing each node's feature dimensions within the LSHTC taxonomy. This is done by removing those features that are helpless in discriminating between class labels (child nodes) in the dimensional reduction phase of the LSHTC framework, which consequently scales up HC model performance (by lowering processing time). Besides, the problem of poor performance that arises due to the problem of losing important information as a result of bad collisions and ranking mismatch was also addressed. The following research issues are addressed in this study:

**(i) How to effectively mitigate the occurrence of bad collisions**

Feature Hashing [28][30][31] is one of the dimensional reduction techniques effectively used in reducing high-dimensional features set. FH-based methods take the original input features space and project each feature into a lower defined index. This technique, which deals with sparse features efficiently, is widely used in scaling-up LSHTC tasks. Hash collision is the main problem associated with the methods based on this technique. A single collision could deteriorate the performance of an HC model. Given some particular value of $k$, such that $k << R$, where $R$ is the dimension of the input feature space and $k$ is the smaller size of the hashed space. The hash scheme of the existing FH, [6][26][27] randomly maps original features $k$ to a lower space $R$. Despite the presence of unused buckets, the methods end up bucketing multiple features with different class distributions into the same bucket. This may lead to some collisions, which result in significant information loss. However, reducing the number of unused buckets by avoiding distinct mapping feature into a single bucket will mitigate the rate of bad collisions, consequently improving prediction accuracy. Therefore, an approach that enhances the hashing scheme of the existing FH by using a term-weight to minimize the rate of collisions was proposed in this study.

**(ii) How to efficiently avoid the possibility of filtering out highly informative features**

Irrelevant features are naturally present in an input features set, and they are unwanted features that do not contribute to the discrimination between classes [36][37]. These features generate a lot of problems for the HC models because their presence increases the dimension of feature space by order of a million. As a result, several issues arise, including very high processing time, high memory utilization, and a high risk of over-fitting [11][38][39]. These issues become more and more complex when they are encountered with LSHTC problems. Numerous FS approaches for reducing the feature dimensions have been integrated into the LSHTC framework to overcome the challenges mentioned above. Among the approaches, multi-strategy filtering approaches have shown to be more effective than single strategy approaches but suffer from the problem of result discrepancy. Due to this problem, the existing multi-strategy methods, [28][29][30][31][32] fail to select those features that are highly informative (those features that are ranked highest by one method and the other merged method fails to rank them higher) when two filter methods are integrated. This increases the error rate for class prediction of any incoming new instance. Therefore, in this study, an improved ranking for the multi-strategy approach has been proposed. This efficiently avoids losing those informative features by considering each feature's vector magnitude score and ranking produced by the integrated filter methods. This will significantly contribute to discriminating among the large number of classes in the LSHTC task.

**(iii) How to efficiently improve the scalability of LSHTC problem**

LSHTC is often considered a dataset comprising thousands of number categories and a disproportionately large number of instances with high-dimensional sparse features presentation. Training HC models in the original features space of the LSHTC to discriminate between large numbers of classes falls into scalability problem due to high processing time. In this study, scalability is defined as:

**Definition 1** "*The ability of a model to successfully execute or handle an increasing large amount of textual data that produces millions of parameter vectors by reducing processing time and at the same time improving or maintaining its performance*" [4][3][26][40].

Now, consider a multi-class classification problem with a linear classifier; given a training dataset {x, y} with *m* instances and $\ell$ label classes presented in a *d*-dimensional feature space, where $x \in R^d, y \in \ell, and |y| = \ell$. Therefore, each document (instance) will result in $\ell d$ parameters to train. Let $\Gamma(x, y) = v_y \otimes x$ be the join input-output mapping, where $\Gamma(x, y) \in R^{\ell d}$ is the tensor product of training instances *x* and vector $v_y \in R^{\ell d}$, and all entries of $v_y$ are zero except the $y^{th}$ entry. By learning a parameter vector $w \in R^{\ell d}$, a learning classifier is achieved, such that the class prediction for every input document *x* is given by:

$$\hat{y} = argmax_y w^T \Gamma(x, y) \qquad (1.1)$$

The parameters size can be enormous (in billions) for the LSHTC problem. Moreover, storing and processing such a large number of parameters in a given memory and within possible lower time could be a complex and challenging problem. However, the current dimensional reduction approaches, [3][27][41][42] integrated into the LSHTC problem sacrifice the performance of HC models due to their difficulties of losing essential features. Reducing features size to a much lower dimension and at the same time maintaining important features will lower processing time and improve performance, thus improving scalability. Therefore, this study proposed an approach based on multi-stage dimensional reduction. The approach integrates efficient FH method and multi-filtering approach into the LSHTC framework. This will improve the scalability of LSHTC problems and, at the same time, avoids losing those features that will significantly contribute to discriminating among a large number of classes.

## 1.3 Research aim and objectives

This research aims to improve the scalability and performance of HC models by sequentially solving a couple of issues in the dimension reduction phase of the LSHTC problem. To achieve this goal, we focus on the following objectives:

1. To propose a Modified Feature Hashing approach (MFH) that uses term weight to eliminate the bad collisions between dissimilar features.
2. To propose a Bi-strategy Filtering Approach (BFA) that uses feature vector magnitude score to minimize the problem of result discrepancy which avoid the problem of losing highly informative features.
3. To propose a Multi-stage Dimensional Reduction method (MDRM) for LSHTC which will improve the scalability of HC models by integrating the approaches proposed in objectives (1) and (2).

## 1.4 Research scope

Considering the numerous challenges associated with each phase of the LSHTC problem, this research work focuses only on improving the dimensionality reduction phase in the LSHTC framework. Among the various issues related to dimensionality reduction approaches, this study will focus on minimising the collision rate associated with FH approaches and improving the ranking mismatch associated with multi-filtering approaches.

HC task can be divided into two (2) major approaches: single-label (every child node has only a single parent within the tree taxonomy) and multi-label (child nodes could have multiple parents within the tree taxonomy). Thus, the single-label approach (multi-class classification) was the only one considered in this study.

Moreover, regarding the experimental datasets, this study considers only secondary datasets with their information organized in hierarchies (parent-child relationship), which include 20NewsGroup (20NG) [43], International Patent Classification (IPC) [44], and Directories Mozilla (DMOZ-small) datasets [45]. The nature and properties of the datasets are illustrated in Table 3.1 (in Chapter 3). These datasets are believed to be large-scale with diverse classes, high-dimensional, and sparse. The study measured the effectiveness of the proposed approaches by focusing on some selected evaluation metrics, including micro-fi, macro-f1, hierarchical-f1, hierarchical-error, and running time.

The proposed methods in this study are limited to handling only text classification with defined hierarchies. Finally, the performance evaluation of the proposed methods will be recorded concerning dimensionality reduction only.

## 1.5 Significance of the study

LSHTC task has become the most efficient way of classifying large-scale text datasets in recent years. The task has grown in popularity to organize text documents in various application domains, such as web document classification, web directories, and international patent classification. However, the exponential growth of documents size, the number of features, and the number of classes have raised difficulties for the applications mentioned above when classifying new instances [46]. As a result, this requires an improved scalable approach to overcome the challenges associated with the LSHTC problem. It is crucial to reduce the feature dimensions to enhance the scalability and prediction performance of the LSHTC framework [47][48][49], even though existing studies reduce the number of features by using different dimensionality reduction approaches and techniques. However, the state-of-art approaches proposed in [6][3] are still inadequate for LSHTC problem. Besides, inefficient dimensional reduction may lead to poor document predictions and sometimes long processing time. For addressing these problems, two main approaches have been previously used within the LSHTC framework in different settings (single-stage [27][6][3][29] and multi-stage [50][28][28]). Both the existing single-stage approaches (specifically based on FH technique) and the multi-stage approaches (specifically based on multi-strategy filtering technique) are inadequate for LSHTC problem. For the existing CFH [6] and HFS [3] approach, when a user issues a query request, the framework utilizes either of the approaches to reduce the input features into a lower space before classification. Therefore, for straight-forward solutions:

- The framework uses the FH method in place of the feature vectorization in the dimensional reduction phase of LSHTC to reduce the input feature dimensions: The method directly uses bag-of-words to map each feature in the input space into a lower index dimension.

## REFERENCES


[1] M. Lu, "Embedded feature selection accounting for unknown data heterogeneity," *Expert Syst. Appl.*, vol. 119, pp. 350–361, 2019.

[2] A. Naik and H. Rangwala, "Large Scale Hierarchical Classification : State of the Art," in *Springer Briefs in Computer Science*, Switzerland: Springer, Cham., 2018, pp. 1–104.

[3] A. Naik and H. Rangwala, "Embedding feature selection for large-scale hierarchical classification," in *Proceedings of IEEE Internation Conference on Big Data (Big Data)*, 2016, pp. 1212–1221.

[4] A. Naik and H. Rangwala, "Filter based taxonomy modification for improving hierarchical classification," *arXiv:1603.00772v3 [cs.AI]*, vol. 3, pp. 1–14, 2016.

[5] M. Hashemi, "Web page classification: a survey of perspectives, gaps, and future directions," *Multimed. Tools Appl.*, vol. 79, no. 17–18, pp. 11921–11945, 2020.

[6] B. Zhao and E. P. Xing, "Hierarchical feature hashing for fast dimensionality reduction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015, pp. 2051–2058.

[7] R. Jain, R. Sharma, P. Nagrath, and R. Jain, "Music genre classification ChatBot," *Lect. Notes Networks Syst.*, vol. 203 LNNS, no. 1, pp. 393–408, 2021.

[8] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. D. Kanmani, C. Venkatesan, and C. S. G. Dhas, "Analysis of DNA sequence classification using CNN and hybrid models," *Comput. Math. Methods Med.*, 2021.

[9] R. K. Roul and J. K. Sahoo, "Text categorization using a novel feature selection technique combined with ELM," in *(eds.) Intelligent Computing Computing Techniques. Lecture notes in Advances in Intelligent Systems and Computing*, 2018, vol. 709, pp. 217–228.

[10] M. Soheili and A. M. Eftekhari-moghadam, “DQPFS : Distributed quadratic programming based feature selection for big data,” *J. Parallel Distrib. Comput.*, vol. 138, pp. 1–14, 2020.

[11] M. Habib ur Rehman, C. Sun Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, and S. U. khan, “Big data reduction methods : A survey,” *Data Sci. Eng.*, vol. 1, pp. 265–284, 2017.

[12] R. Babbar, L. Partalas, E. Gaussier, M. Amini, and C. Amblard, “Learning taxonomy adaptation in large-scale classification,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–37, 2016.

[13] L. Wang, Y. Wang, and Q. Chang, “Feature selection methods for big data bioinformatics: A survey from the search perspective,” *Methods*, vol. 111, pp. 21–31, 2017.

[14] I. Yaqoob *et al.*, “Big data: From beginning to future,” *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 1231–1247, 2017.

[15] D. R. Devi and S. Sasikala, “Feature selection and classification of big data using MapReduce framework,” in *(Eds.) Intelligent Computing , Information and Control Systems. Lecture notes in Advances in Intelligent Systems and Computing*, 2020.

[16] A. Unnikrishnan, U. Narayana, and S. Joseph, “Performance analysis of various supervised algorithms on big data,” in *IEEE Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 2293–2298.

[17] M. A. Khan, M. F. Uddin, and N. Gupta, “Seven V ’s of big data understanding: big data to extract value,” in *Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)*, 2014, pp. 1–5.

[18] A. Naik and H. Rangwala, “Integrated framework for improving large-scale hierarchical classification,” in *proceeding of 16th EEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 281–288.

[19] J. Golay and M. Kanevski, “Unsupervised feature selection based on the morisita estimator of intrinsic dimension,” *Knowledge-Based Syst.*, vol. 135, no. Nov, pp. 125–134, 2017.

[20] K. Ikeuchi, *Computer Vision: A reference Guide,* 2014th ed., vol. 2. Boston, USA: Springer, 2014.

[21] M. Li, H. Wang, L. Yang, Y. Liang, and Z. Shang, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Syst. Appl.*, vol. 150, no. July, pp. 1–10, 2020.

[22] J. Luengo, D. García-gil, S. Ramírez-gallego, S. Garcia, and F. Herrera, *Big data preprocessing: Enabling smart data*, Eds. Cham: Springer, 2020.

[23] R. Krishnan, V. A. Samaranayake, and S. Jagannathan, "A hierarchical dimension reduction approach for big data with application to fault diagnostics," *J. Big Data Res.*, vol. 18, p. 100121, 2019.

[24] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2859–2900, 2015.

[25] C. B. Freksen, L. Kamma, and K. G. Larsen, "Fully understanding the hashing trick," in *Prceedings of the International Conference on Neural Information Processing System, NIPS*, 2018, pp. 5394–5404.

[26] Q. Shi *et al.*, "Hash Kernels," in *Proceedings - Machine Learning Research, MLR*, 2009, pp. 496–503, [Online]. Available: http://proceedings.mlr.press/v5/shi09a.html.

[27] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, "Feature hashing for large scale multitask learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09*, 2009, pp. 1113–1120.

[28] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3105–3114, 2017.

[29] K. D. Rajab, "New hybrid features selection method : A case study on websites phishing," *Secur. Commun. Networks*, vol. 2017, no. March, pp. 1–10, 2017.

[30] A. Onan and K. Serar, "A feature selection model based on genetic rank aggregation for text sentiment classification," *J. Inf. Sci.*, vol. 43, no. 1, pp. 25–38, 2015.

[31] F. Kamalov and F. Thabtah, "A feature selection method based on ranked vector

scores of features for classification," *Ann. Data Sci.*, pp. 1–20, 2017.

[32] D. Agnihotri, K. Verma, and P. Tripathi, "Variable global feature selection scheme for automatic classification of text documents," *Expert Syst. Appl.*, vol. 81, pp. 268–281, 2017.

[33] P. Gil, "6 predictions for the $203 billion big data analytics market, Forbes," 2017.

[34] A. Naik, "Hierarchical classification with rare categories and inconsistencies," George Mason University, 2017.

[35] I. Czarnowski and J. Piotr, "An approach to data reduction for learning from big datasets : Integrating stacking ,rotation , and agent population learning techniques," *J. Comlexity*, vol. 2018, no. 1, pp. 1–13, 2018.

[36] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Inf. Fusion*, vol. 59, no. January, pp. 44–58, 2020.

[37] M. Rong, D. Gong, and X. Gao, "Feature selection and its use in big data: Challenges, methods, and trends," *IEEE Access*, vol. 7, pp. 19709–19725, 2019.

[38] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," in *2008 ACM Conference on Recomemender Systems, RecSys'08*, 2008, pp. 259–266.

[39] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, 2016th ed. Greensboro, USA: Springer, 2017.

[40] A. Naik and H. Rangwala, "A ranking-based approach for hierarchical classification," in *proceedings of IEEE Internation Conference on Data Science and Adavanced Analytics (DSAA)*, 2015, pp. 1–10.

[41] A. Rehman, K. Javed, and H. A. Babri, "Feature selection based on a normalized difference measure for text classification," *Inf. Process. Manag.*, vol. 53, no. 2, pp. 473–489, 2017.

[42] H. Zhou, S. Han, and Y. Liu, "A novel feature selection approach based on document frequency of segmented term frequency," *IEEE Access*, vol. 6, pp. 53811–53821, 2018.

[43] K. Albishre, M. Albathan, and Y. Li, "Effective 20 Newsgroup dataset cleaning," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent*

*Agent Technology (WI-AIT)*, 2016, pp. 98-101.

[44] WIPO, “Guide to the International Patent Classification,” 2016.

[45] I. Partalas *et al.*, “LSHTC : A benchmark for large-scale text classification,” *CoRR*, vol. abs/1503, pp. 1–9, 2015.

[46] J. Gao, B. Chin Ooi, Y. Shen, and W.-C. Lee, “Cuckoo feature hashing : Dynamic weight sharing for sparse analytics,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 2135–2141.

[47] C. Caragea, A. Silvescu, and P. Mitra, “Combining hashing and abstraction in sparse high dimensional feature spaces,” in *Proceedings of the 26th AAAI National Conference on Artificial Intelligence*, 2012, pp. 3–9.

[48] D. Singh and C. K. Mohan, “Projection-SVM : Distributed kernel support vector machine for big data using subspace partitioning,” in *proceeding of IEEE International Conference on Big Data (Big Data)*, 2018, pp. 74–83.

[49] Y. Mu, X. Liu, Z. Yang, and X. Liu, “A parallel C4 . 5 decision tree algorithm based on MapReduce,” *Concuracy Comput. Pract. Experiance*, vol. 29, no. 02, pp. 1–12, 2017.

[50] F. Salo, A. B. Nassif, and A. Essex, “Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection,” *Comput. Networks*, vol. 148, pp. 164–175, 2019.

[51] N. Khan, M. S. Husain, and M. R. Beg, “Big data classification using evolutionary techniques : A survey,” in *Preceedings of IEEE International Conference on Engineering and Technology (ICETECH)*, 2015, no. March, pp. 243–247.

[52] S. Nagarajan and R. M. Chandrasekaran, “Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques,” *Indian J. Sci. Technol.*, vol. 8, no. 8, pp. 771–776, 2015.

[53] R. Kumari and S. Kr., “Machine learning: A review on binary classification,” *Int. J. Comput. Appl.*, vol. 160, no. 7, pp. 11–15, 2017.

[54] L. W. Santoso, B. Singh, S. S. Rajest, R. Regin, and K. H. Kadhim, “A genetic programming approach to binary classification problem,” *EAI Endorsed Trans. Energy Web*, vol. 8, no. 31, pp. 1–8, 2021.

[55] F. Gurcan, “Multi-Class classification of Turkish texts with machine learning

algotirthms," *ISMSIT 2018 - 2nd Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, pp. 18–22, 2018.

[56] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," *Futur. Gener. Comput. Syst.*, vol. 101, pp. 341–371, 2019.

[57] R. Alan, P. A. Jaques, and J. Francisco, "An analysis of hierarchical text classification using word embeddings," *Inf. Sci. (Ny).*, vol. 471, pp. 216–232, 2019.

[58] S. Ray, "A quick review of machine learning algorithms," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 35–39, 2019.

[59] A. I. Taloba, D. . Eisa, and S. I. Ismail, "A comparative study on using principle component analysis with different text classifiers," *Int. J. Comput. Appl.*, vol. 180, no. 31, pp. 1–6, 2018.

[60] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *J. Data Anal. Inf. Process.*, vol. 07, no. 04, pp. 190–207, 2019.

[61] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019.

[62] P. . Chaitra and K. Saravana, "A review of multi-class classification algorithms," *Int. J. Pure Appl. Math.*, vol. 118, no. 14, pp. 17–26, 2018.

[63] A. I. Kadhim, "Survey on supervised machine learning techniques," *Artif. Intell. Rev.*, vol. 52, no. 01, pp. 1–20, 2019.

[64] M. Allahyari *et al.*, "A brief survey of text mining: classification, clustering and extraction techniques," in *Proceedings of KDD Bigdas*, 2017, pp. 1–13, [Online]. Available: http://arxiv.org/abs/1707.02919.

[65] Q. Li *et al.*, "A survey on text lassification: from shallow to deep learning," *ACM Trans. Intell. Syst. Technol.*, vol. 37, no. 4, 2020, [Online]. Available: http://arxiv.org/abs/2008.00364.

[66] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "Comparative analysis of text classification algorithms for automated labelling of Quranic

verses," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 4, pp. 1419–1427, 2017.

[67] A. Faraz, "An elaboration of text categorization and automatic text classification through mathematical and graphical modelling," *Comput. Sci. Eng. An Int. J.*, vol. 5, no. 2/3, pp. 1–11, 2015.

[68] F. Fabris, "New Prpbabilitic Graphical Models and meta-laerning approaches for hierarchical classification with applications in bioinformatics and ageing," Unversity of Kent, 2017.

[69] C. Anveshi and H. Rangwala, "HierCost : Improving large scale hierarchical classification with cost sensitive learning," in *Machine Learning and knowledge Discovery in Databases. ECML PKDD. Lecture Notes in computer science*, Eds., vol. 9284, Springer, Cham, 2015, pp. 675–690.

[70] X. Qi and B. D. Davison, "Hierarchy evolution for improved classification," in *International Conference on Information and Knowledge Management, Proceedings*, 2011, pp. 2193–2196.

[71] L. Tang, J. Zhang, and H. Liu, "Acclimatizing taxonomic semantics for hierarchical content classification," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, vol. 2006, pp. 384–393.

[72] A. Naik and H. Rangwala, "Inconsistent node flattening for improving top-down hierarchical classification," in *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, 2016, pp. 379–388.

[73] K. Nitta, "Improving taxonomies for large-scale hierarchical classifiers of web documents," in *International Conference on Information and Knowledge Management, Proceedings*, 2010, pp. 1649–1652.

[74] C. Dong, B. Zhou, and J. Hu, "A hierarchical SVM based multiclass classification by using similarity clustering," in *proceedngs of IEEE International Joint Conference on neural Networks (IJCNN)*, 2015, pp. 1–6.

[75] M. Ramírez-corona, L. E. Sucar, and E. F. Morales, "Hierarchical multilabel classification based on path evaluation," *Int. J. Approx. Reason.*, vol. 68, pp. 179–193, 2016.

[76] M. Sugiyama, *Introduction to Statistical Machine Learning*. Waltham: Elsevier

Inc., 2015.

[77] N. Pilnenskiy and I. Smetannikov, “Feature selection algorithms as one of the python data analytical tools †,” *Futur. internet Artic.*, vol. 54, no. 12, pp. 1–14, 2020.

[78] S. Zhang, X. Chen, and P. Li, “Principal Component Analysis algorithm based on mutual information credibility,”in *Proceeedings of the 2019 International Conference on Computation and Information Sciences, ICCIS*, 2019, pp. 536–545.

[79] A. Juvonen, T. Sipola, and T. Hämäläinen, “Online anomaly detection using dimensionality reduction techniques for HTTP log analysis,” *Comput. Networks*, vol. 91, pp. 46–56, 2016.

[80] G. Kraemer, M. Reichstein, and M. D. Mahecha, “dimRed and coRanking-unifying dimensionality reduction in R,” *R J.*, vol. 10, no. 1, pp. 342–358, 2018.

[81] X. Huang, L. Wu, and Y. Ye, “A Review on Dimensionality Reduction Techniques,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 10, pp. 1–23, 2019.

[82] A. Guru and A. Parveen, “Classification of text data using feature clustering algorithm,” *Int. J. Res. Eng. Technol.*, vol. 3, no. 3, pp. 231–232, 2014, [Online]. Available: http://www.ijret.org.

[83] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny, and R. R. Mostafa, “Improved feature selection model for big data analytics,” *IEEE Access*, vol. 8, pp. 66989–67004, 2020.

[84] D. Sisiaridis and O. Markowitch, “Reducing data complexity in feature extraction and feature selection for big data security analytics,” in *the Proceedings of the 2018 1st International Conference on Data Intelligence and Security, ICDIS 2018*, 2018, pp. 43–48.

[85] G. T. Reddy *et al.*, “Analysis of dimensionality reduction techniques on big data,” *IEEE Access*, vol. 4, pp. 54776–54788, 2020.

[86] A. Meyer-Baese and V. Schmid, *Pattern Recognition and Signal Analysis in Medical Imaging*, 2nd ed. Elsevier Inc., 2015.

[87] W. Sharif, N. A. Samsudin, M. M. Deris, and S. K. A. Khalid, “A technical study on feature ranking techniques and classification algorithms,” *J. Eng. Appl. Sci.*, vol. 13, no. 9, pp. 7074–7080, 2018.

[88] N. Gu, M. Fan, L. Du, and D. Ren, "Efficient sequential feature selection based on adaptive eigenspace model," *Neurocomputing*, vol. 161, pp. 199–209, 2015.

[89] W. Zhao, S. Han, W. Meng, D. Sun, and R. Q. Hu, "BSDP: Big sensor data preprocessing in multi-source fusion positioning system using compressive sensing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8866–8880, 2019.

[90] Y. Yang, Y. li, T. Zhang, P. Gadousey, and J. Liu, "Feature clustering dimensionality reuction based on affininty propagation," *J. Intell. Data Analys.*, vol. 22, pp. 309–1208, 323, 2018.

[91] R. Xu and S. Lee, "Dimensional reduction by feature clustering for regression problems," *J. Info.. Scie.*, vol. 299, pp. 1-12, 2016.

[92] S. Lhazmir, I. El Moudden, and A. Kobbane, "Feature extraction based on principal component analysis for text categorization," in *preceedings of the 6th IFIP International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks, PEMWN*, 2018, pp. 1–6.

[93] Z. M. Hira and D. F. Gillies, "A Review of feature selection and feature extraction methods applied on micraarray data," *Adv. Bioinformatics*, vol. 2015, pp. 1–13, 2015.

[94] A. Subasi, *Practical Guide For Biomedical Signals Analysis Using Machine Learning Techniques*, 1st ed. Elsevier Inc., 2019.

[95] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proceedings of the 2014 Science and Information Conference, SAI*, 2014, pp. 372–378.

[96] M. A. Shayegan and S. Aghabozorgi, "A new dataset size reduction approach for PCA-Based classification in OCR application," *Math. Probl. Eng.*, vol. 2015, pp. 1–14, 2015.

[97] H. Xi, L. Sum, C. Lyu, and X. Wang, "Quantum Locally Linear embedding for nonlinear dimensionality reduction," *J. Quant. Info. Proc.*, vol. 19, no. 309, 2020.

[98] Y. Yan, G. Liu, E. Ricci, and N. Sebe, "Multi-task linear discriminant analysis for multi-view action recognition," *IEEE Trans. IMAGE Process.*, vol. 23, no. 12, pp. 2842–2846, 2015.

[99] W. Li, F. Feng, H. Li, and Q. Du, "Discriminant analysis-based dimensional

reduction for hyperspectral images classification: A survey of the most recent advances and experimental comparison of different techniques," In *IEEE Geosicence and Remote Sensing magazine.* vol. 6, no. 1, pp. 15–34, 2018.

[100] F. Reverter, E. Vegas, and J. M. Oller, "Kernel-PCA data integration with enhanced interpretability," *BMC Syst. Biol.*, vol. 8, no. 2, p. S6, 2015.

[101] D. P. Francis and k. Raimond, "Major advancement in kernel function approximation," *Artif. Intell. Rev.*, vol. 54, pp. 843-876, 2021.

[102] M. Yousaf, T. U. Rehman, and L. Jing, "An extended ISOMAP approach for nonlinear dimensional reduction," *SN Comput. Scie.*, vol. 1, no. 160, pp. 1–17, 2020.

[103] M. Jagadeesan, "Understanding sparse JL for feature hashing," in *Proceeding of Advances in Neural Information Processing Systems, NeurlPS 2019*, 2019, pp. 1–31.

[104] S. Vora and H. Yang, "A comprehensive study of eleven feature selection algorithms and their impact on text classification," in *proceeding of Computing Conference*, 2017, no. July, pp. 440–449.

[105] L. Zhang, L. Jiang, and C. Li, "A new feature selection approach to Naive Bayes text classifiers," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 30, no. 2, pp. 1–17, 2017.

[106] K. Li, M. Yu, L. Liu, T. Li, and J. Zhai, "Feature selection method based on weighted mutual information for imbalance data," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 28, no. 8, pp. 1177–1194, 2018.

[107] Y. Wei, L. Jiao, R. Mehmood, H. Liu, A. Umek, and A. Kos, "Hierarchical feature reduction with max relevance and low dimensional embedding strategy and its application in activity recognition with multi-sensor," *Procedia Comp. Scie., vol. 129*, pp. 284-290, 2017.

[108] H. Liu, S. Member, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Autom. Sin.*, vol. 6, no. 3, pp. 703–715, 2019.

[109] S. A. Shahee and U. Ananthakumar, "An effective distance based feature selection approach for imbalance data,"*Appl. Intell.*, vol. 50, pp. 717–745, 2020.

[110] Z. Hongfang, W. Xiqian, and Z. Yao "Feature selection based on weighted conditional mutual information," *Appl. Compt. and Info.*, vol. 17, no. 2, 2020.

[111] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang "Benchmark for filter methods for feature selection in high-dimensional classification data," *J. Compt. Statis. Data Anal.*, vol. 143, p. 106839, 2020.

[112] L. N. H. Nam and H. B. Quoc, "A combined approach for filter feature selection in document classification," in *Proceedings of the International Conference on Tools with Artificial Intelligence, ICTAI*, 2016, vol. 2016-Janua, pp. 317–324.

[113] L. Ma, M. Li, Y. Gao, T. Chen, X. Ma, and L. Qu, "A Novel Wrapper Approach for Feature Selection in Object-Based Image Classification Using Polygon-Based Cross-Validation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 409–413, 2017.

[114] A. K. Das, S. Sengupta, and S. Bhattacharyya, "A Group Incremental Feature Selection for Classification using Rough Set Theory based Genetic Algorithm," *Appl. Soft Comput. J.*, vol. 64, no. April, pp. 400–411, 2018.

[115] G. Jesus and Q. G. John "A new multi-objective wrapper method for feature selection - accuracy and stability analysis for BCI," *Neurocomputing*, vol. 333, no. 1–2, pp. 407–418, 2017.

[116] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, 2015.

[117] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, "Feature selection methods and genomic big data : a systematic review," *J. Big Data*, vol. 6, no. 79, pp. 1–24, 2019.

[118] R. Chakraborty and N. R. Pal, "Feature selection using a neural framework with controlled redundancy," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 1, pp. 35–50, 2017.

[119] T. Kari *et al.*, "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm," *IET Gener. Transm. Distrib.*, vol. 12, no. 21, pp. 5672–5680, 2018.

[120] L. Cui, L. Bai, Z. Zhang, Y. Wang, and E. R. Hancock, "Identifying the most

informative features using a structurally interacting elastic net," *Neurocomputing*, vol. 336, pp. 13–26, 2019.

[121] D. Loanmi, "A review on variable selection in regression analysis," *Econometrics*, vol. 6, no. 25, pp. 1–23, 2018.

[122] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, pp. 225–231, 2020.

[123] S. Haris B. and B. Revanasidappa M., "A comprehensive survey on various feature selection methods to categorize text documents," *Int. J. Comput. Appl.*, vol. 164, no. 8, pp. 1–7, 2017.

[124] Y. Liu, S. Ju, J. Wang, and C. Su, "A new feature selection method for text classification based on independent feature space search," *Math. Probl. Eng.*, vol. 2020, pp. 1–14, 2020.

[125] A. Ado, N. A. Samsudin, M. M. Deris, and A. Ahmed, "Comparative analysis of integrating multiple filter-based feature selection methods using vector magnitude score on text classification," in *Proceedings of 11th Annual International Conference on Industrial Engineering and Operations Management (IEOM)*, 2021, pp. 4664–4676.

[126] M. Pérez-Ortiz, M. Torres-Jiménez, P. A. Gutiérrez, J. Sánchez-monedero, and C. Hervás-Martínez, "Fisher score based feature selection for ordinal classification: A social survey on subjective well-being," in *(eds.) Hybrid Artificial Intelligent Systems, HAIS*. Lecture Notes in Computer Science, vol. 9648, 2016, Springer, Cham.

[127] S. Fan, Y. Jia, and C. Jia, "A feature selection and classification method for activity recognition based on an inertial sensing unit," *Inf.*, vol. 290, no. 10, pp. 1–21, 2019.

[128] A. Ado, M. M. Deris, S. Noor Azah, and A. Aliyu, "A new feature filtering approach by integrating IG and T-test evaluation metrics for text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 500–510, 2021.

[129] N. O. Essied, I. Othman, and A. H. Osman, "A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification," *Res. J. Appl. Sci. Eng. Technol.*, vol. 7, no. 3, pp. 625–638, 2015.

[130] D. Jain and V. Singh, "An efficient hybrid feature selection model for dimensionality reduction," in *Proceedings of the International Conference on Computational and Data Science*, 2018, vol. 132, pp. 333–341.

[131] U. Ryan, M. Meeker, W. LaCava, R. Olson, and J. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Infor.*, vol. 85, 2017.

[132] A. Wosiak and D. Zakrzewska, "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis," *Complexity*, vol. 2018, pp. 1–11, 2018.

[133] A. M. Kowshalya, R. Madhumathi, and N. Gopika, "Correlation based feature selection algorithms for varying datasets of different dimensionality," *Wirel. Pers. Commun.*, vol. 108, no. 3, pp. 1977–1993, 2019.

[134] X. Y. Liu, Y. Liang, S. Wang, Z. Y. Yang, and H. S. Ye, "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection," *IEEE Access*, vol. 6, pp. 22863–22874, 2018.

[135] A. Brankovic, M. Hosseini, and L. Piroddi, "A distributed feature selection algorithm based on distance correlation with an application to microarrays," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 16, no. 6, pp. 1802–1815, 2019.

[136] H. Tan, G. Wang, W. Wang, and Z. Zhang, "Feature selection based on distance correlation: a filter algorithm," *J. Appl. Stat.*, vol. 49, no. 2, pp. 411–426, 2022.

[137] Y. Wang and L. Feng, "A new hybrid feature selection based on multi-filter weights and multi-feature weights," *Appl. Intell.*, vol. 49, no. 12, pp. 4033–4057, 2019.

[138] X. Wang, B. Guo, Y. Shen, C. Zhou, and X. Duan, "Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization," *IEEE Access*, vol. 7, pp. 151525–151538, 2019.

[139] W. Gao, L. Hu, P. Zhang, and F. Wang, "Feature selection by integrating two groups of feature evaluation criteria," *Expert Syst. Appl.*, vol. 110, pp. 11–19, 2018.

[140] D. S. Guru, M. Suhil, L. N. Raju, and N. V. Kumar, "An alternative framework for univariate filter based feature selection for text categorization," *Pattern Recognit. Lett.*, vol. 103, no. January, pp. 23–31, 2018.

[141] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A chi-square statistics based

feature selection method in text classification," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2019, vol. November, pp. 160–163.

[142] A. Ado, N. A. Samsudin, and M. Mat Deris, "A New feature hashing approach based on term weight for dimensional reduction," in *proceedings - IEEE Internation Congress of Advance Technology and Engineering (ICOTEN)*, 2021, pp. 1–7.

[143] B. Dalessandro, "Bringing the noise : Embracing randomness the key to scaling up machine learning algorithms," *Big Data*, vol. 1, no. 2, pp. 110–112, 2014.

[144] D. Soren, B. Mathias, K. Tejs, and T. Mikkel, "Practical hash functions for similarity estimation and dimensionality reduction," in *Proceedings of 31st International Conference on Neural Information Processing Systems, NIP'17*, 2017, pp. 6618–6628.

[145] V. Akoto-Adjepong, M. Asante, and S. Okyere-Gyamfi, "An enhanced non-cryptographic hash function," *Int. J. Comput. Appl.*, vol. 176, no. 15, pp. 10–17, 2020.

[146] Tanjent, "MurmurHash First Announcement,", 12-Dec-2018, [Online] Available: *Tanjent.livejournal.com.*

[147] P. Saxena, "Analysis of various hash function," *Int. J. Innov. Sci. Res. Technol.*, vol. 3, no. 5, pp. 217–220, 2018, [Online]. Available: www.ijisrt.com.

[148] A. Pranckevičius, "Hash functions all the way down,", 2-Aug-2016, [Online] Available*: https://aras-p.info/blog/2016/08/02/Hash-Functions-all-the-way-down*

[149] C. Caragea and A. Silvescu, "Protein sequence classification using feature hashing," 2011.

[150] P. Li, A. Shirvastava, J. Moore, and C. K. Arnd, "Hashing algorithms for large-scale learning," in *Prceedings of the 24th International Conference on Neural Information Processing System, NIPS'11*, 2011, pp. 2672–2680.

[151] L. Zhou, M. Li, D. G. Andersen, and A. J. Smola, "Cuckoo linear algebra," in *Proceedings of The 5th ACM International Conference on KDD*, 2015, no. 1, pp. 1553–1562.

[152] H. Uguz, "Knowledge-based systems: A two-stage feature selection method for

text categorization by using information gain , principal component analysis and genetic algorithm,” vol. 24, pp. 1024–1032, 2011.

[153] S. Sarode and G. Jayant, “Hybrid dimensionality reduction approach for web page classification,” in *IEEE International Conference on Communication, Information & Computing Technology (ICCICT)*, 2015, pp. 1–6.

[154] S. Sasikala, S. Appavu alias Balamurugan, and S. Geetha, “Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set,” *Appl. Comput. Informatics*, vol. 12, no. 2, pp. 117–127, 2016.

[155] C. Wan, Y. Wang, Y. Liu, J. Ji, and G. Feng, “Composite Feature Extraction and Selection for Text Classification,” *IEEE Access*, vol. 7, pp. 35208–35219, 2019.

[156] N. Priatna, A. F. Huda, Q. . Safitri, and W. Darmalaksana, “Hybrid reduction dimension on clustering text of english hadith translation,” in *Proceedings of IEEE 5th International Conference on Wireless and Telematics (ICWT)*, 2019, pp. 1–5.

[157] E. Demeke and T. Tegegne, “Designing a hybrid dimension reduction for improving the performance of Amharic news document classification,” *PLoS One*, vol. 16, no. 5, pp. 1–14, 2021.

[158] M. Anandarajan, C. Hill, and T. Nolan, “Text preprocessing,” in *[Advances in Analytics and Data Science] Practical Text Analytics (Maximizing the Value of Text Data)*, vol. 2, Switzerland: Springer Nature, 2019, pp. 45–59.

[159] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving Text preprocessing for student complaint document classification using sastrawi,” in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 874, no. 1.

[160] A. I. Kadhim, “An evaluation of preprocessing techniques for text classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.

[161] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, *Empirical evaluation and study of text stemming algorithms*, vol. 53, no. 8. Springer Netherlands, 2020.

[162] K. M. M. Rajashekharaiah, S. S. Chikkalli, P. K. Kumbar, and P. S. Babu, “Unified framework of dimensionality reduction and text categorisation,” *Int. J. Eng.*

*Techology*, vol. 7, no. November 2018, pp. 648–654, 2018.

[163] S. Rahamat Basha, J. Keziya Rani, and J. J. C. Prasad Yadav, "A novel summarization-based approach for feature reduction enhancing text classification accuracy," *Eng. Technol. Appl. Sci. Res.*, vol. 9, no. 6, pp. 5001–5005, 2019.

[164] N. Sun, B. Sun, J. D. Lin, and M. Y. Wu, "Lossless pruned Naive Bayes for Big data classifications," *Big Data Res.*, vol. 14, no. May, pp. 27–36, 2018.

[165] B. Paula, T. Lu´ıs, and R. Rita P., "Relevance-based evaluation metrics for multi-class imbalanced domains," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10235 LNAI, pp. 698–710, 2017.

[166] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-Class classification: An overview," *ArXiv[ML.cs]*, vol. abs/2008.0, pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756.

# VITA

The author was born to the family of Late Alhaji Ado Rogo of Zoo Road Quarters, opposite Zoo Logical Garden, Tudum Maliki, Kano, Nigeria, on 26th July, 1984. He attended Iman Nursery and Primary School from 1990 to 1995. He obtained West African Senior Secondary Certificate Examination (WASSCE) in 2006 from Dawakin Kudu Science College (DKSC), Kano. He was offered admission into Abubakar Tafawa Balewa University, ATBU, Bauchi, in 2002/2003 Academic Session to Study Computer Science. He was awarded B.Tech (Hons) Computer Science in 2009 by ATBU, Bauchi. Upon graduation, He served the Nigerian Nation under the mandatory one year National Youth Service Corp (NYSC) between 2009 and 2010, in Kano state Board of Internal Revenue in Nigeria. The author was offered an appointment as Graduate Assistant with the Department of Computer Science, in 2012, by the management of Yusuf Maitama Sule Uiversity (YUMSUK), Kano formerly known as Northwest University (NWU), Kano. He earned Masters in Computer Science and Technology from Liaoning University of Technology (LUT), Jinzhou, China. He was offered an admission by UTHM to undergo PhD in Information Technology in 2019/2020 academic session. He is a registered Nigerian Computer Society (NCS) and an Associate Member of Computer Professional of Nigeria (CPN). He has published several journal and conference papers. He has numerous academic awards and prizes in his honour. The author is happily married to Ruqayya Bello Gaya. The marriage is blessed with Ameer formally known as Adam.