Feature Selection Techniques for High-Dimensional Data Analysis: Applications, Challenges, and Future Directions

Philemon Uten Emmoh[1*], Christopher Ifeanyi Eke[2], Timothy Mosese[3], Agushaka J. Ovre[4]

[1]Department of Computer Science, Federal University Wukari, Wukari, Taraba State, Nigeria.
[2,3,4]Department of Computer Science, Federal University of Lafia, Lafia, Nasarawa State, *Nigeria*.

ABSTRACT: The curse of dimensionality is a major concern in high-dimensional data. As the challenge persists, researchers have adapted the feature selection process to reduce the dimensionality by selecting the important features. This research aims to conduct a systematic literature review on feature selection techniques in high-dimensional data. The goal is to examine trends in feature selection techniques between 2015 and 2024 to help guide researchers in conducting future research. The study examined five (5) research questions on feature selection techniques in high-dimensional datasets to discover prominently used feature selection techniques between 2015 and 2024. The study also examined machine learning models adapted over the years and feature selection issues identified from other studies suggested solutions to these issues, identified the application for the feature selection techniques and carried out potential future research directions. Five internet-based databases were used to select 40 primary studies to carry out the comprehensive study. The findings reviewed 50 feature selection techniques with the chi-square technique as the most prominent technique used by researchers. A total of 37 machine learning models were identified with support vector machine as the most prominent model used by researchers. This review can help researchers to enhance the efficacy of their techniques and models in future investigations.

KEYWORDS: Application area**,** feature selection techniques, future directions, high-dimensional data, machine learning models.

[Received Sept. 29, 2024; Revised Feb. 26, 2025; Accepted March 14, 2025] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

This High-dimensional data is a dataset in which the number of features in a dataset is larger than the number of instances in the dataset (Narisetty, 2020). As data collection and storage capabilities expand, high-dimensional data is becoming more common across industries. Conventional analysis methods may fail to identify complex relationships and trends in data because there are more features to take into consideration. The challenges of the curse of dimensionality include degradation of predictive accuracy, overfitting, increased computational complexity, and increased learning time (Karimi, *et al*., 2023). Overfitting refers to a condition in which a model performs well on trained data but do not perform well on new data (Edlund et al., 2021). Many scholars are attempting to address this problem by suggesting techniques to reduce the data dimensionality to extract meaningful information from this massive volume of data, which is a major challenge (Hashemi et al., 2022).

Feature selection is one of the effective methods used in high-dimensionality data reduction which selects a subset of the features that have the highest relevance to the target class and the lowest inner similarity. Through the process of getting rid of irrelevant, redundant, or data that is noisy, the dimensionality of the data is reduced (Rostami et al., 2021). Since the feature selection technique is essential in high dimensional data reduction, researchers find it difficult to decide the appropriate feature selection technique and machine learning model to be used during the data pre-processing stage and also find it difficult to identify the current issues with various feature selection techniques and machine learning models for researchers take appropriate decision when dealing with feature selection. This study will help researchers to identify the most frequently used feature selection technique and machine learning model used by other researchers from 2015 to 2024 for future research studies. The three basic feature selection methodologies in the literature are the filter, wrapper, and embedded methods by Pudjihartono *et al.*, (2022); Theng and Bhoyar (2024). Filter methods select features based on statistical measures to evaluate the important features from the unimportant features. They are used in pre-processing steps without depending on any machine learning algorithms (Hopf & Reifenrath, 2021). Following the filter-based feature selection procedure, any classifier may be used to assess the high quality of the features that have been selected from a dataset. As a result, the filter-based approach is quick and easy to use. Wrapper methods use the predicted precision of a preset learning method to evaluate the quality of the selected subsets (Bouchlaghem *et al.* 2022). Embedded feature selection (FS) techniques incorporate the FS machine learning

*Corresponding author: philemon@fuwukari.edu.ng *doi:* http://dx.doi.org/10.4314/njtd.v22i1.2943

algorithm as a learning algorithm component, allowing feature selection and classification to operate simultaneously. The importance of feature selection in a high-dimensional dataset is not only to improve the predictive accuracy performance of a model but to also improve the efficiency of the machine learning algorithm (Khan *et al*.2022).

There are so many articles that reviewed dimensionality reduction. Alhassan and Wan (2021) mentioned three dimensionality reduction techniques, which are, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Generalized Discriminant Analysis (GDA). Muhammad *et al*. (2019) used the PCA dimensionality reduction technique toperform a polytomous variables categorization that can be used to classify variable-based diabetes and other chronic diseases into meaningful clusters. The PCA results were modelled using the k-nearest neighbors KNN model to produce an optimal classification process. According to the study by Rasitha (2016), the LDA technique was used for the detection of hypothyroid disease prediction accuracy. They applied the LDA technique alongside the K-fold cross-validation on the hypothyroid disease dataset to obtain the prediction accuracy of the disease. The LDA technique gave an accuracy of 99.62%. In their research, they could not apply the LDA technique to other disease datasets to determine its prediction power. Bahmaninezhad and Hansen (2016) applied Generalized Discriminant Analysis (GDA) on an i-vector-based speaker recognition using Probabilistic Linear Discriminant Analysis (PLDA) to improve the occurrence of the regular test utterances and in the evaluation of for short duration segments on a 10-sec test condition. The equal error rate for the cosine kernel employed using the technique was 20%. In their research work, the feature selection technique was not applied to select the most important features for better accuracy predictions. A study by Uten, *et al*., (2025) proposed a feature selection technique based on statistical lift measure to select important features from a dataset. The proposed technique is a generic approach that can be used in any binary classification dataset problem. The proposed technique was tested on a lung cancer dataset and also on a happiness classification dataset. The effectiveness of the proposed technique in selecting important features subset was evaluated and compared with other existing techniques, namely Chi-Square, Pearson Correlation, and Information Gain. The proposed and the existing techniques were evaluated on five machine learning models using four standard evaluation metrics such as accuracy, precision, recall, and F1-score. The proposed technique outperformed the existing techniques in both the lung cancer and happiness classification datasets. Eke *et al.* (2021) proposed a multi-feature fusion framework for sarcasm identification on Twitter data using machine learning models to construct the lexical feature extraction using the bag-of-words (BoW) technique, and trained using five standard classifiers, including support vector machine (SVM), decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR) and random forest (RF). SVM, DT, KNN, LR, and RF, to predict the sarcastic tendency. Information gain and Pearson correlation techniques were used to select the important features. Their results show that, the classification models based on the proposed new feature fusion model obtained higher performance result with a precision of 0.947 when using Random Forest classifier. The results were compared with the results of three baseline approaches and outperformed the three baseline approaches. It is important to use machine learning model after the selection of the important features by the feature selection techniques to perform an optimal accuracy prediction. According to *Eke et al.*, (2019), supervised machine learning models are the most prominent used models, this is because, and the trained data helps the system to learn and categorized the results according to the inputs.

The key contributions of this study are as follows:

1. To identify the frequently used feature selection techniques in high dimensional data between 2015 and 2024 to help researchers make better decisions when selecting an appropriate technique to use for prediction modelling
2. To identify the current issues with some feature selection techniques on high dimensional data to help researchers make appropriate decision when dealing with feature selection.
3. To get to know researchers, academics, and professionals with the concept of systematic reviews and meta-analyses on feature selection techniques for high-dimensional data analysis.
4. To examine machine learning models applied between 2015 – 2024 on various application domains and to identify which of the models was prominently used on different application domains
5. To identify the current trends and future directions of feature selection on high dimensional data within the last decade

The remainder of this research is as follows: Section II discussed the methods that were employed to conduct this research work. Section III presents the data extraction and synthesis of the research, while Section IV reviews the findings and discussions of results. The conclusion of the work is presented in Section V.

II. RESEARCH METHODOLOGY

We adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines in this systematic literature review. However, there are so many guidelines available to conduct a systematic literature review such as the Cochrane Reviews protocol which is mostly used in the health domain (Higgins & Green, 2008), and the proposed protocols that focus on software engineering (Kitchenham, 2007). This research work adopted the systematic literature review guidelines proposed by Kitchenham and Brereton (2013) to carry out the systematic review. This guideline has become so common that most researchers adopt it. It is noticed that the systematic literature review is mostly used in the software engineering research area, but many other research domains have adopted the systematic approach guidelines in conducting a systematic literature review. The PRISMA gives a well-organized framework to communicate the procedures and to draw a better conclusion in a systematic literature review study. It follows

the principle of planning, conducting, and reporting used in systematic literature review process phases. Table 1 explains these 3 phases.

Table 1: Summary of research methodology

SLR Planning	-	**Define Research Questions**
	-	**Identify Data Sources**
	-	**Search Strategy**
SLR Conducting	-	Selection and Execution
	-	Quality Assessment
	-	Data Extraction and Synthesis
SLR Reporting	-	Review Findings and Discussions
	-	Conclusion

A. Research Questions and Objectives

The purpose of this systematic review study is to conduct an overview of research in feature selection techniques in high-dimensional data to identify the feature selection techniques, their impact, challenges, and possible solutions, application area, and research trends to pick out the total and kind of studies and available results. We used the research questions in Table 2 to conduct our systematic literature review.

Table 2. Research questions and research objectives.

	Research questions	Research Objectives
RQ1	What are the most commonly used feature selection techniques for high-dimensional data analysis?	To identify the feature selection (FS) techniques used from 2015-2024
RQ2	What are the most prominently used machine learning models for high-dimensional data analysis?	To examine the machine learning models employed over the last decade
RQ3	What are the primary challenges associated with feature selection in high-dimensional data analysis, and how are these challenges addressed in the literature?	To investigate feature selection issues experienced and possible solutions to the issue
RQ4	What are the main applications of feature selection techniques in various domains dealing with high-dimensional data?	To identify the application areas of FS techniques
RQ5	What are the emerging trends and prospects of feature selection for high-dimensional data analysis as identified in recent research?	To find out trends and future directions in feature selection for high-dimensional data

B. Search Strategy

We used different components of search techniques such as the search keyword, the search strategy, and the literature tools in this research. The search keyword was developed using the Population, Intervention, Comparison, and Context (PICOC) standards developed by (Mohamed *et al*., 2021). We also adopt the principle of Eke *et al*. (2023) to begin a search by applying keywords to query articles for a particular study. We obtained the relevant articles from five internet-based databases shown in Table 3.

Table 3: Internet-Based Databases

	Databases	Links
1	Science Direct	https://www.sciencedirect.com/
2	IEEE Xplore	https://ieeexplore.ieee.org/
3	Scopus	https://scopus.com
4	Google Scholar	https://scholar.google.com/
5	Springer Link	https://link.springer.com/

All the keywords were included in the Search Query (SQ) which are illustrated as follows:

Query 1 = (Feature Selection Techniques OR Feature Selection approach OR Feature Selection algorithm)

Query 2 = ("Machine learning models OR classifiers")

Query 3 = ("Feature selection Issues OR challenges OR drawbacks AND solutions")

Query 4 = ("Application areas OR domain areas")

query 5 = ("Feature selection trends OR future direction")

We therefore have:

Search Query = Query 1 AND Query 2 AND Query 3 AND Query 4 AND Query 5

The query can also be stated as (Feature selection techniques OR approach OR algorithm) AND (Machine Learning Models OR classifiers) AND (Feature Selection Issues OR Challenges OR drawbacks OR solutions AND Solutions to Feature Selection Issues) AND (Application areas OR domain areas) AND (feature selection trends OR future directions). The systematic literature covers published articles from 2015-2024. Conducting a systematic literature review (SLR) requires a thorough comprehensive study of all the relevant materials. Five internet-based databases were employed to search for research papers using previously created search phrases. Because the search engines inside each database apply different search term syntax, keywords are changed to fit each database. The literature tools adopted to carry out this comprehensive research work were taken from reputable databases such as Science Direct, IEEE Xplore, Scopus, Google Scholar, and Springer Link.

C. Selection Criteria

The resources used in this research were taken from different online databases articles that deal with feature selection, machine learning models, feature selection issues, solutions to feature selection issues, application domain, and future directions from between 2015-2024. A specific criterion has to be established to ascertain if an article can be used in the systematic literature review or not. The most relevant articles were found using a specific criterion for including and excluding articles. Kitchenham and Charters (2007) explain publication bias as a problem when positive results are more likely to be published than negative results. As such the selection criteria are chosen in advance to avoid the chance of bias. Studies that met the purpose of the research study were carefully selected, while those that did not meet the research goals were discarded.

This research study employed the PRISMA guidelines to report the search results to include or exclude primary studies in the research analysis. After the identifications and compilations of the related studies obtained from the databases, the search results were uploaded to Mendeley Reference Manager to automate the citations and bibliography of this

research work. At the end of the article queries, all duplicated articles were identified and deleted immediately. To carry out the article's selection process, 40 primary studies were chosen for this systematic literature review that includes 9 conference proceedings and 31 journals from 2015-2024. The analysis of these papers is depicted in Figure 1.

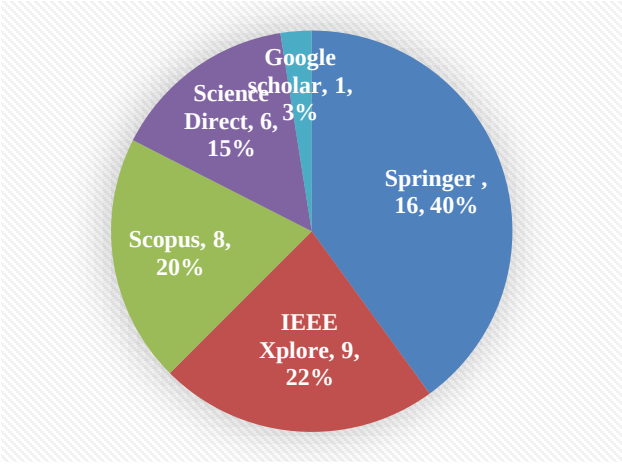

Figure 1: Distribution of Internet-based databases with the selected papers

In the distribution of papers published from 2015-2024, two primary selected papers were published in 2015 and 2016 each, five were published in 2017, three were published in 2018, seven were published in 2019, and six primary studies were selected in 2020. Meanwhile, five primary studies were selected in 2021, three were selected in 2022, four primary studies were selected in 2023 and finally, two were selected in 2024. The distribution of the selected papers over the last decade is shown in Figure 2. For an article to be included or excluded, the criteria in Table 4 depict the processes required to include and exclude a study.

D. Selection Execution

This study adopted the systematic literature review guideline established by Petersen *et al.* (2008) that required the study of the articles' abstracts, searching the articles using the research keywords to select the required papers that match the research questions and research objectives. We therefore used the following selection process steps to pick the relevant papers for the research.

1. Discarding duplicate papers: with the help of the Mendeley referencing manager software, all duplicated papers were identified and removed
2. Title and abstract filtering: during the search process, we carefully looked at the title and the abstract of the paper to ensure that it captured one or more of the research keywords used in the research questions and research objectives
3. Full-text filtering: the full content of the papers was downloaded and scrutinized as it relates to the study selection criteria

From the total of 190 papers reviewed,19 papers were detected as duplicates and were removed, 52 papers were screened after reading the title and the abstract, while 40 papers were later screened after reading the full text of the papers. At the end of the selection process, 119 papers were excluded while 40 papers fulfilled the inclusion criteria used in the study. This process of identifying, screening, and selecting primary studies can be seen in Table 5. The PRISMA flow diagram for this research study is shown in Figure 3.

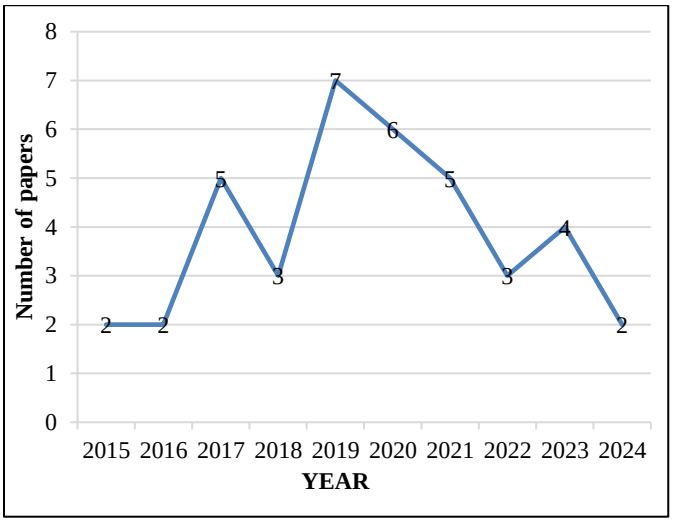

Figure 2: Number of selected studies over the years

We chose to use the following research databases for this study because they are the most popular search engines that researchers are most interested in. We included Google Scholar because it will pick up some references that might not be in other publisher databases. Science Direct, IEEE Xplore and Springer are all publisher databases while Google Scholar and Scopus are each scholarly databases that include papers from lots of publishers.

It is necessary to use these databases in this study because Science Direct offers access to a wide range of Scientific and technical research articles in the fields of physical sciences, life sciences, health sciences, and engineering that includes a large number of journals, conference proceedings, and many others. Scopus is a multidisciplinary abstract and citation database that covers a wide range of fields such as arts, humanities, social science, and science that includes a large number of high-quality journals, conference proceedings. Google Scholar offers a broad range of materials such as journal articles, conference papers, theses, books, and preprint; IEEE Xplore is also one of the most prominent professional organizations in the field of electrical engineering and computer science with journal papers and conference proceedings that are highly regarded in the academic sectors. Springer also typically covers a broad range of topics within electrical engineering. By the quality and versatility of these databases which cut across different fields of research, we adopted them for this study since our study covers a multidisciplinary field of study.

Table 4: Inclusion and Exclusion criteria

Inclusion	Criteria	Exclusion	Criteria
In.1	Feature selection techniques papers published between 2015-2024	En.1	Paper that does not reflect the research topic
In.2	Only peer-reviewed articles or conferences	En.2	Not peer-reviewed articles or conferences
In.3	Only paper written in the English language	En.3	Papers published before 2015
In.4	Papers that answered all the research questions	En.4	Papers not written in the English language
In.5	Only primary study articles are selected	En.5	Papers not accessible
		En.6	Papers that do not answer all the research questions

Table 5: Research papers identified for the systematic literature review

Internet-based database	Initially reviewed papers	Screened based on title and abstract	Screened based on full-text reading
Science Direct	41	9	6
Scopus	19	11	8
Springer	53	17	16
IEEE	32	14	9
Google Scholar	5	1	1
Total	150	52	40

E. Quality Assessment (QA)

Quality assurance is very important in this research because that will add value to the huge number of articles researched. Yang *et al.* (2021) stated that there is no agreed standard definition of study quality, as such assessment quality of study varies, most especially between secondary and tertiary studies. To evaluate the rigor, accuracy, and relevance of associated research, a set of quality evaluation questions was created. We evaluate the quality of papers at this stage. Therefore, the approaches used to assess the quality of the 40

papers are adopted from Kitchenham and Charters (2007) guidelines. For us to be sure of the findings of this research survey, we conducted an in-depth discussion about relevant quality research to ascertain the reliability of the research conclusions. We reviewed and chose pertinent research based on the quality assessment questions listed in Table 6.

Table 6: Quality Assessment criteria used in this study

QA	Description
QA1	Does the study use any feature selection techniques?
QA2	What machine learning algorithms were used in the study?
QA3	What are the feature selection issues observed?
QA4	How are these issues resolved?
QA5	What are the future directions of feature selection in the study?

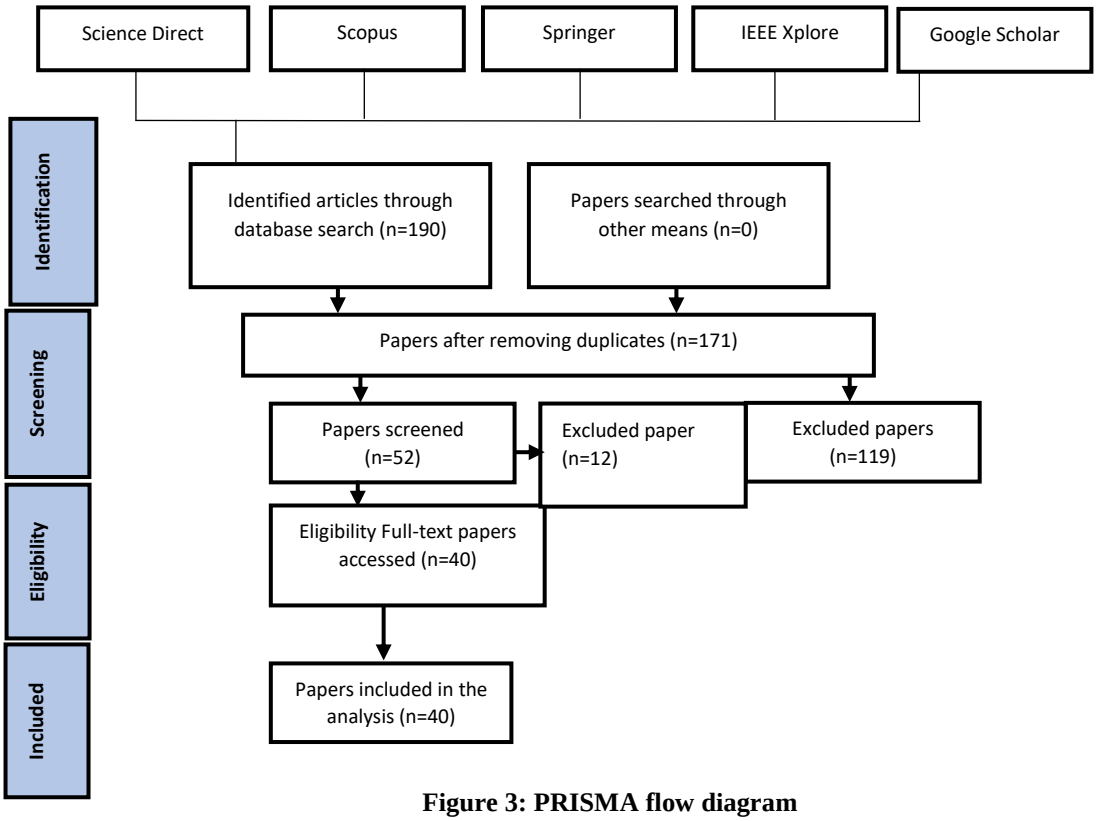

Figure 3: PRISMA flow diagram

III. DATA EXTRACTION AND SYNTHESIS

In this section, we investigate all the final 40 selected papers. The data is synthesized to address the five research questions. At this stage, the process of data extraction is documented through the PRISMA flow diagram. We created a data extraction form for the 40 papers as seen in Table 7 to collect relevant information from the 40 papers.

Table 7: Description of data synthesis

S/N	Class	Explanation
1	Paper ID	The paper's unique assigned number
2	Author(s)	Article author's names
3	Year	Year of article publication
4	Aim of the research study	The purpose that prompted the study
5	Feature selection technique(s)	The FS techniques used in the study
6	Machine learning models	The classifier(s) employed in the study
7	FS challenges	Issues observed in the study
8	FS solutions	How the FS challenge was resolved
9	Application domain	Field(s) where the FS technique(s) being applied
10	Paper types	The type of journals and conference proceedings
11	Databases	Databases from which papers were sourced
12	Results obtained	Analysis of results obtained in the study

The data synthesis phase in the systematic literature review involves the summary and interpretation of information gathered from the selected papers (Al-sharif *et al.*, 2024). It purposely targets the research questions through analysis, discussion, and various forms of representation by using charts, tables, and other representation forms. We therefore carefully applied the data extraction form on the 40 selected articles as shown in Table 7. We adapted the standard information extraction form from the work of Kitchenham *et al*. (2013).

A. *Publication Source*

About 40 papers were selected for this systematic literature review study which includes 9 conference proceedings and 31 journals with their respective authors, publication years, domain area, and database source of publication as shown in Table 8.

B. Review of the Publication Year

Figure 2 depicts the 40 articles publication year and number of papers between 2015 and 2024 which is a research that covers the last decade. We started in 2015 with little interest in the research area, which also extends to 2016. In 2017, there was increasing research interest in the research area as can be seen in Figure 2. Interest declined in 2018 but later

rose in 2019. In 2020, the interest continued to decline sequentially till 2023, when it slightly increased. In 2024, there was a decrease in the research area. In our research, we noticed that the least published articles were done in 2024 which can be assumed that the year has not ended and could accommodate more publications. The highest published articles were done in 2019, meaning, there was high interest in the research area.

IV. REVIEW OF FINDINGS AND DISCUSSION

From Table 9, 40 primary study papers were chosen for this systematic literature review, and the papers were used to answer the research questions in this study. All the internet-based databases used produced at least one paper each. All the papers answered the research questions listed in this study.

A. Review of Findings

To answer RQ1: What are the most commonly used feature selection techniques for high-dimensional data analysis?

We noticed that all the feature selection techniques used in all the selected papers improved the prediction accuracy. As is clearly shown in Figure 4, chi-square was most frequently used, followed by information gain, whereas fast correlation technique, relief, gain ratio, and Boruta having 4 number of studies each, analysis of variance (ANOVA) and minimum redundancy maximum relevance (MRMR) having 3 number of studies each while symmetrical uncertainty (SU), fisher filtering, gradient boosting and Recursive feature selection technique are having 2 number of studies each.

Table 8: Summary of the primary selected papers synthesis

	Ref.	Database /Year	Feature Selection Techniques	ML techniques	FS Impact on ML model performance	FS Application Domain	Publication Type
1	(Subbiah *et al*.)	IEEE 2022	Boruta	SVM, LDA, CART, random forest	The model is enhanced with a random forest classifier from 99% to 99.9%	Network Intrusion Detection	Conference
2	(Lv *et al*.)	Springer 2023	Reweight Boosting Feature Selection (RBFS)	XGBoost	This model improves the accuracy prediction performance of the proposed FS technique.	Healthcare	Journal
3	(Spencer *et al*.)	Google Scholar 2020	Chi2, ReliefF, PCA, Symmetrical uncertainty	BayesNet, Logistic, Stochastic Gradient Descent (SGD), KNN, Adaboost M1 with Decision Stump, Adaboost M1 with Logistic	The most accurate model is accuracy 85.0%, precision 84.73%, and recall 85.56%, using Chi-square feature selection with BayesNet classifier	Healthcare	Journal
4	(Noroozi *et al*.)	Springer 2023	Information gain, Symmetrical uncertainty	Bayes net, Naïve Bayes (BN), Multivariate Linear Model (MLM), Support Vector Machine (SVM), logit boost, j48, and Random Forest.	SVM-based filtering methods have a best-fit accuracy of 85.5	Healthcare	Journal
5	(Ay *et al*.)	Springer 2023	Meta-heuristic algorithms (Flower Pollination Algorithm (FPA), Whale Optimization Algorithm (WOA), and Harris Hawks Optimization (HHO) algorithms	SVM, KNN, Random Forest, Naïve Base, Logistic Regression	The impact was obtained from KNN with an F-score of 97.45%	Healthcare	Journal
6	(Khourdifi & Bahaj)	Springer 2019	Fast Correlation-Based Feature Selection (FCBF), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) approach.	KNN, SVM, RF, NB, MLP	The proposed optimized model by FCBF, PSO, and ACO achieved an accuracy score of 99.65% with KNN.	Healthcare	Journal

7	(Zaini & Awang)	Springer 2023	Chi-squared and analysis of variance (ANOVA)	Logistic regression, KNN, decision tree, random forest, Gaussian naive Bayes, extra gradient boosting, support vector classifier, multilayer perceptron, stochastic gradient descendent, and additional tree classifier	The performance of a stacking ensemble yields the best result with 93.44% using logistic regression.	Healthcare	Journal
8	(Semwal *et al*.)	Springer 2017	ANOVA, incremental feature selection	ANN, SVM, KNN, classifier fusion and DNN	classifier fusion provides satisfactory results (92.23 %) compared with other classifiers	Healthcare	Journal
9	(Oreski *et al*.)	Scopus 2017	Info gain, gain ratio, Relief, Linear forward selection, and Voting feature selection	Neural Networks, Discriminant Analysis, and Decision Trees	The DT model improves the accuracy	Data science	Journal
10	(TAKCI)	Scopus 2018	Backward logit, Forward logit, Fisher filtering, ReliefF	C4.5, C-RT, SVM, ID3, K-NN, MLP, Naïve Bayes and Logistic regression	The best model is support vector machine algorithm with the linear kernel with 84.81%	Healthcare	Journal
11	(Shafiq *et al*.)	Springer 2018	WMI_AUC weighted mutual information (WMI) metric and area under ROC curve (AUC) and Robust feature selection (RFS) algorithm	C4.5 decision tree, Random Forest, AdaBoost, Naïve Bayes, Bagging, R/tree, SMO, Bays Net, OneR, PART, and Hoeffding	Random Forest and C4.5 decision tree ML classifiers enhance the model accuracy by 100% and 99.99% respectively.	Network Intrusion Detection	Journal
12	(Aksu *et al*.)	Springer 2018	Fisher Score algorithm	Support Vector Machine (SVM), K Nearest Neighbour (KNN) and Decision Tree (DT)	SVM with an accuracy of 0.9997% enhances the model success rates	Network Intrusion Detection	Journal
13	(Masoudi-Sobhanzadeh *et al*.,)	Springer 2019	FeatureSelect	SVM, ANN, and DT	The SVM model was adopted to developFeatureSelect based on LIBSVM library its uses.	Researchers	Journal

14	(Zhao *et al*.,)	IEEE 2019	mRMR	Random Forests, Naïve Bayes, and Logistic regression	Their model-based variants RFCQ and RFRQ provide optimal results for Random Forests classification models with 0.811% each	e-commerce	Conference
15	Zhang)	Scopus 2019	Global Sensitivity Analysis (GSA)	Random forest	This novel FS method was ensembled with random forest (RF) to determine the optimum feature subsets.	Engineering	Journal
16	Nithya&Ilango)	Springer 2019	Boruta algorithm, SA and tree()	C5.0, part, Random Forest, SVM and KNN algorithms	C5.0 and Random Forest classifiers impact with an accuracy of 97% and 96.9% respectively.	Healthcare	Journal
17	(Trivedi)	Scopus 2020	Information-gain, Gain-Ratio and Chi-Square	Naïve Bayes, Bayesian, Random Forest, Decision Tree (C5.0) and SVM (Support Vector Machine	Random forest enhanced the best performance compared to other models with 93% accuracy.	Financial Institution	Journal
18	(Toğaçar *et al*.)	Scopus 2020	minimum Redundancy Maximum Relevance	Decision tree, K-nearest neighbors, Linear Discriminant Analysis, Linear Regression, and Support Vector Machine	linear discriminant analysis impacts the best results with an accuracy of 99.41%	Healthcare	Journal
19	(Ali *et al*.)	IEEE 2020	Correlation-based Feature Selection (CFS)	Multilayer Perceptron (MLP) and instance-Based Learning algorithm (IBL)	IBL enhanced the model accuracy with 99.87%	Network Intrusion Detection	Conference
20	(Uçar)	Scopus 2020	P-Score	DT, NB, SVM	The impact of SVM enhances the performance of P-Score FST for classification performance.	Healthcare and Text classification	Journal
21	(Upadhyay *et al*.)	IEEE 2021	Gradient Boosting Feature Selection, Chi-Square, and Principal Component Analysis (PCA)	DT	DT improves the accuracy with 97.66%	Network Intrusion Detection	Conference

22	(Mahindru & Sangal)	Springer 2021	Correlation-based feature selection, Rough set analysis (RSA) Chi-squared test, information gain, oneR	Least Square Support Vector Machine (LSSVM)	LSSVM improves detection accuracy of 98.75%	Network Intrusion Detection	Journal
23	(Srinivasa Rao *et al*.)	Science Direct 2020	Hybrid Binary Chemical Reaction Optimization-Particle Swarm Optimization (HBCRO-BPSO) Binary Chemical REACTION Optimization (BCRO) and	Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Random Forest (RF	KNN classifier produces an accuracy of 5.01% for APRF, 3.83% for APIA	Researchers	Journal
24	(Sivaranjani *et al*.)	IEEE 2021	Forward and Backward Feature Selection	Support Vector Machine (SVM) & Random Forest (RF)	The RF accuracy of 83% gives the highest enhanced prediction result	Healthcare	Conference
25	(Thirumoorthy & Muneeswaran)	Springer 2022	Term Frequency Distribution Measure (TFDM)	Naive Bayes and SVM	NB performed better in the WebKB dataset with 89.47% while also better in BBC News datasets with 96.7%	Text classification	Journal
26	(Mohy-eddine et al., 2023)	Scopus 2023	Principal component analysis (PCA), univariate statistical test, and genetic algorithm (GA	K-Nearest Neighbors (K-NN)	the model provided 99.99% ACC for prediction accuracy	IoT	Journal
27	(Sultana & Islam)	Springer 2023	RFE, Boruta, and LASSO	KNN, RF, DT, SVM, AB, and GB	The ML model improves the prediction results using an RF classifier, with 99% accuracy.	Healthcare	Journal
28	(Begum *et al*.)	IEEE 2015	Consistency Based Feature Selection (CBFS), Fuzzy Preference Based Rough Set (FPRS)	K Nearest Neighbour (kNN)	The KNN model enhances the performance of CBFS with 94.28%	Healthcare	Conference

			and Kernelized Fuzzy Rough Set (KFRS)				
29	(Ramya & Kumaresan)	IEEE 2015	Information Gain, Gain Ratio and Chi-Square	No classifier used	no ML model impact	Financial Institutions	Conference
30	(Tang *et al*.)	Science Direct 2016	ANOVA	SVM	the model could achieve an overall prediction accuracy of 83.6%	Bioinformatics	Journal
31	(Jing Sun *et al*.)	IEEE 2016	Minimum Redundancy-Maximum Relevance (mRMR), Max Relevance (MaxRel), Quadratic Programming Feature Selection (QPFS) and Partial Least Squared (PLS)	K-Nearest-Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN).	The KNN model improves the accuracy by 1.0%	Healthcare	Conference
32	(Emmanuel *et al*.)	Springer 2024	Information Gain (IG)	RF, GB, XGB, KNN, ANN, Stacked and DT	The stacked model enhances the best accuracy with 86.23%	Financial Institutions	Journal
33	(Pathak *et al*.,)	Science Direct 2024	Gradient boosting FI	KNN, NB, DT and RF	Random Forest (RF) classifier impacts the prediction accuracy with 91.39%.	Intrusion Detection System	Journal
34	SumaiyaThaseen&AswaniKumar,)	Science Direct 2017	Chi-square	SVM	the model improves the prediction accuracy of 98.0% with SVM	Intrusion Detection System	Journal
35	(Xu *et al*.)	Science Direct 2021	MFeature	K-Nearest Neighbor (KNN)	This model impacts the accuracy prediction of the feature technique	Researcher	Journal
36	(Alejandre *et al*.)	IEEE 2017	Genetic Algorithm (GA)	C4.5	The model improves the detection rate on ISOT with 99.46%, ISCX with 95.58% and ISOT + ISCX with 96.52%	Botnet Detection	Conference

37	(Tuba *et al*.)	Science Direct 2019	Brain Storm Optimization (BSO)	SVM	SVM enhanced the accuracy of the proposed BSO-SVM method on the Hepatitis dataset from 94.92% to 97.16% and on the liver dataset, 82.55% compared to 84.31% while on the diabetes dataset from 89.97% to 91.46%.	Healthcare	Journal
38	(Garcia-Chimeno *et al*.)	Springer 2017	L1 penalty	SVM, Boosting and Naive Bayes	SVM impacted the ML model with 95% compared with the other two models	Healthcare	Journal
39	(Arora & Kaur)	Scopus 2020	Bolasso(Bootstrap-Lasso), Gain ratio, chi-square, ReliefF	Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and K-Nearest Neighbors (K-NN)	BS-RF has higher impact classification accuracy that outperformed other methods in terms of Accuracy with 0.988% on the LC dataset, 0.823% on the Kaggle dataset, and 0.84% on the German dataset resulting in effectively improving the decision-making process of lenders	Financial Institution	Journal
40	(Naik & Mohan)	Springer 2019	Boruta	ANN	The ML model enhanced the decreasing error rate in the prediction by 12%.	Stock Exchange	Journal

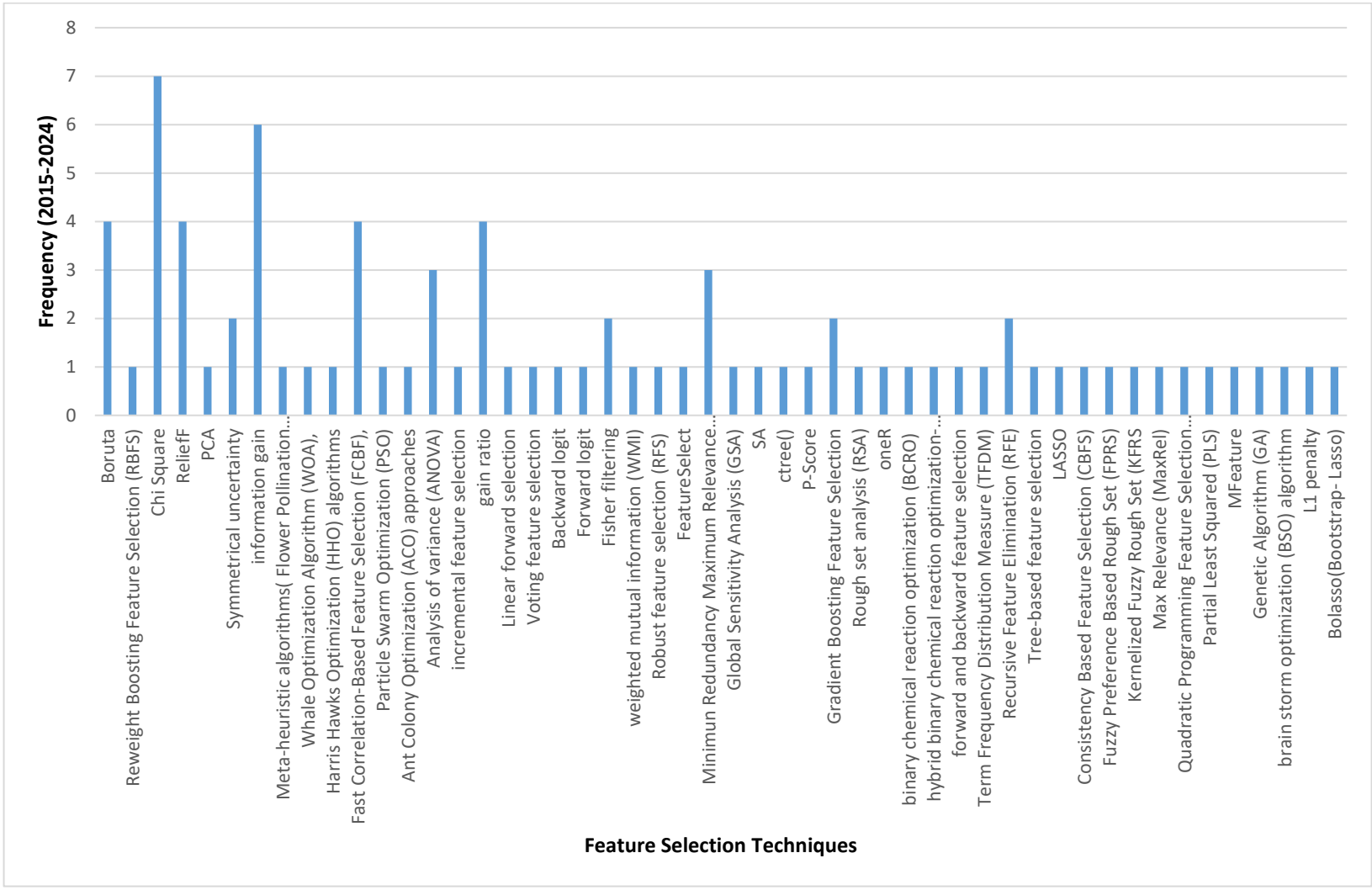

Figure 4: The distribution of feature selection techniques

Other feature selection techniques have only one number of studies over the selected primary studies used from 2015-2024. A total of 50 feature selection techniques used by the studies were recorded. The system literature review answered RQ1 by identifying the feature selection techniques employed by the primary selected studies from 2015-2024.

To answer **RQ2:** we need to know what machine learning is all about and how useful it is in prediction models. According to (Mahesh, 2020) machine learning is a scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Machine learning models help to provide systems with the ability to learn and enhance from experience automatically without being specifically programmed (Sarker, 2021).The feature selection techniques cannot effectively perform well without applying a machine learning model to its operations. As such, feature selection techniques impact the performance of machine learning algorithms by identifying the machine learning models employed by the researchers over the last decade and how the feature selection techniques were implemented for accuracy prediction using the machine learning models shown in Figure 5.

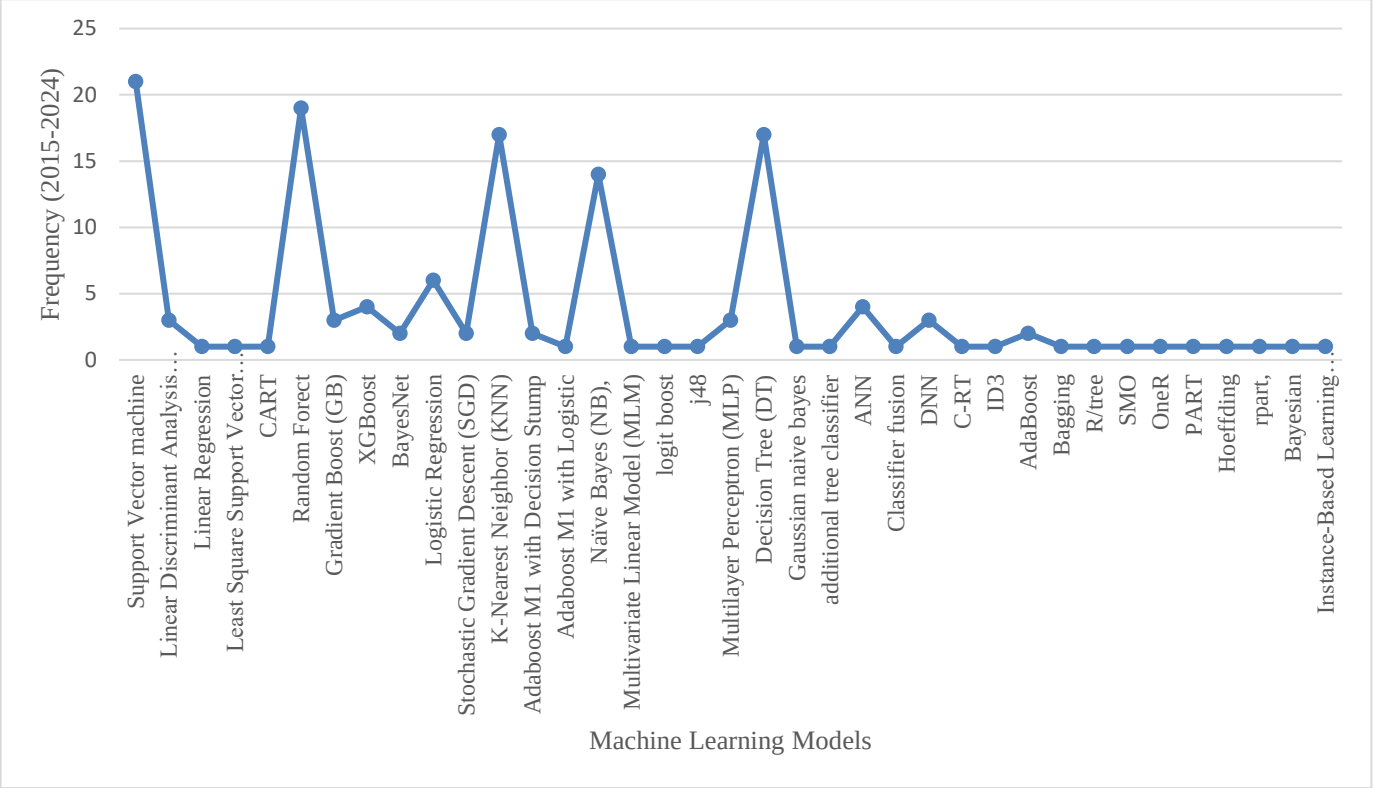

Figure 5: The distribution of machine learning models

As can be seen in Figure 5 above, 37 machine-learning algorithms were employed by the selected primary studies over the years. The most appropriately used machine learning model is the Support Vector Machine (SVM) with (21) studies, followed by Random Forest (RF) with (19) studies, next K-Nearest Neighbour (KNN) and Decision Tree (DT) with (17) studies each.Others can be seen as displayed in Figure 5. It can be concluded that feature selection techniques cannot be directly used to predict the accuracy of a model unless with the employment of any machine learning models.

To answer **RQ3**: most of the challenges noted from the literature reviews are summarized in Table 9.

Table 9. Summarized studies on the challenges and possible solutions

Study ID	Challenge identified	How they were addressed	References
S1	The technique does not take cognizant of the training time	The time of computation should be highly considered	(Subbiah et al., 2022)
S2	The computational time is high	Using another classifier may give better results	(Lv et al., 2023)
S3	Chi2 achieved better results than others used, but it did poorly in a large sample of data.	Novel or recently developed FS techniques can perform better	(Spencer et al., 2020)
S4	Only one dataset was used.	Multiple datasets should be used	(Noroozi *et al.*, 2023)

S5	Small sized two heart disease dataset was used	A large size dataset with high features should be used	(Ay, Eknci and Garip, 2023)
S6	it was tested on only one dataset	A variety of different datasets should be used	(Khourdifi & Bahaj, 2019)
S7	The feature size in the dataset is too small.	Large-sized datasets should be employed to test the efficacy of the techniques.	(Zaini & Awang, 2023)
S8	Other recognition systems such as biometrics were not tested	Other measures of the recognition system should be applied	(Semwal *et al*., 2017)
S9	Other existing techniques can do better than these techniques.	It should be applied to a variety of recently developed techniques.	(Oreski *et al.*, 2017)
S10	The dataset is a low dimensional size. It can only do well in a dataset with small features.	It should be applied in a large dimensional dataset	(TAKCI, 2018)
S11	Using only two datasets is not sufficient to ascertain the technique's performance.	Multiple datasets should be applied.	(Shafiq et al., 2018)
S12	The technique can detect only DDOS attacks	Other network attacks should be implemented	(Aksu et al., 2018)
S13	Having so many techniques embedded in one technique usually causes a computational delay.	One or two techniques joined perform better.	(Masoudi-Sobhanzadeh *et al*., 2019)
S14	The techniques do not consider the interaction between the features	Feature interaction should be looked into	(Zhao *et al*., 2019)
S15	Only one classier was used.	Trying Multiple classifiers would have gotten a better result	(Zhang, 2019)
S16	The techniques were not tested on other types of cancer disease other than cervical cancer prediction.	Other cancer disease datasets should be used.	(Nithya & Ilango, 2019)
S17	Only German credit data was used in their research	The research should be extended to other credit datasets	(Trivedi, 2020)
S18	Only the pneumonia dataset was used to implement the technique	Other related datasets should be employed	(Toğaçar et al., 2020)
S19	Applied only on one dataset	It should be extended to other datasets	(Ali et al., 2020)
S20	Few classifiers used	Many classifiers should be used	(Uçar, 2020)
S21	Several classifiers would have been employed. The techniques only use one classifier	Multiple classifiers should be employed	(Upadhyay et al., 2021)
S22	The study applied only one classifier. Also, was conducted can only detected whether an app is malware or benign without considering if each feature is capable of detecting malware or not.	More classifiers should be applied and the study should be extended to detect each feature should be able to detect malware.	(Mahindru & Sangal, 2021)
S23	The techniques were not experimented on high-dimensional datasets.	The hybrid approach should be implemented on high-dimensional dataset with an additional classifier.	(Srinivasa Rao et al., 2021)
S24	One dataset is not sufficient. Two classifiers are not enough	More datasets and classifiers should be employed	(Sivaranjani et al., 2021)
S25	Only two datasets were used and only two classifiers	It should be applied to other datasets from different domain areas with more classifiers	(Thirumoorthy & Muneeswaran, 2022)
S26	Only two datasets and three classifiers	It should be extended to other connected device datasets with more additional classifiers for better results.	(Fatima *et al*. 2022)
S27	Applying only primitive FS cannot give a convincing accuracy result. Also, only one dataset was used.	More recent FST should be used and many datasets.	(Sultana & Islam, 2023)
S28	The research was implemented on only one dataset and one classifier	More datasets and classifiers should be used	(Begum *et al.*, 2015)
S29	No classifier used	ML algorithms should be used to ascertain the effectiveness of the technique.	(Ramya & Kumaresan, 2015)

S30	One dataset, one Technique and one classifier cannot analyze a convincing performance of the techniques.	Multiple datasets, FST and classifiers should be experimented	(Tang et al., 2016)
S31	The computational complexity of the techniques was not considered	The computational complexity of the techniques should be taken into consideration for the technique's better performance.	(Sun *et al.*, 2016)
S32	The research was applied to only one Feature Selection Technique (FST)	More FST should be experimented on the same datasets with more recently developed techniques.	(Emmanuel *et al*., 2024)
S33	Other FS techniques should have been used for better accuracy prediction.	Several techniques should be employed for better performance evaluation.	(Pathak *et al*., 2024)
S34	Only one technique, dataset and classifier which cannot give a befitting result	Multiple techniques, datasets and classifiers	(Thaseen & Kumar, 2017)
S35	The technique was not tested on popular learning algorithms such as SVM, extreme Boost, GBoost.t.c	It should be evaluated with more learning algorithms	(Xu et al., 2021)
S36	Fewer datasets were used. With one technique	More datasets with large-sized features should be used. More techniques should be used too.	(Alejandre *et al*., 2017)
S37	The 3 healthcare datasets used are not large	Large datasets should be employed with more learning algorithms	(Tuba et al., 2019)
S38	The technique was not compared with existing techniques for evaluation purposes.	It should be compared with other existing techniques on the same datasets with the same classifiers.	(Garcia-Chimeno et al., 2017)
S39	Was only applied to two credit datasets with only a few learning algorithms.	It should be experimented with more learning algorithms on more related datasets.	(Arora & Kaur, 2020)
S40	The technique was not compared with other existing techniques. One learning algorithm is not sufficient.	More techniques and classifiers should be experimented with.	(Naik & Mohan, 2019)

It was observed that some employed feature selection techniques used in the literature did not take cognizance of the training time of the model and applying only one technique for prediction is challenging only because some other existing techniques might perform better predictive accuracy. It was noticed that some studies applied only one machine learning algorithm for prediction whereas, when several machine learning models are applied, the best model will be identified. Also, it was discovered that some feature selection techniques could only do well in a low-dimensional dataset rather than ina high-dimensional dataset. It was also observed that some studies experimented on only one dataset with minimal machine-learning models. All the challenges mentioned, can be solved by applying several feature selection techniques on multiple datasets with diverse machine learning algorithms to identify the perfect and better techniques, machine learning algorithms, and better datasets for accurate prediction. This is illustrated in Figure 6.

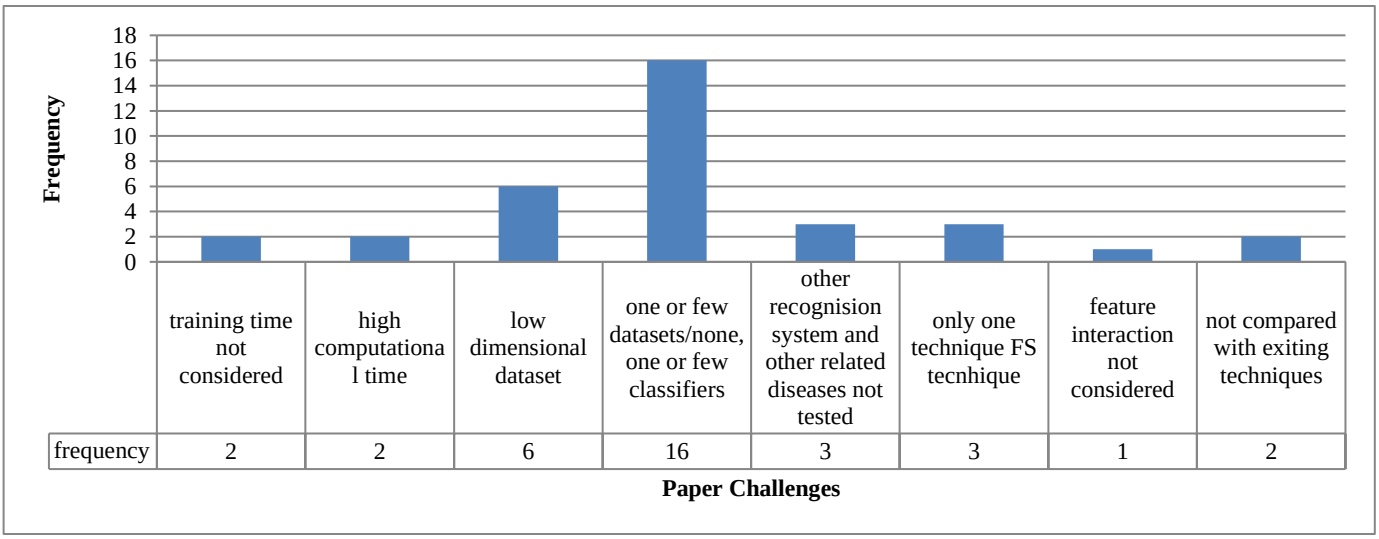

Figure 6: Distribution of studies challenges

The challenges noted in Table 9 are summarized in Figure 6; 16 papers used either only one or few datasets and none implementation of machine learning model to validate the feature selection techniques or uses one or few machine learning models to performed their experiments. This is a challenge because, for instance, when predicting heart disease, most of the heart disease datasets or even not all need to be experimented to ensure a better heart disease prediction. About 6 studies fall within those that experimented on a low-dimensional or small datasets with few features. This is a challenge because, applying those feature selection techniques or machine learning models on high–dimensional datasets may not give an optimal result. About 3 studies experiment fall within using either only one recognition system instead of trying other recognition system such as biometric or using only one disease instead of trying it on other related diseases. Another 3 studies applied only one feature selection techniques rather than applying more to selected the feature selection technique with the best result. About 2 studies fall within feature selection studies that do not take cognizance of the training time. Another 2 studies fall within high computational time and another 2 studies fall within feature selection techniques that were not compared with other existing techniques. Lastly, 1 study falls within feature selection techniques that do not take into consideration the interaction among the features in the datasets. It is suggested that, researchers should apply feature selection techniques on so many related classification problem datasets with many classifiers to identify the classifier that can give the best result.

Concerning **RQ4,** the application research areas are shown in Figure 7. About 12 application areas were identified and analysed as follows in accordance with the way they were explained in Table 8:

1. Healthcare: studies S3, S4, S5 S6, S7, and S10 used their feature selection techniques and classification models to predict heart disease. A variety of classification algorithms were used to create models that are then compared to seek the optimal feature combinations, to improve the correct prediction of heart conditions. Study S8 used their study to diagnose Parkinson's disease, and S16 used their techniques to analyse risk factors associated with cervical cancer that help in recognizing the importance of test features of cervical cancer by classifying the patients based on the result predicted. Study S18 used the techniques to detect pneumonia disease. S24 uses it for the prediction of diabetes disease. Study S27 used features selection techniques for thyroid disease classification and to identify the associated risk factors. Study S28 used it to predict Leukaemia issues which can be explained as a primary disorder of bone marrow. They are malignant neoplasms of hematopoietic stem cells; Study S37 used their techniques to predict Hepatitis, Liver disorder and Diabetes diseases. Their obtained results were compared to other state-of-the-art methods and their proposed approach achieved better classification accuracies for all three considered datasets while reducing or keeping the same number of features that are used. Study S38 used their feature selection technique to distinguish between patients with sporadic migraine, chronic migraine, and patients at risk of medication overuse is possible via feature selection techniques and machine learning. Study S31 experimented on how an improved performance microarray data analysis affect the prediction of colon and leukemia disease when the important features were selected.
2. Data mining: Study S9 used their technique to examine experimentally how dataset characteristics affect both the accuracy and the time complexity of feature selection. Their experiment was conducted on 128 datasets to examine their characteristics and how they affect the performance of feature selection techniques. The study revealed an intrinsic relationship between dataset characteristics and feature selection techniques' performance that will help researchers in data mining by providing a roadmap for future research on feature selection and a significantly wider framework for comparative analysis. Study S13 developed software called FeatureSelect that depends on machine learning models to address any feature selection problem in any kind of research. This software will select important features for the model to give an optimal accuracy prediction. Study S23 proposed a feature selection technique based on binary chemical reaction optimization and hybrid binary chemical reaction optimization-binary particle swarm optimization approach that can be used by data scientists to select relevant features in the dataset. Study S35 developed a model to select important features using an innovative binary version of moth-flame optimizer. This model helps to address some feature selection challenges that may result in not picking the optimal feature subsets.
3. E-commerce: Study S14 researched online product offerings and marketing strategies using feature selection techniques. The authors used the mRMR feature selection technique to evaluate both synthetic datasets and three real-world marketing datasets. The model-free FCQ (F-test correlation quotient) variant shows robust performance for different classification models as well as high computation efficiency as compared to other models experimented with.
4. Researchers: Study S20 developed a feature selection technique named P-Score that can select features based on their classification performance. The technique developed was based on decision support vector classification performance. This technique has been determined as a method which can be used in the field of machine learning. The study helped to describe the implementation and deployment of the mRMR method in a large-scale automated machine-learning platform for marketing purposes.
5. Engineering: Study S15 Proposed GSA-RF model that accurately captured tunnelling-induced settlement and determined the significance of input attributes to the settlement that provides a construction roadmap and predicts risks in engineering practice.
6. Financial Institution: Studies S17, S29, S32, and S39 experimented on credit scoring prediction by identifying defaulters of credit for an accuracy prediction purpose.

Their study will help financial institutions build an automated model for better credit scoring that decides potential candidates to give a line of credit identifying the right people without any credit risk.

7. Text Classification: Study S25 applied their technique on text classification to reduce the challenges confronting an unstructured text document such as high dimensional feature space, noise, and irrelevant information for better accuracy performance of text classifier.
8. Bioinformatics: Study S2 used their technique to predict the interaction of different species between long non-coding RNAs and protein by selecting the important feature in the dataset for accurate and optimal prediction. Study S30 predicted cell-penetrating peptides using their techniques. The model separates cell-penetrating peptides from non-cell penetrating peptides.
9. Stock Exchange: Study S40 considered the issue with short-term trading which often fluctuates demand and supply in the stock market by applying feature selection techniques on stock prices that predict technical indicators in the stock exchange.
10. Network Intrusion Detection: Studies S1, S11, S12, S19, S21, S22, S33, S34 performed experiments on wireless sensor networks (WSNs) attacks that are aimed to mitigate the network's ability to perform its expected functions. Due to the incredible development in computer-related applications and massive Internet usage, it is indispensable to provide host and network security. As such, the development of hacking technology tries to compromise computer security through intrusion. The studies employed an Intrusion detection system (IDS) with the help of machine learning (ML) Algorithms to detect intrusions in the network.
11. Botnet Detection: Study S36 used a feature selection technique to select important feature subsets from ISOT and ISCX datasets to identify the type of botnets using C4.5 machine learning model.
12. Internet of Things (IoT): Study S26 integrated an intrusion detection system using the K-Nearest Neighbors model to enhance IoT security. By applying feature selection and KNN model, their model had proven 99.99% accuracy in security enhancement.

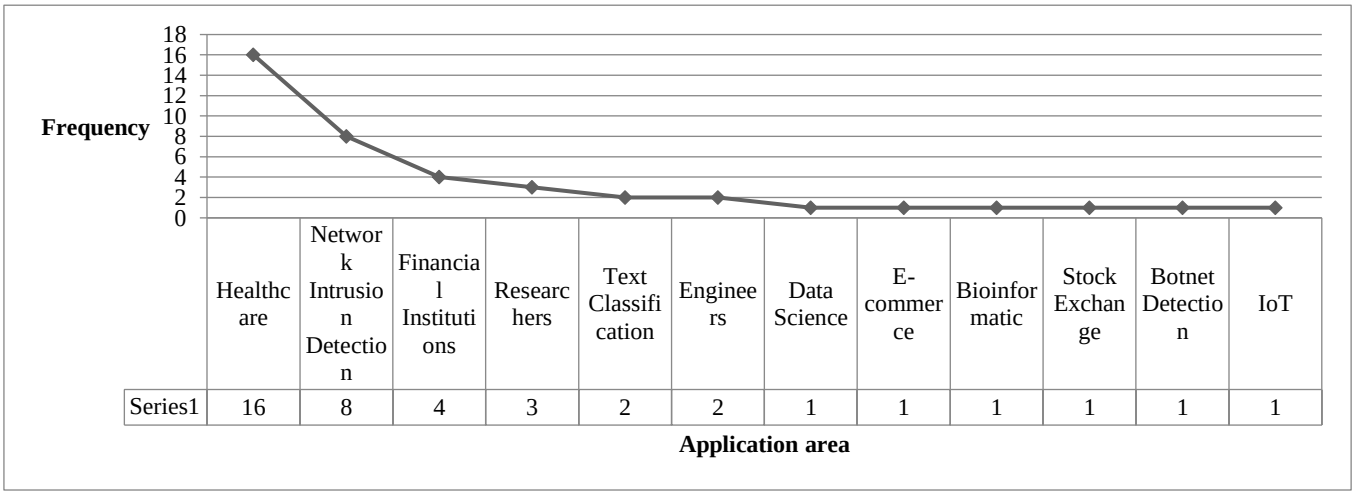

Figure 7: Distribution of application areas

Regarding **RQ5,** the computational time of the techniques and the machine learning models shall be put into our future focus. One of the major benefits of applying feature selection techniques in high dimensional data is to reduce the computational time of the model. The techniques help to remove the unimportant features of all the models and focus on the important features of the dataset which reduces the issue of overfitting. Removing the unimportant features increases the prediction accuracy and reduces the computational time of the information. Therefore, when dealing with dimensionality reduction, considering the kind of feature selection techniques and machine learning models to be adopted is very important. When developing feature selection techniques, attention should be given to how fast the model can predict the accuracy of the classification problems.

The results of recently proposed feature selection techniques (2020 till date) shall be compared to the results of years that precede 2020. Several feature selection techniques and machine learning models shall be experimented on multiple datasets of different application domains for better accuracy predictions. Finally, more research should be carried out to develop better feature selection techniques that can outperform existing techniques.

B. Discussion

RQ1: This systematic literature review identified the most commonly used feature selection techniques used for high dimensional datasets between 2015 and 2024. The study examined 50 feature selection techniques used by 40 primarily selected articles with 7 studies that employed the chi-square feature selection technique which is the most prominently used, followed by the information gain technique with 6 studies that implemented their research studies, 4 studies used fast correlation technique, reliefF, gain ratio, and Boruta. Analysis of variance and minimum redundancy maximum relevance (MRMR) have the same number of studies with 3 studies, while 2 studies use the symmetrical uncertainty, fisher

filtering, gradient boosting and Recursive feature selection technique. Other feature selection techniques have only one number of studies over the selected primary studies used from 2015-2024.

RQ2: The most frequently used machine learning models were identified in this systematic literature review. This study recorded 37 machine learning models used by the 40 primary selected studies over the last decade. The most appropriately employed machine learning model identified in this research study is the Support Vector Machine (SVM) with 21 studies, 19 studies that implemented the Random Forest (RF), next by 17 studies that adopted the K-Nearest Neighbour (KNN) and Decision Tree (DT) which can be seen as displayed in Figure 5. About 14 studies used Naïve Bayes, 6 studies applied logistic regression, 4 studies used XGboost and Artificial Neural Network (ANN) models, while 3 studies adopted the Linear Discriminant Analysis (LDA), Gradient Boost (GB) and Multilayer Perceptron (MLP). Here, 2 studies used BayesNet, Adaboost M1 with the decision, Stochastic Gradient Descent, Neural Networks and Adaboost while 1 study each used linear regression, Least square support vector, CART, Adaboost M1 with logistic, multivariate linear model, logit boost, j48, additional tree classifier, classifier fusion, DNN, C-RT, ID3, bagging, R/tree, SMO, OneR, PART, Hoeffding, rpart, Bayesian and instance-based learning (IBL).

RQ3: The feature selection challenges identified in this research article are categorized in 8 as can be seen in figure 6. Some techniques used in intrusion detection can detect only one type of attack in the network intrusion attack while other similar attacks cannot be detected. Some techniques do not consider the interaction between the features in the dataset, some use only one classifier whereas other classifiers can do better than that and a study does not use any machine learning model to predict the model accuracy. To solve some of these issues, when developing a feature selection technique or employing any technique for feature selection, the researcher should consider the computational time of the technique and also consider using several related datasets of the same domain area. Most of the suggested solutions have been expressed in Table 8.

RQ4: The 40 primary selected studies applied to solve a classification problem in 12 different application domains between 2015 and 2024. From all the experiments carried out by the 40 selected studies, we noticed that most of the studies focussed on healthcare care problems rather than other fields of research. Considering Figure 7, out of the 40 selected articles between 2015 and 2024, healthcare has about 17 studies, followed by network intrusion Detection with 9 studies. Financial institution takes 4 studies while 3 studies were related to research applications. About 2 studies conducted their studies on engineering and text classification application areas. The other application areas such as data science, e-commerce, bioinformatics, stock exchange, botnet detection, and IoT have 1 number of studies each as can be seen in Figure 7. From the analysis and observations noticed from this study, it was discovered that healthcare, network intrusion detection, and financial institutions need special attention in applying feature selection techniques to enhance their prediction capabilities.

RQ5: The emerging Trends and prospects in feature selection for high-dimensional data analysis as identified in recent research can be categorized as follows:

A. Application Domains

This study categorized these techniques into two major groups and 12 application domains, which helped highlight dominant techniques and domains. The healthcare domain takes the highest frequency with 17 studies to solve problems in the healthcare domain, which is the highest source of high dimensional data. This is followed by the network instruction detection domain with 8 studies that investigated the domain area, which is also a good source of high-dimensional data. Researchers covered 3 studies in the research domain. Financial institutions' detection of fraud attempts has increased drastically with about 4 studies done between 2015-2024, which makes fraud detection more important than ever. Text classification and engineers have 2 studies each, while data science, e-commerce, bioinformatics, stock exchange, botnet detection and IoT have 1 study each. More attention should be given to those domain areas with prolific sources of high-dimensional data.

B. Machine Learning Models

Regarding ML models used in this study, researchers focused on supervised classification methods such as SVM (21 papers). On the other hand, Random Forest appeared in 19 papers, KNN and Decision Tree were used 17 papers each, while Naïve Bayes and Logistic Regression, Xgboost, ANN, LINEAR regression, gradient boost, multilayer perceptron, Bayes net, Adaboost M1 with decision stump, stochastic gradient descent, Neural Network and Adaboost were used in 14, 6, 4, 4, 3, 3, 3, 2, 2, 2, 2, and 2 papers, respectively. Others, as seen in Figure 5, appeared in 1 paper each. All the studies applied supervised classification methods. In prospect, both supervised and unsupervised tasks as deep learning techniques should be experimented with. Additionally, clustering is beneficial for investigating latent space or uninterruptable patterns and future studies focusing on unsupervised practices can help uncover new insights.

V. CONCLUSION

This systematic literature review was conducted to identify the feature selection techniques for high-dimensional data analysis and machine learning models used for the period of 2015-2024 to help researchers take decision on the appropriate approach to use when carrying out research and to investigate the current issues and future direction of feature selection in high-dimensional data. Five popularly known research databases were accessed in this study in order to select relevant articles such as journals or conferences proceedings. Between 2015 and 2024, 40 studies were selected, and five research issues were examined and analyzed. The selected primary papers demonstrate how feature selection techniques combined with machine learning algorithms can increase prediction accuracy. 50 feature selection techniques were examined among which chi-square was identified as the most frequently used technique and 37 machine learning algorithms were applied to the 50 feature selection techniques for accuracy

prediction and support vector machine was identified as the most prominently used machine learning model. These feature selection techniques chosen from all the primary studies were applied in 12 different application areas and most of the techniques were applied on healthcare domain. This study could not separate the appropriate feature selection techniques and machine learning model for each of the application domain. In our future work, we will identify suitable feature selection technique and machine learning model for each of the 12 application domains.

AUTHOR CONTRIBUTIONS

Phillemon Uten Emmoh Resources, formal analysis, Data Curation, Conceptualization, Methodology, Software, Writing-Original draft preparation: **C.I. Eke:** Data curation, Writing-Original draft preparation, Supervision Project administration, Writing-Reviewing and Editing: **Timothy Moses:** Visualization, Supervision, Investigation, Software, Validation: **Agushaka J. Ovre:** Supervision, Investigation**.** All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENT(S)

I acknowledged everyone that contributed to this research work.

REFERENCES

Aksu, D.; S. Üstebay, M.A. Aydin & T. Atmaca. (2018). Intrusion Detection with Comparative Analysis of Supervised Learning Techniques and Fisher Score Feature Selection Algorithm. 24th IFIP World Computer Congress, WCC, Poznan, Poland, 141–149. https://doi.org/10.1007/978-3-030-00840-6_16

Al-shalif, S. A.; N. Senan, F. Saeed, W. Ghaban, N. Ibrahim, M. Aamir & W. Sharif. (2024). A systematic literature review on meta-heuristic based feature selection techniques for text classification. PeerJ Computer Science, 10, e2084. https://doi.org/10.7717/peerj-cs.2084

Alejandre, F.V., Cortes, N.C. & Anaya, E.A. (2017). Feature selection to detect botnets using machine learning algorithms. International Conference on Electronics, Communications and Computers (CONIELECOMP) (pp. 1–7), Cholula, Mexico. https://doi.org/10.1109/CONIELECOMP.2017.7891834.

Alhassan, A.M. & Wan Zainon, W.M.N. (2021). Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis., *IEEE Access, 9*, 87310–87317.
https://doi.org/10.1109/ACCESS.2021.3088613.

Ali, A.; S. Shaukat, M. Tayyab, M.A. Khan, J.S. Khan, Arshad, & J. Ahmad. (2020). Network Intrusion Detection Leveraging Machine Learning and Feature Selection. IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), 49–53, Charlotte, NC, USA. https://doi.org/10.1109/HONET50430.2020.9322813

Arora, N. & Kaur, P.D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing, 86*(1), 105936. https://doi.org/10.1016/j.asoc.2019.105936.

Ay, Ş.; E. Ekinci & Z. Garip. (2023). A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *The Journal of Supercomputing,79*(11), 11797–11826. https://doi.org/10.1007/s11227-023-05132-3

Bahmaninezhad, F. & Hansen, J. H. L. (2016). Generalized Discriminant Analysis (GDA) for Improved i-Vector Based Speaker Recognition. Interspeech 2016, 3643–3647. https://doi.org/10.21437/Interspeech.2016-1523

Begum, S.; D. Chakraborty & R. Sarkar. (2015). Data Classification Using Feature Selection and kNN Machine Learning Approach. 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 811–814. https://doi.org/10.1109/CICN.2015.165

Bouchlaghem, Y.; Y. Akhiat & S. Amjad. (2022). Feature Selection: A Review and Comparative Study. E3S Web of Conferences, 351, 01046. https://doi.org/10.1051/e3sconf/202235101046

Edlund, C.; T. R. Jackson, N. Khalid, N. Bevan, T. Dale, A. Dengel, S. Ahmed, J. Trygg & R. Sjögren. (2021). LIVECell—A large-scale dataset for label-free live cell segmentation. *Nature Methods, 18*(9), 1038–1045. https://doi.org/10.1038/s41592-021-01249-6

Eke, C.I., Norman, A. A., Shuib, L. & Nweke, H. F. (2019). A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. *IEEE Access, 7*(3), 144907–144924. https://doi.org/10.1109/ACCESS.2019.2944243.

Eke, C. I.; A. A. Norman & M. Mulenga. (2023). Machine learning approach for detecting and combating bring your own device (BYOD) security threats and attacks: a systematic mapping review. *Artificial Intelligence Review, 56*(8), 8815–8858. https://doi.org/10.1007/s10462-022-10382-3

Eke, C.I., Norman, A.A. & Shuib, L. (2021). Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach. PLOS ONE, 16(6), 0252918. https://doi.org/10.1371/journal.pone.0252918.

Emmanuel, I.; Y. Sun & Z. Wang. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data, 11*(2), 23. https://doi.org/10.1186/s40537-024-00882-0

Garcia-Chimeno, Y.; B. Garcia-Zapirain; M. Gomez-Beldarrain; B. Fernandez-Ruanova & J. C. Garcia-Monco. (2017). Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data. *BMC Healthcare Informatics and Decision Making, 17*(1), 38-40. https://doi.org/10.1186/s12911-017-0434-4

Hashemi, A.; M. Joodaki; N.Z. Joodaki and M.B. Dowlatshahi. (2022). Ant colony optimization equipped with an ensemble of heuristics through multi-criteria decision making: A case study in ensemble feature selection. *Applied Soft Computing, 124*, 109046. https://doi.org/10.1016/j.asoc.2022.109046

Higgins JPT & Green S. (2008). Cochrane Handbook for Systematic Reviews of Interventions (J. P. Higgins & S. Green (eds.)). Wiley. https://doi.org/10.1002/9780470712184

Hopf, K. & Reifenrath, S. (2021). Filter Methods for Feature Selection in Supervised Machine Learning Applications - Review and Benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2111.12140

Sun, J., Passi, K., & Jain, C. K. (2016). Improved microarray data analysis using feature selection methods with machine learning methods. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1527–1534, Shenzhen. https://doi.org/10.1109/BIBM.2016.7822748

Karimi, F.; M.B. Dowlatshahi & A. Hashemi. (2023). SemiACO: A semi-supervised feature selection based on ant colony optimization. *Expert Systems with Applications, 214*, 119130. https://doi.org/10.1016/j.eswa.2022.119130

Khan, F.; I. Tarimer; H.S. Alwageed; B.C. Karadağ; M. Fayaz; A.B. Abdusalomov & Y.I. Cho. (2022). Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms. *Electronics, 11*(21), 3518. https://doi.org/10.3390/electronics11213518

Khourdifi, Y. & Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Engineering and Systems, 12*(1), 242–252. https://doi.org/10.22266/ijies2019.0228.24

Kitchenham, B. & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology, 55*(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010

Kitchenham B, C. S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report EBSE-2007-01.

Lv, G.; Y. Xia; Z. Qi; Z. Zhao; L. Tang; C. Chen; S.Yang; Q. Wang & L. Gu. (2023). LncRNA–protein interaction prediction with reweighted feature selection. *BMC Bioinformatics, 24*(1), 410. https://doi.org/10.1186/s12859-023-05536-1

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research(IJSR), 9*(1), 381–386. https://doi.org/10.21275/ART20203995

Mahindru, A. & Sangal, A. L. (2021). FSDroid:- A feature selection technique to detect malware from Android using Machine Learning Techniques. *Multimedia Tools and Applications, 80*(9), 13271–13323. https://doi.org/10.1007/s11042-020-10367-w

Masoudi-Sobhanzadeh, Y.; H. Motieghader & A. Masoudi-Nejad. (2019). FeatureSelect: a software for feature selection based on machine learning approaches. *BMC Bioinformatics, 20*(1), 170. https://doi.org/10.1186/s12859-019-2754-0

Mohamed Shaffril, H. A.; S. F. Samsuddin & A. Abu Samah. (2021). The ABC of systematic literature review: the basic methodological guidance for beginners. *Quality & Quantity, 55*(4), 1319–1346. https://doi.org/10.1007/s11135-020-01059-6

Mohy-eddine, M.; A. Guezzaz, S. Benkirane and M. Azrour. (2023). An efficient network intrusion detection model for IoT security using K-NN classifier and feature selection. *Multimedia Tools and Applications, 82*(15), 23615–23633. https://doi.org/10.1007/s11042-023-14795-2

Muhammad, M. U.; R. Jiadong; N. S. Muhammad; M. Hussain & I. Muhammad. (2019). Principal Component Analysis of Categorized Polytomous Variable-Based Classification of Diabetes and Other Chronic Diseases. *International Journal of Environmental Research and Public Health, 16*(19), 3593. https://doi.org/10.3390/ijerph16193593

Naik, N., & Mohan, B. R. (2019). Optimal Feature Selection of Technical Indicator and Stock Prediction Using Machine Learning Technique. International Conference on Emerging Technologies in Computer Engineering 985(1) 261–268. https://doi.org/10.1007/978-981-13-8300-7_22

Narisetty, N. N. (2020). Bayesian model selection for high-dimensional data Handbook of Statistics 43(1). 207–248. https://doi.org/10.1016/bs.host.2019.08.001

Noroozi, Z.; A. Orooji & L. Erfannia. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports, 13*(1), 22588. https://doi.org/10.1038/s41598-023-49962-w

Oreski, D.; S. Oreski & B. Klicek. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing, 52*, 109–119. https://doi.org/10.1016/j.asoc.2016.12.023

Pathak, A.; U. Barman, and T. S. Kumar. (2024). Machine learning approach to detect android malware using feature-selection based on feature importance score. Journal of Engineering Research. https://doi.org/10.1016/j.jer.2024.04.008

Petersen, K.; R. Feldt, S. Mujtaba & M. Mattsson. (2008). Systematic Mapping Studies in Software Engineering. Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, 68-77, Bari, Italy.

Pudjihartono, N.; T. Fadason; A.W. Kempa-Liehr & J. M. O'Sullivan. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics, 2*. https://doi.org/10.3389/fbinf.2022.927312

Ramya, R. S. & Kumaresan, S. (2015). Analysis of feature selection techniques in credit risk assessment. 2015 International Conference on Advanced Computing and Communication Systems,1–6. https://doi.org/10.1109/ICACCS.2015.7324139

Rasitha, G. (2016). Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique. *Communications on Applied Electronics, 4*(1), 4–6. https://doi.org/10.5120/cae2016651990

Rostami, M.; K. Berahmand; E. Nasiri & S. Forouzandeh. (2021). Review of swarm intelligence-based feature selection methods. *Engineering Applications of Artificial Intelligence, 100*, 104210. https://doi.org/10.1016/j.engappai.2021.104210

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN*

Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Semwal, V. B.; J. Singha; P. K. Sharma; A. Chauhan & B. Behera. (2017). An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification. *Multimedia Tools and Applications, 76*(22), 24457–24475. https://doi.org/10.1007/s11042-016-4110-y

Shafiq, M.; X. Yu; A. K. Bashir; H. N. Chaudhry & D. Wang. (2018). A machine learning approach for feature selection traffic classification using security analysis. *The Journal of Supercomputing, 74*(10), 4867–4892. https://doi.org/10.1007/s11227-018-2263-3

Sivaranjani, S.; S. Ananya; J. Aravinth & R. Karthika. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 141–146. https://doi.org/10.1109/ICACCS51430.2021.9441935

Spencer, R., F. Thabtah, N. Abdelhamid & M. Thompson. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health, 6*, 205520762091477. https://doi.org/10.1177/2055207620914777

Srinivasa Rao, P. C.; A. J. Sravan Kumar, Q. Niyaz, P. Sidike & V. K. Devabhaktuni (2021). Binary chemical reaction optimization based feature selection techniques for machine learning classification problems. *Expert Systems with Applications, 167*, 114169. https://doi.org/10.1016/j.eswa.2020.114169

Tang, H.; Z.-D. Su; Wei, H.-H..W. Chen, & H. Lin. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochemical and Biophysical Research Communications, 477*(1), 150–154. https://doi.org/10.1016/j.bbrc.2016.06.035

Theng, D. & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems, 66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Thirumoorthy, K. & Muneeswaran, K. (2022). Feature Selection for Text Classification Using Machine Learning Approaches. *National Academy Science Letters, 45*(1), 51–56. https://doi.org/10.1007/s40009-021-01043-0

Toğaçar, M.; B. Ergen, Z. Cömert & F. Özyurt. (2020). A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models. *IRBM, 41*(4), 212–222. https://doi.org/10.1016/j.irbm.2019.10.006

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society, 63*, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Tuba, E.; I. Strumberger, T. Bezdan, N. Bacanin & M. Tuba (2019). Classification and Feature Selection Method for Healthcare Datasets by Brain Storm Optimization Algorithm and Support Vector Machine. *Procedia Computer Science, 162*, 307–315. https://doi.org/10.1016/j.procs.2019.11.289

Uçar, M. K. (2020). Classification Performance-Based Feature Selection Algorithm for Machine Learning: P-Score. *IRBM, 41*(4), 229–239. https://doi.org/10.1016/j.irbm.2020.01.006

Upadhyay, D.; J. Manero, M. Zaman and S. Sampalli. (2021). Gradient Boosting Feature Selection With Machine Learning Classifiers for Intrusion Detection on Power Grids. *IEEE Transactions on Network and Service Management, 18*(1), 1104–1116. https://doi.org/10.1109/TNSM.2020.3032618

Uten E.P., Eke, C.I. and Moses, T. (2025). A feature selection and scoring scheme for dimensionality reduction in a machine learning task. J. Nig. *Soc. Phys. Sci, 7*, 2273. https://doi.org/10.46481/jnsps.2025.2273.

Xu, Y.; H. Huang; A.A. Heidari; W. Gui; X. Ye; Y. Chen; H. Chen & Z. Pan. (2021). MFeature: Towards high performance evolutionary tools for feature selection. *Expert Systems with Applications,186*, 115655. https://doi.org/10.1016/j.eswa.2021.115655

Yang, L.; H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong & D. Shao. (2021). Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective. *Information and Software Technology, 130*, 106397. https://doi.org/10.1016/j.infsof.2020.106397

Zaini, N. A. M., & Awang, M. K. (2023). Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction. *International Journal of Advanced Computer Science and Applications, 14*(2). https://doi.org/10.14569/IJACSA.2023.0140220

Zhang, P. (2019). A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model. *Applied Soft Computing, 85*, 105859. https://doi.org/10.1016/j.asoc.2019.105859

Zhao, Z.; R. Anand & M. Wang. (2019). Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 442–452. https://doi.org/10.1109/DSAA.2019.00059