Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776566

STUDY OF FEATURE SELECTION ALGORITHMS IN BIG DATA USING MACHINE LEARNING

Jagadish Kalava1, Dr. Pramod Pandurang Jadhav2
Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University, Indore, (M.P.) - 452010

Abstract
In many facets of our daily lives, machine learning has been widely adopted and applied. However, as the big data era approaches, certain conventional machine learning techniques are unable to meet the demands of real-time processing for significant data quantities. Machine learning must redesign itself in reaction to massive data. In this article, we evaluate current studies that have used machine learning for big data processing. First, a review of big data is provided, and then an analysis of the new features of machine learning in relation to large data follows. Then, using machine learning methods, we suggest a workable reference framework for managing massive data. The pre-processing steps are explained in the following chapter. This research work is carried out based on three approaches, namely filter feature selection approach, hybrid approach and ensemble feature selection approach. The mentioned approaches are analyzed and the results obtained are presented.
Keywords: *Feature Selection Algorithms, Big Data, Machine Learning, Big Data Challenges, Machine Learning Algorithms*

1. Introduction

1.1. Title Description

In the subject of machine learning, feature selection techniques are essential. Feature selection is also known as variable selection; attribute selection, and variable subset selection. The process of choosing the most important features from an input dataset to create a model is known as feature reduction. It's crucial to distinguish feature selection from feature extraction (A.Gandomi and M. Haider, 2015). While feature selection includes and excludes the attributes already existing in the data without changing them, feature extraction is a dimensionality reduction technique that creates new combinations of attributes. Modern machine learning techniques are already under danger from big data due to the development of more effective algorithms for analyzing vast and ever-changing data sets. The ability to make instantaneous judgments based on such streaming data has the potential to unlock significant value through data monetization. The amount of data created globally is expected to exceed 180 zettabytes (or 180 trillion gigabytes) by the year 2025. When put in context with the 10 zettabytes of data created in 2015, this number suddenly seems astronomical. Deep learning, which eventually led to AI, emerged as a result of the availability of massive datasets.

1.2. Background of Issue/ Problem

Massive amounts of data require complex analysis that would take too long and are impractical to complete. Therefore, methods for data reduction such data cube aggregation, attribute subset

selection, dimensionality reduction, numerosity reduction, data discretization, and concept hierarchy generation are employed. Any data reduction methodology used by the user must reduce the dataset without losing any important data. The accuracy of the classifier is another issue with classification. The feature selection process and the classification algorithm work together to determine how accurate the classifier is (Boston, 2018). The complexity may increase if incorrect and irrelevant features are chosen. The primary factor increasing classification efficiency is the feature selection approach. Many feature selection algorithms have failed to carry over the high dimensionality issue. Hence, these algorithms need an enhancement to reduce dimensionality and to improve accuracy in both ML Algorithms and predictions.

1.3. Basic Concept

The study of machine learning aims to comprehend the computational principles by which experience can improve performance by enabling computers to learn without being explicitly programmed. It is a very interdisciplinary field that draws inspiration from numerous various types of fields. Machine learning has nearly completely taken over every aspect of our lives in recent years; it is now so prevalent that you probably use it hundreds of times every day without even realising it. Through the adoption of a wide range of applications, which has had a significant impact on society and research, it is largely influencing the larger world. Many machine learning techniques, including neural networks, decision trees, support vector machines, k-nearest-neighbor, genetic algorithms, Q-learning, etc., have been presented in recent years. They have been applied in a variety of fields, including robotics, natural language processing, pattern recognition, and autonomous control systems.

Machine learning is a form of efficient mathematics that uses statistical algorithms to examine a variety of large-scale data sets. The accumulation of data sets is so vast and intricate, nevertheless, that it is challenging to manage using conventional data processing methods and models as the time for big data approaches. As a result, certain classic machine learning algorithms are inappropriate for this situation and unable to meet the demands of big data's real-time processing and storage. Therefore, in order to evaluate and manage massive data, we need to investigate some novel techniques that make use of parallel computing and distributed storage. In earlier study, academics mostly concentrated on two characteristics of studies: I One task involved creating a platform or framework for distributed parallel computing that could handle huge data quickly, such as Map Reduce, Dryad], Graph Lab, Hardtop, Haloop, and Twister, etc.; ii) Task three involved proposing a new class of algorithms to address specific big data issues. He Q et al. used a parallel extreme learning machine, for instance, to solve regression issues using Map Reduce. To deal with the imperfect streaming huge data, the authors have developed a low-complexity subspace learning method. For the sparse representation of large amounts of data, several academics have also used dictionary learning.

Constantly and at an increasing rate, data is being created. A lot of data is produced by modern technology, including mobile phones, social media, medical imaging, smart gadgets, etc. They must be kept in storage and put to use for analysis (C. Rudin, 2014). The advent of digital devices and sensors automatically generates data, necessitating the need for real-time

processing to be monitored and stored. It is tough to simply store the enormous amount of data, and it is far more difficult to evaluate it. To find relevant patterns and meaningful information, the data is not saved using the standard structure and semantics. These problems create a push for modernization in the fields of science, education, business, and government, among others.

According to McKinsey, in order to gain insights from big data, it is necessary to use new technical infrastructures and analytical tools because big data has problems with scale, dispersion, and timeliness. It makes it abundantly evident that a company requires new data structures and tools, new analytical techniques, and the integration of numerous abilities into the new position of the data scientist. Big Data's properties prevent competent analysis using simply conventional databases or techniques. These cutting-edge tools and technologies enable the creation, exploitation, and management of massive datasets and the storage systems that address big data concerns, and are required to store, manage, and realize the commercial benefits. The biggest sources of big data are in the government, health, and social media industries.

The greatest sources are nontraditional which derive benefits of large and complex data analysis leads a way to create a new powerful analytical capabilities. Hence, the main objective of a big data system is to extract insights from huge volumes of dissimilar data which is not promising using traditional methods.

2. Literature review

2.1. Work in the relevant field

K.Kalpana et al (2019) since we live in the information era, we gather a massive amount of data from a wide variety of sources, with sizes ranging from petabytes to exabytes. Data is a resource because it conceals important information and knowledge. Better decision making can be achieved through the use of data analytics to gain in-depth understanding and spot subtle patterns. Data analytics is used to gain insights from data, and machine learning is at its core. Data that has more tuples and features introduces new difficulties for machine learning systems. Feature selection is sometimes carried out as a first step in the preprocessing of data before actual machine learning is carried out in order to increase the efficacy and precision of a machine learning algorithm. This serves to reduce the data's dimensionality by eliminating superfluous features. In this research, we investigate a number of different feature selection algorithms and evaluate their suitability for use in sentiment analysis of large datasets.

Khawla Tadist, et al (2019) Feature-selection approaches are expected to revolutionize the way genetic data is collected, analyzed, and used in the modern era by drastically simplifying the data and making it more amenable to further processing (Chun-WeiTsai, 2020). In the following decade, scientists will likely move toward examining the genomes of all known forms of life, making genomics the primary source of information. It is anticipated that feature selection methods will prove to be a game-changer by greatly simplifying genetic data, making it simpler to interpret and translate into actionable information. Without first doing a comprehensive study of the field, researchers have no way of knowing how their work compares to previous studies or what it can add to the field as a whole. We present a thorough

and organized review of the literature on feature-selection tactics for large-scale genomic data analytics in this paper.

Jianyu Miao (2016) Feature selection is a method for reducing the number of features that need to be considered by eliminating superfluous, redundant, or noisy ones. Improvements in learning performance, such as increased learning accuracy, decreased computational cost, and enhanced interpretability of models, are often the result of careful feature selection. Many different feature selection methods have been presented and demonstrated to be useful by researchers in computer vision, text mining, and related fields in recent years. The goal of this study is to give a thorough review of how these methods are currently being used. Additionally, a thorough experiment is run to determine whether or not feature selection, taking into account some of the strategies stated in the literature, can enhance the performance of learning. The experimental outcomes demonstrate the advantages of unsupervised feature selection techniques for machine learning applications, particularly for enhancing the efficiency of clustering.

2.2. Common methodology / experimental setup/ materials, in others work

Mohamad, M et al (2021) In order to deal with large and difficult datasets, this research proposes a novel strategy to data extraction by integrating the correlation-based feature selection (CFS), the best first search (BFS), and the dominance-based rough set approach (DRSA) techniques. By deleting irrelevant and inconsistent data, this study hopes to increase the classifier's functionality in decision analysis. The proposed method, CFS-DRSA, consists of a number of successive phases, the first two of which involve two important feature extraction tasks. Data reduction is the first step, which employs a CFS strategy and a BFS algorithm. The final, highly refined data set is then created using a DRSA during a data selection process. For that reason, the primary objectives of this research are to improve classification accuracy and reduce the computing time complexity. The experimental methodology used several datasets with varying properties and volumes to assess the reliability of the suggested strategy.

Kumar (2014)this work aims to increase the accuracy of classifiers by discussing various feature selection methods and how they have been used to diverse datasets for the purpose of selecting the necessary features for binary and multi class classification. Recent studies in medical diagnosis employ several classification algorithms to aid in the prognosis of illness. The classification method used to forecast the disease returns a binary class. Keeping the multiclass data in its unaltered form without excessive reduction allows the technique to be applied to the dataset for optimal precision. The optimal feature selection algorithm must be used to reduce the space and time complexity of a dataset with a large number of attributes, such as thousands. How accurately a classification system predicts labels for a given dataset can be used to gauge its effectiveness. By selecting the appropriate features from the original dataset, the accuracy bound is largely decided. The application of feature selection strategies can considerably enhance classification performance (D. Che, 2019). "Feature selection" refers to one of the initial stages of categorization. The efficiency of several feature selection algorithms is compared in this study using a range of datasets. By examining different feature

selection techniques and how they have been applied to diverse datasets for the purpose of picking the essential features for binary and multi class classification, this article tries to improve the accuracy of classifiers. Several classification algorithms have been used recently in medical diagnosis research to help in illness prognosis. The disease is classified using a binary class that is returned by the forecasting technique. Only if the multiclass data is preserved in its original form without maximum reduction can the method be used to the dataset to provide the maximum accuracy. To reduce the size and processing time of a dataset with millions of attributes, the best feature selection technique should be applied to cherry-pick the most crucial ones.

2.3. Tool used in past to solve similar problems and their results

Mohamad, M et al (2021) Standard evaluation measures and comparisons to well-established approaches like deep learning were used to verify the method's efficacy (DL). With the NN classifier achieving an accuracy rate of 82.1% compared to the SVM's 66.5% and the DL's 49.96%, the proposed study showed overall that it may help the classifier provide a substantial outcome. The proposed method can be used as an alternative extraction tool by people who don't have access to expensive large data analysis tools or who are otherwise uninitiated in the field of data analysis, according to the statistical results of a one-way analysis of variance (ANOVA).

Kumar (2014)to make use of a classification algorithm for prediction, a multiclass dataset must first be reduced to a binary class using one of several available data reduction techniques. When data reduction is used on the original dataset, sometimes the quality of the data and the algorithm's output can suffer. Measures of a classification algorithm's efficacy include their ability to correctly predict instances of a given class within a given dataset. The accuracy bound is mostly determined by picking the right features from the original dataset. Classification performance can be improved greatly with the use of feature selection techniques. Selecting features to use in the classification is a preprocessing step. This study compares the effectiveness of various feature selection algorithms across a variety of datasets.

2.4. Research gap

In today's era of "big data," there are usually many things to choose from in the troves of information at your disposal. Each object is described by tens of thousands, if not hundreds of thousands, of properties. It also gets difficult to analyse and use the information from such data. In various industries, including social media, e-commerce, healthcare, bio informatics, transportation, computer vision, text mining, etc., it has grown to be a significant issue. Dimensionality reduction aids data mining and machine learning algorithms in concentrating on the most crucial elements of the data (D. Yu, 2019). By shortening the learning time and boosting the learning performance, the decreased data aids to provide substantial benefits for decision making.

2.5. Problem Statement

There are significant difficulties with big data analytics. High cognitive analytical skills are required for the several types of analyses that must be performed on the vast quantities and types of data. The main focus of this study is on developing better feature selection algorithms to alleviate the curse of high dimensionality and increase precision while working with massive data. Feature selection is a crucial method for reducing the "curse of high dimensionality" by choosing the most informative features. There have been several attempts to solve the high dimensionality problem with feature selection algorithms, but they have all failed. Thus, these algorithms require a refinement to lessen the impact of dimensionality and raise the bar for the precision with which ML Algorithms and predictions can be made. Improved precision thanks to machine learning techniques.

2.6. Objectives

The main goal of this research is to increase the machine learning's accuracy. Algorithms while reducing the high dimensionality issues of big data by using the following approaches with objectives.

3. Research Methodology

It explains the research objectives and projected contributions. The methodology diagram mentioned in this chapter gives the picture of the overall process of the research work (Grünauer A, 2018). This chapter also describes the implementation setup, software applications and dataset used for the research work.

3.1. Details of experimental setup and material/ instrument used

- **Feature Selection Approaches**

Although feature selection has proven beneficial in numerous applications, big data's properties pose difficulties. It has been demonstrated that feature selection, a form of dimension reduction strategy, is effective and efficient in managing high dimensional data. Three distinct strategies are suggested and put into practice to lessen the high dimension curse. These approaches are performed and the feature selection occurs to reduce the dimension in order to improve the ML Algorithm accuracy.

- **Enhanced Filter Feature Selection Approach**

Modelling Machine Learning Algorithm with high dimension data remains a challengeable task. The filter feature selection method is used to strengthen and increase the precision of machine learning algorithms. With fewer rounds and a greater reduction in unnecessary and redundant features, the suggested strategy has been used. The most useful options are to rank and weight the features before updating depending on prediction criteria and iterations. ML algorithms have been used to construct a classifier model using the chosen feature subset.

- **Hybrid Feature Selection Approach**

To enable hybrid feature selection model, the proposed feature selection algorithm are applied on big data. To be able to achieve this, Map Reduce algorithms are developed to split the data set. Furthermore, hybrid feature selection approach applied then combines the solution which gives the output of interesting feature sub set from the big data set (J. Ekanayake, 2010). ML algorithms have been used to construct a classifier model using the chosen feature subset. With relation to the phase of Map and Reduce, execution time is also taken into account.

- **Ensemble Feature Selection Approach**

The suggested approach will be used with huge data to enable ensemble feature selection models. The linked works have amply demonstrated that the ensembles in the heterogeneous method do not have enough algorithms created for them. In this study, a unique rank and score aggregation method is presented together with a multi-filter ensemble feature selection methodology. This method reduces the majority of superfluous features. ML algorithms have been used to construct a classifier model using the chosen feature subset. To achieve more precision, the threshold values are selected.

3.2. Simulation and Software Details

Hadoop, a tool for building the map-reduce architecture, and Java, a programming language, are employed in this study to actualize the hybrid strategy. Hadoop's distributed file system, HDFS, takes its cues for its user interface from the Unix/Linux file system. HDFS primarily compartmentalises metadata and application data. NameNode is a specialized server that houses the HDFS metadata. DataNode is the server where all of the application's data are kept. The technology that actually processes data is called "Map Reduce." With Map Reduce, multiple jobs run in parallel; each job is effectively a self-contained Java programme that releases into the data stream and begins pulling information as it is required. Map Reducer's ability to process data in parallel is an essential part of the Hadoop framework (J.L. Liang, 2014). It's useful for analyzing massive amounts of data on a network of computers. The Anaconda distribution includes a desktop GUI called Anaconda Navigator that may be used to run apps and manage conda packages, environments, and channels without resorting to the command line. The navigator's backend is a Jupyter Notebook. There is a growing trend toward using Jupyter Notebooks, which are files that can be viewed, updated, and used in a web browser and contain all of the components of a traditional notebook, including code, descriptive text, output, graphics, and interactive interfaces. The tools utilized are Anaconda 1.9.12 and Jupyter Notebook 6.0.1. Python's interpretive nature, along with its dynamic type and clean syntax, make it a fantastic language for quick application development across many domains and platforms.

3.3. Details of new amendment in material/ design/ or mathematical correlation/ Chemical formula/ parts or product design as per new research requirement

The average Classification accuracy, Sensitivity and Specificity are used to find the impact of the proposed approaches. The metrics are discussed in the following sub-sections.

- **Confusion Matrix**

A 2x2 confusion matrix can be used to assess a binary classification task by comparing the expected labels to the actual labels. The measurements are computed for each class in a multi-class problem before the average is taken into account (Miao, 2016). The best way to determine the sensitivity, specificity, and accuracy is through the confusion matrix. There are four possible results for each of the confusion matrix's predictions: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The two by two confusion matrix is shown in table 3.1.

1. TP, or True Positive (Correctly predicted to be a positive class)
2. True Negative (Correctly predicted to not be a positive class)
3. Misleading Positive (FP) (Incorrectly predicted to be a positive class)
4. Four False Negatives (Incorrectly predicted to not be a positive class)

Table: 1. Confusion Matrix

		Predicted Positives	**Predicted Negatives**
Predicted	**True**	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

- **Sensitivity**

Sensitivity measures the proportion of actual positives that are correctly identified. The equation for calculating the sensitivity is given in the Equation 3.1.

$$\textbf{sensitivity} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

- **Specificity**

The true negative rate, often known as the ratio of true negatives to all other possible negatives, is a measure of specificity. The equation for calculating the specificity is given in the Equation 3.2.

$$\textbf{specificity} = \frac{\textbf{TN}}{\textbf{FP} + \textbf{TN}}$$

- **Accuracy**

The closeness of measuring results to the actual value is known as accuracy. The formula for calculating the accuracy is given in the Equation 3.3

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

- **ROC Curve**

Receiver operating characteristic (ROC) curves are used to compare categorization models visually. For a variety of diagnostic test results, the sensitivity vs. false positive rate is plotted on the ROC curve. The accuracy of the model is determined by the area under the ROC curve. One parameter, the Area under the Curve, can be used to represent all conceivable pairings of sensitivity and specificity that can be obtained by altering the test cutoff value (AUC).

The test's accuracy increases with larger AUC.

The test is 100 percent accurate if the AUC is 1.0.

The ROC curve is a straight diagonal line with an AUC of 0.5 (50%) and serves as the "perfect poor test," which can only ever be accurate by chance.

3.4. Implementation of methodology/ experimental setup establishment

In this research work the Bench Mark dataset is used (al, 2019). The data set is described in the forthcoming chapters. The Figure 1 depicts the methodological diagram of the proposed research work.

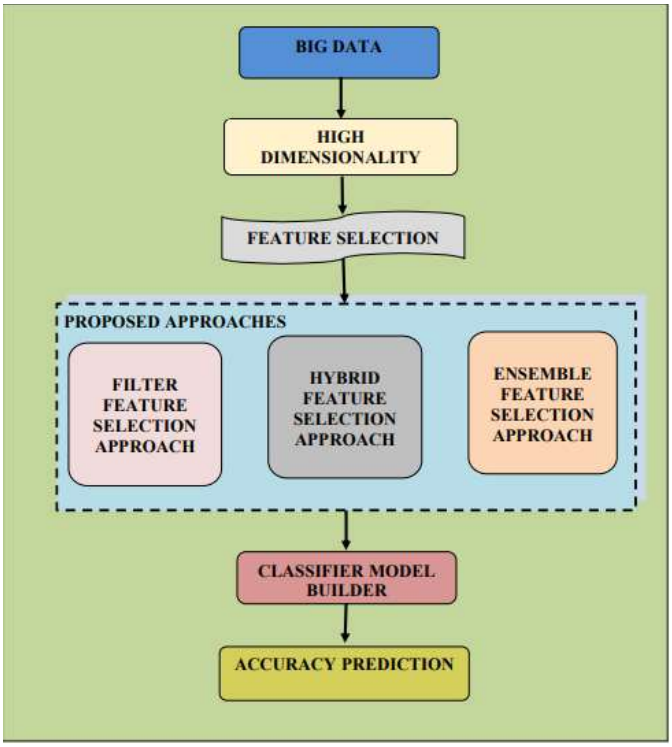


Figure: 1. Methodological Diagram

Data is collected from open source repository and simple missing values replacement technique has been done as pre-processing. Three different approaches are proposed to select the relevant features beside dimension reduction and to improve ML algorithms accuracy. These three approaches are explained in the succeeding chapters (al K. T., 2019). Sensitivity, Specificity, Accuracy are considered as performance measures to evaluate the improvement in the algorithms. The evaluation measurements are discussed in the following sections.

- **Data Set**

More over ten different data sets with high dimensions, different classes and instances are downloaded from the Open Source Repositories like UC Irvine (UCI) Machine Leaning Repository and they are considered for the proposed work.

4. Result Analysis

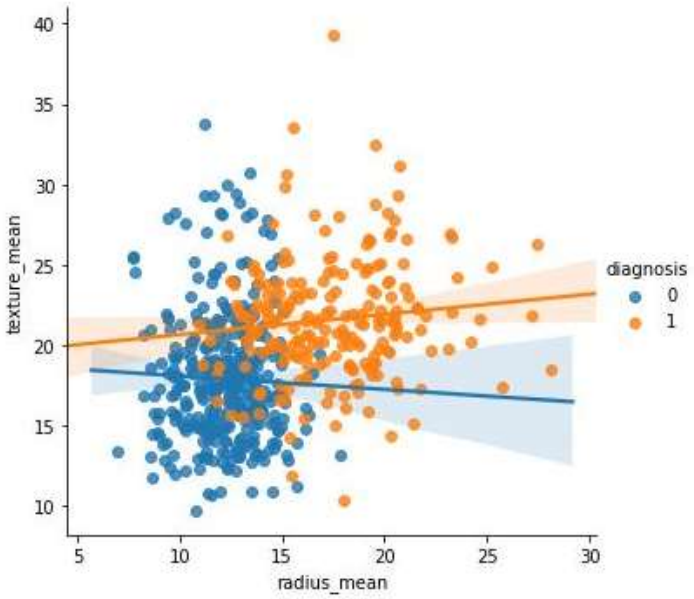
The experimental findings from the suggested approach are covered in this section. On an Intel (R) Core i7 processor, Python is used for all calculations. Open Source Repositories are used to collect five benchmark datasets (OSR). The specifics of the data set are shown in Table 4.1. The input data sets' sparsely In Figure 4.2, scatter plots for prostate cancer, weather forecast, diabetes, heart disease, and sonar are displayed.

4.1. Data generated through various experimentation

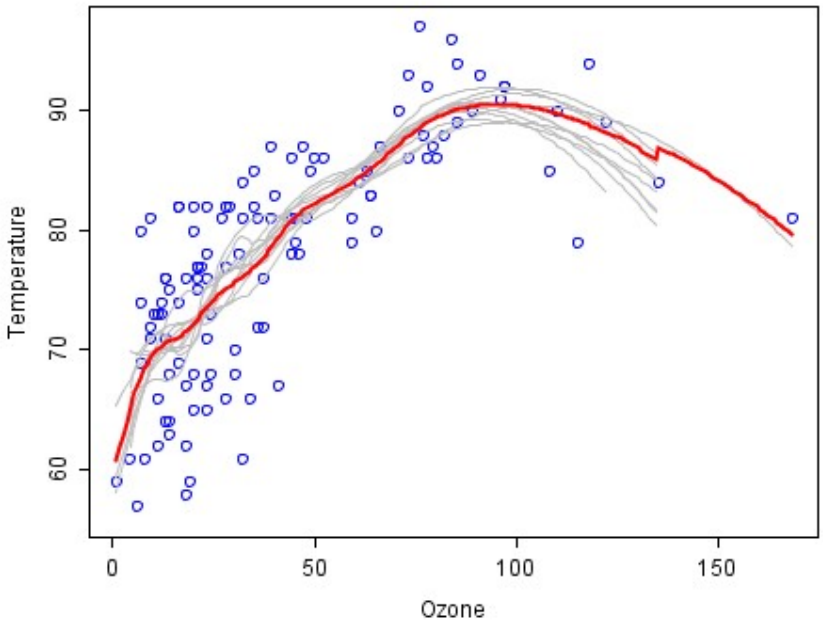
Table 4.1 Data Set Description

Data Set	Instances	Features	Class
Prostate cancer	100	9	2
Weather forecast	1000	16	4
Diabetic	65536	50	4
Heart disease	304	14	2
Sonar	625	60	2

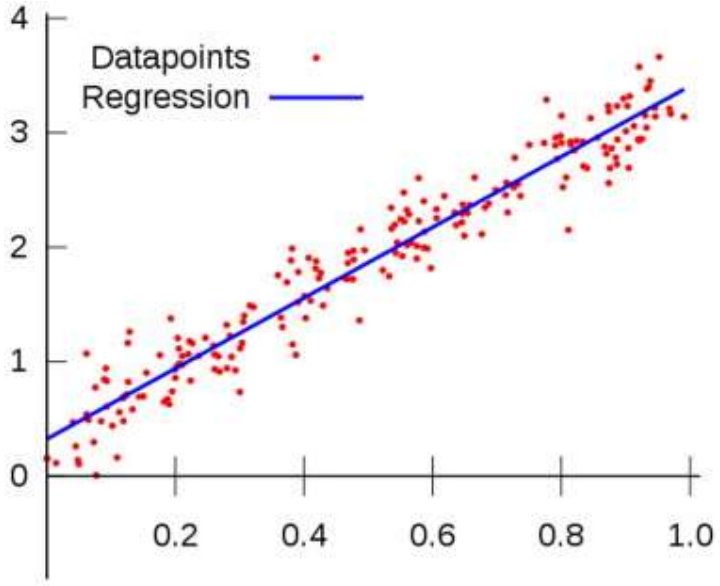
4.2. Data Representation through Graphs



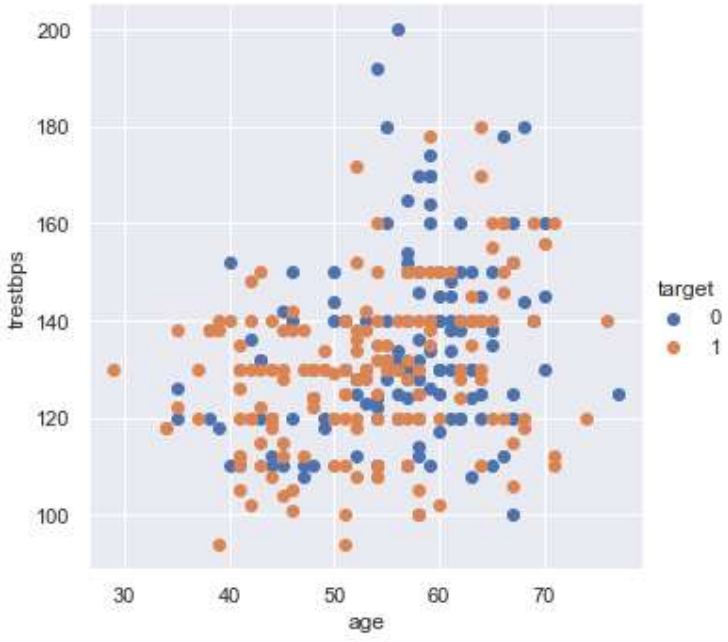
Prostate cancer



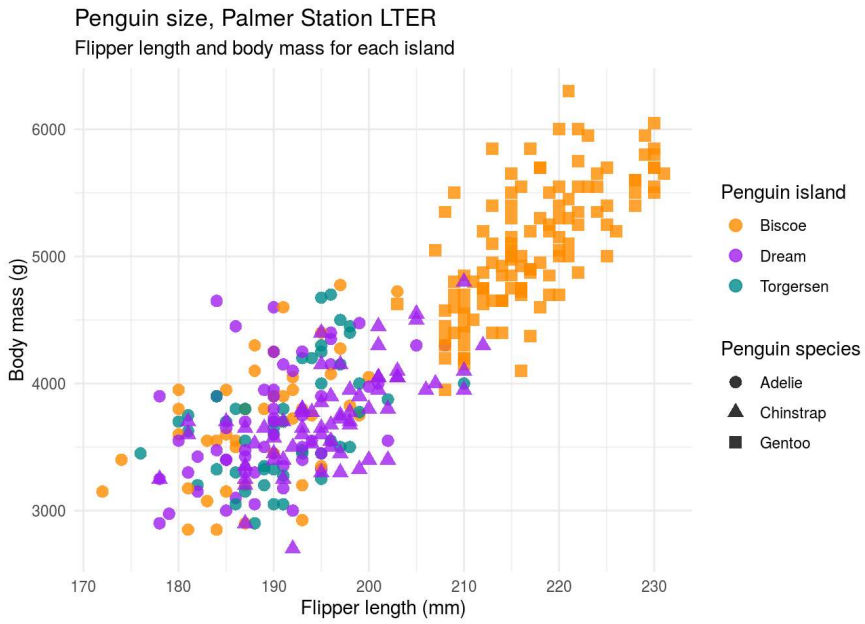
Weather for caste



Diabetic



Heart Disease



Sonar

Figure 2 Scatter Plot of Input Data Sets Dimension

4.3. Comparison charts and their analysis

Table 4.3 Comparison Result of Existing and Proposed Feature Selection

Data Set	Avg of Feature Selection Existing (Relief)	Avg of Feature Selection Proposed (E-Relief)
Prostate cancer	66.87	33.33
Weather forecast	78.40	60.00
Diabetic	65.00	50.00
Heart disease	76.82	65.23
Sonar	80.00	57.22

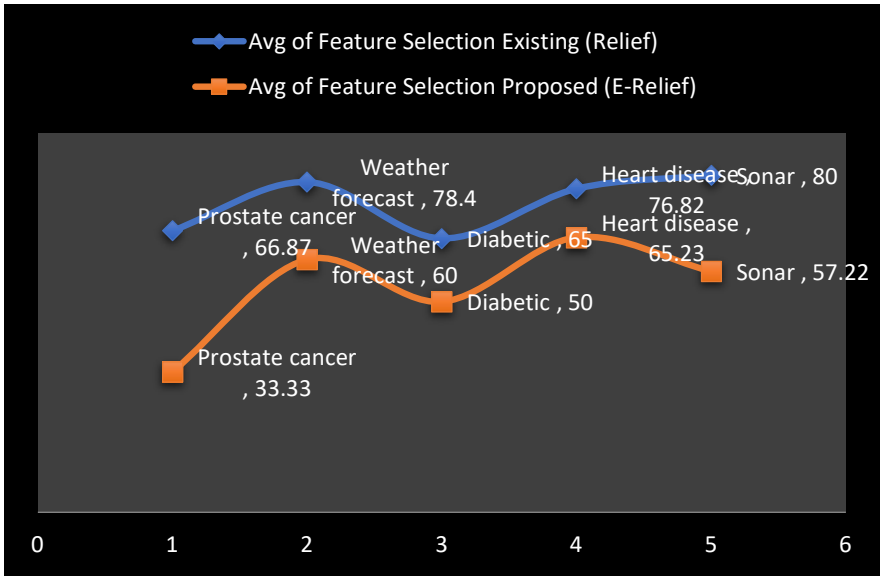


Figure: 3. Comparison Result of Existing and Proposed Feature Selection

Table 4.4 Comparative Results of ML Algorithms

Data Set	DT			KNN			NB			SVM		
	Accuracy	Sensitivity	specificity	Accuracy	Sensitivity	specificity	Accuracy	Sensitivity	specificity	Accuracy	Sensitivity	specificity
Prostate cancer	70	83	80	82	76	200	53	64	57	72	63	82
Weather forecast	82	73	200	66	40	74.4	75	64	200	87	34	83.6
Diabetic	67	63.7	82	63	53	66	73	62.3	77	72	70	72
Heart disease	67	55	40	66.3	55	82	53.4	50	53.4	70	70	72.34
Sonar	72	64	82	80	74	200	53	84	40	73	60	84

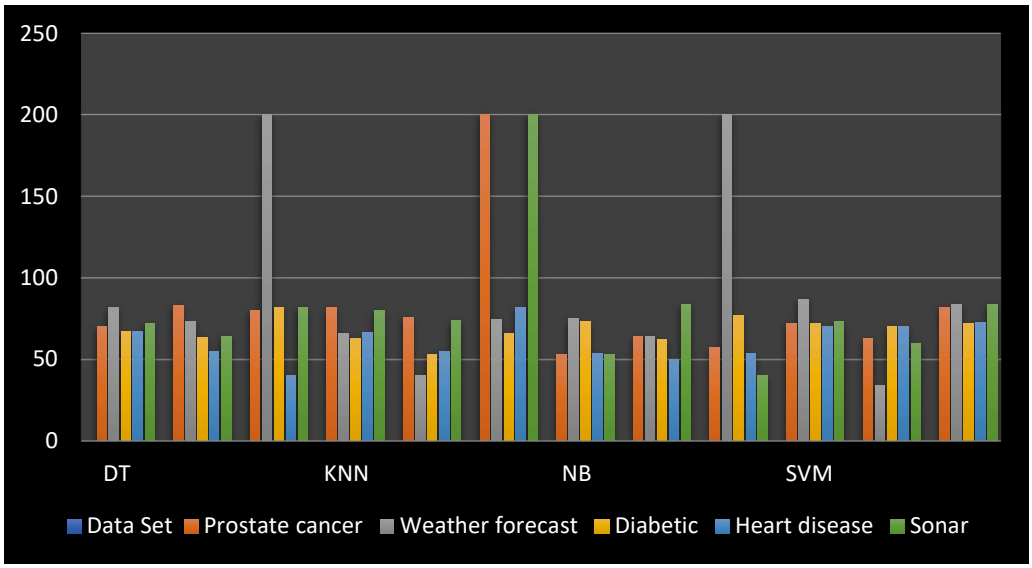


Figure: 4. Comparative Results of ML Algorithms

Table 4.5 Accuracy Comparison of ML Algorithms

Data Set	**Feature Selection**	**DT**	**KNN**	**NB**	**SVM**
		Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
Prostate cancer	Relief (Existing)	62	64	50	73
	E-Relief (proposed)	70	82	73	92
Weather forecast	Relief (Existing)	72	73	72	50
	E-Relief (proposed)	82	66	73	87
Diabetic	Relief (Existing)	62	50	53	60
	E-Relief (proposed)	67	63	94	72
Heart disease	Relief (Existing)	53	72	62	60
	E-Relief (proposed)	67	66.3	53.4	70
Sonar	Relief (Existing)	53	70	67	66

	E-Relief (proposed)	72	80	53	73

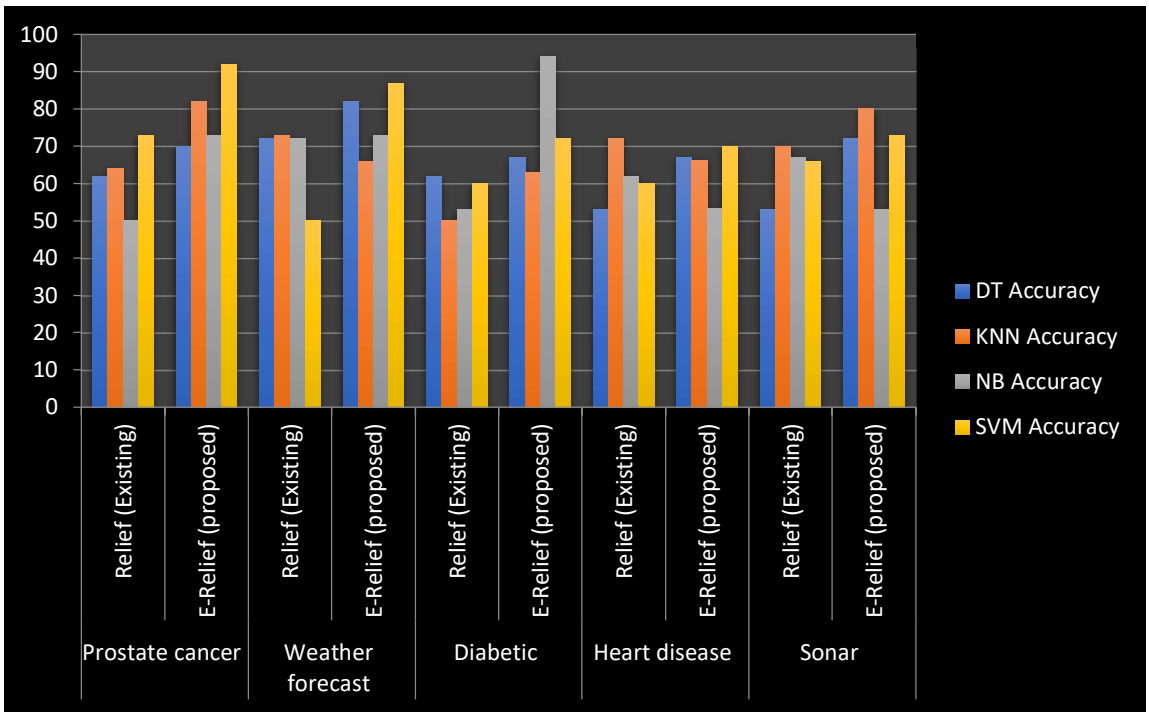


Figure: 5. Accuracy Comparisons of ML Algorithms

The table 2 and figure 3 are the comparative results of the existing and proposed approach regarding the average feature selection. The proposed approach eliminates redundancy. Hence, the numbers of features are reduced compared to the existing approach (Kumar, 2014). The comparative outcomes of machine learning algorithms are displayed in table 3 below. Ten-fold cross validation is used to divide the data into training and testing groups. Calculated metrics include classification accuracy, sensitivity, and specificity averages. The number of characteristics has been decreased in the suggested method, increasing classification accuracy and efficiency. The classification outcomes of machine learning techniques are shown in Figure 4. The accuracy enhancement of the suggested feature selection is shown in Table 4. The illustration of an accuracy comparison is shown in Figure 5.

5. Conclusion

We began by summarizing the types and characteristics of massive information and providing an overview of large data. We then looked at the latest developments in AI with massive information to highlight the differences of AI methods with reference to enormous information. In light of AI techniques, we also put forth a reference structure that combines the power of distributed hoarding and equal processing to handle large amounts of data. Finally, we discussed a few exploratory challenges and open questions (L.Z. Wang, 2020). We hope that

this overview will spark additional interest in creative work on AI-based solutions for handling massive amounts of information.

Since we live in a technological era, enormous amounts of information are being gathered quickly. An important component of navigation is recognizing the information buried in enormous amounts of information. Due of the characteristics of large data, the conventional ML calculations don't work well. The presence of noise and inconsequential components is one of the main challenges in managing massive amounts of information. Before applying AI, preprocessing must be completed on massive amounts of data. Therefore, pretreatment of large amounts of data must be completed using various dimensionality reduction techniques that have been advocated in writing (M. Chen, 2018). For the purpose of doing an opinion analysis on person-to-person communication destinations, we looked at how all element subset choosing processes would apply to vast amounts of data in this study. By applying element determination procedures, the ML calculation's exhibition and precision can be increased. We can also use the profound learning models in the future to reduce the dimensionality of vast amounts of data.

6. References

1. *A.Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2016), pp.137-144*
2. *Boston, MA: Houghton Mifflin Harcourt, 2018*
3. *C. Rudin, and K.L. Wag staff, Machine learning for science and society, Mach Learn. 95 (2014) 1-9.*
4. *Chun-WeiTsai, Chin-Feng Lai, Han Chieh Chaoand Athanasios V. Vasilakos "Big Data analytics: a survey Journal of Big Data (2015) 2:21*
5. *D. Che, M. Safran, and Z.Y. Peng, From big data to big data mining: challenges, issues, and opportunities, in Proc. of the 18th International Conference on Database Systems for Advanced Applications Lecture Notes in Computer Science (LNCS), Wuhan, 2013, pp. 1-15.*
6. *D. Yu, and L. Deng, Deep learning and its applications to signal and information processing, IEEE Signal Proc. Mag. 28 (2011) 145-154*
7. *Grünauer A, Vincze M. Using dimension reduction to improve the classification of high-dimensional data. ArXiv preprint arXiv:1505.06907. 2015.*
8. *J. Ekanayake, H. Li, B.J. Zhang, T. Gunarathne, S.H. Bae, J. Qiu, and G. Fox, Twister: a runtime for iterative MapReduce, in Proc. of the 19th ACM International Symposium on High Performance Distributed Computing (HDPC), Chicago, 2010, pp. 810-818.*
9. *J.L. Liang, M.H. Zhang, X.Y. Zeng, and G.Y. Yu, Distributed dictionary learning for sparse representation in sensor networks, IEEE Trans. Image Process. 23 (2014) 2528-2541.*
10. *Jianyu Miao (2016)," A Survey on Feature Selection", Information Technology and Quantitative Management*
11. *K. Kalpana et al (2019)," Feature Selection for Machine Learning in Big Data", International Journal of Innovative Technology and Exploring Engineering*
12. *Khawla Tadist et al (2019)," Feature selection methods and genomic big data: a systematic review",*

13. *Kumar, Dr K Ramesh & Vanaja,. (2014). Analysis of Feature Selection Algorithms on Classification: A Survey. International Journal of Computer Applications. 96. 28-35. 10.5120/16888-6910.*
14. *L.Z. Wang, K. Lu, P. Liu, R. Ranjan, and L. Chen, IK-SVD: dictionary learning for spatial big data via incremental atom update, Computer. Sci. Eng. 16 (2014) 41-52.*
15. *M. Chen, S. Mao, and Y. Liu, Big data: a survey, Mobile Newt. Appl. 19 (2018) 171-209*
16. *M. Mardani, G. Mateos, and G.B. Giannakis, Subspace learning and imputation for streaming big data matrices and tensors, IEEE Trans. Signal Process. 63 (2015), 2663-2677.*
17. *M. A. Beyer and D. Laney,,,,,The importance of „big data":A definition,"" Gartner Research, Stamford, CT, USA, Tech. Rep. G00235055, 2012*
18. *M.M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N.Seliya,R. Wald, and E. Muharemagic, „„„Deep Learning Applications and Challenges in Big Data Analytics,"" J. Big Data, vol. 2, no. 1, p. 1, Feb. 2015.*
19. *Mohamad, M.; Selamat, A.; Krejcar, O.; Crespo, R.G.; Herrera-Viedma , E.; Fujita, H. Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection. Electronics 2021, 10, 2984.*
20. *P. B. Dongre and L. G. Malik, „„„A review on real time data stream classification and adapting to various concept drift scenarios, "" in Proc. IEEE Int. Adv. Computer. Conf. (IACC), Feb. 2018, pp. 533–537*
21. *S. F. Ding, X.Z. Xu, and R. Nie, Extreme learning machine and its applications, Neural Computer. App. 25 (2019) 549-556.*
22. *S. R. Sukumar, "Machine Learning in the Big Data Era: Are We There Yet?" ACM Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good, Oak Ridge National Laboratory, pp. 1-5, December 2018.*
23. *Y.F. Zhang, and S.M. Chen, i2MapReduce: incremental iterative Map Reduce, in Proc. of the 2nd International Workshop on Cloud Intelligence, Riva del Garda, 2019.*
24. *Y.F. Zhang, Q.X. GAO, L.X. Gao, and C.R. Wang, iMapReduce: a distributed computing framework for iterative computation, J. Grid Comput. 10 (2018) 47-68.*
25. *Ying He,F.RichardYu,NanZhao,VictorC.M.Leung,HongxiYin"Soft ware-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach"IEEE Communication Magazine(Volume: 55, Issue: 12, DECEMBER 2017)*