**How to Cite**:
Varghese, J., & Selvan, P. T. (2022). A feature reduction based LDA with SVM classification on dimensionality reduction for Big Data. *International Journal of Health Sciences*, *6*(S2), 9415–9431. https://doi.org/10.53730/ijhs.v6nS2.7461

# A feature reduction based LDA with SVM classification on dimensionality reduction for Big Data

**Jijo Varghese**
Research Scholar, Department of CS, CA&IT, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India
Corresponding author email: jijo.22@gmail.com

**Dr. P. Tamil Selvan**
Research Supervisor, Department of CS, CA&IT, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India
Email: tmselvanin@gmail.com

***Abstract*---**Analytics of Big data research has been entering the latest processes of "fast-data", in which every second many Giga Bytes of data arriving towards a massive structure of data. Based on the number, speed, importance, variation, uncertainty and veracity of the collected data, current Big data applications gather dynamic data sources and thus create massive unstructured Big Data. Data sources that are decreased and significant are deemed more valuable than raw, repetitive, unreliable, and noisy data set. A further prospect for reducing the big data whereas the thousands of attributes in large data sets are the cause of the dimensionality which takes infinite computing resources to expose working patterns of information. Not each feature in the generated datasets is essential for the training of computer algorithms. Any characteristics do not influence the effects of the forecast and some may be negligible. The ignorance of this trivial or less important characteristics lowers the pressure on the algorithms of Machine Learning (ML). The MapReduce technology in existing has also been used to decrease dimensionality, but without decreasing irrelevant features it takes all data for a direct reduction, which contributes to lower classification precision. In this study, the most common ML algorithm Support Vector Machine (SVM) Classifier has collaborated with the feature reduction method Linear Discriminant Analysis (LDA). In this work, the raw dataset is first processed by the LDA for feature reduction and then the dataset was further classified by the SVM classifier for good accuracy. The evaluation metrics such as Information-Ratio for Dimension-

Reduction, Accuracy, and Recall, indicate that LDA with the SVM method established a better outcome than the Map-Reduce.

***Keywords*---**Big Data, data reduction, map reduce, linear discriminant analysis, support vector machine.

## Introduction

A representation of Bigdata, as shown in Figure 1, may be described as an accumulation of enormous and multiple formats of streaming data, which come from diverse data providers into a single pool of data [1]. Storage capacity is one of the most obvious features of big data since it is typically described as volume by the number of storage slots one acquires in massive level data warehouses and database servers. The diversity in the datasets that are driven by the vast quantity of big data not merely has the negative effect of causing data heterogeneous and leads to the multidimensional complexity in the datasets. As a result, measures must be taken to minimize the quantity throughout evaluating large data properly [2]. Furthermore, huge amounts of data feeds must be handled instantly to prevent consuming lateral resources for processing and storage.

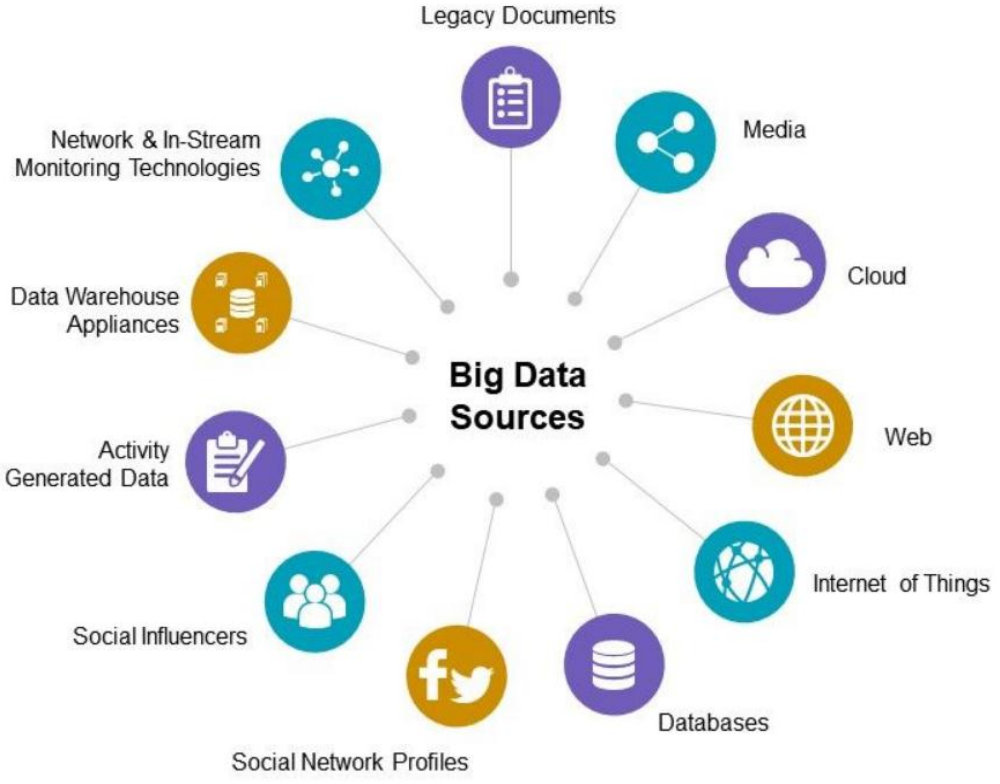


Figure 1: Sources of Bigdata

The velocity of Bigdata is the second most important feature. It relates to the regularity of streaming data, which can only be controlled in particular to properly manage large amounts of data. In this case, solar kinematics observatories produce almost one terabyte of data each day, which would make it almost impossible to interpret the information quickly, which must be reduced or simplified first [3].

Big data inherited the 'dimensionality curse' on the other side. For a deeper understanding, it may also say that if users want to reveal the greatest knowledge patterns, they need billions of dimensionalities (parameters, characteristics, descriptors) to be properly minimized [4]. The online consumers, for instance, have dense, high-dimensions behaviour profiles that consist mostly of searching, webpage hits and bookmark data, with thousands of extra phrases and links [5]. In addition to increasing the velocity and volume of data, personalized large genome analysis additionally contributes to the higher dimensionality of data [6]. As a result, it's critical to limit the number of dimensions despite keeping the most significant and valuable information.

While building successful dimensions reductions algorithms, several factors need to be discussed computing effectiveness, accuracy rate, and consistency with noise information. The lesser the dimensionality, the better the effectiveness from a computing standpoint. When comparing to others that use bigger multidimensional datasets, the lowered dimensions must keep essential characteristics required to enable matches, though perhaps appropriate superior accuracy of classification. Mostly from standpoint of consistency, the method must exhibit little or no efficiency decline although when dealing with noisy affected datasets, which would be common in actual environment information collection. Dimension Reduction (DR) approaches might certainly assist with the diversity and vastness of large amounts of data through condensing millions of variables together into reasonable level [7].

For preparation, clustering frequency information deduplication, repetition removal, and application of networking (graphs) theoretical ideas, DR approach for large data range with purely dimensionality reduction approaches to compressed dependent DR approaches and other strategies. Those methods are normally used after the data has been collected. Likewise, clustering deduplication and repetition removal techniques that reduce data duplication for fast data analysis and valuable information extraction are generally used after the data has been collected [8].

Over the last 2 decades, Machine Learning (ML) has been one of the most rapidly increasing technology solutions. It seems to have a wide range of applicability in domains such as image recognition, genomics, digital marketing, medicine, finance, criminal identification, and trends forecasting, among others. Those would have an impact on data sampling and the structures present among the information. Such methods have been employed in many environments to anticipate and categorize testing information in an attempt to generate reliable findings.

Every single technique employs a single criterion either from an unsupervised and supervised standpoint, with supervised techniques assuming a previous understanding of classes allocation for training samples, whilst unsupervised techniques do not. The hybrid technique, but at the other extreme, combines both requirements. In comparison to a wide range of single DR techniques, the hybrid techniques have so far been convincing because it uses both of the supervised characteristic, which produces mapping feature vector designed to improve the accuracy of classification, and the unsupervised set of standards, which produces mapping feature vector which effectively encompasses the initial information.

Feature extraction would be a crucial pre-processing phase that facilitates the extraction of customized features from original data, which simplifies the ML framework and enhances the classifying method's outcomes. The proportion of ML professionals' effort is spent on filtering and feature extraction [9].

**The following are a few instances of feature extraction**

- Categorical feature decomposition
- Date-Time feature decomposition between hours of each day, a section of each day, days of each week, and so on.
- To simplify the dimensionality of the incoming data, features are replaced with quantitative data.

The problem statement of this research is to discover lower dimensionality frameworks throughout a higher dimensionality collection of data. Due to various recent advances in domains which including image processing and computer vision, bioengineering, online predictive analytics, and other fields in which the dimensions of the observing region may quickly exceed large numbers.

The main motivation of this research is to explore how DR methods affect the effectiveness of ML methods. In today's competitive environment, practically every industry generates a large volume of data. The ML techniques have been deployed to uncover significant patterns in that data, that would empower administrators and entrepreneurs to make better choices. The temporal complexities of the training process may be greatly reduced using DR approaches, which lessens the workload on the computations. The LDA has been implemented to extract the basic features for dimensionality reduction in this research, and it was being formulated toward the extensively leveraged ML supervised classifier SVM by employing freely available data sources from the UCI machine-learning library.

**The following is the research contribution**

- The presented work's initial move is by using LDA on single datasets to identify the most significant variables for DR.
- The collected features are then used for training with the SVM method.
- The effectiveness of the proposed approach would then be especially in comparison with that of the existing Map-Reduce technique employing a multitude of metric criteria.

The rest of this article is categorized as follows whereas Section 2 collects the related review approaches regarding the problem statement of this research, Section 3 briefs about the methodologies of both existing and proposed methods in details, Section 4 deals with the experimental comparison results for existing and the proposed model and finally, Section 5 concludes the article with future scope.

**Related Works**

In [10] the authors described different alternatives for training arbitrary labelling-based criteria for DR. Its initial strategy is often reverse-engineered stacked technique that could be implemented on edge of a regular ruling training method whereas the other technique moves another level even more by modifying the well-utilized divide and conqueror technique for training multiple labelled constraints.

In [11] the authors introduced a fast and robust combination approach relying on Farther-Distance Based on Synthetic-Minority-Oversampling Strategies (FD SMOTE) plus Principal-Component-Analysis (PCA), that effectively decreases the large dimensions and maintains the imbalanced data. In [12] the authors implemented the technique for classification merging with all of the semantic consistency assessment predicated on internal dependability and comparative dependability ideas in which the internal dependability has been portrayed by a matrix and categorizes the possibility of the central component to one category whenever it is categorized into some other category for DR.

In [13] the authors introduced a unique multiple labelling subset of features for DR that classifies labelling into 2 classes: autonomous labelling and based labelling and analyses the distinctions among autonomous labelling and based labelling by inventing a novel feature significance concept, such that, the contingent similarity measure among both nominee features and every label must have chosen to allocate those certain labelling. In [14] the authors developed a DR approach and detecting characteristics of cardiovascular illness by using a selective features methodology and found that Chi-square and PCA with Random-Forest (RF) had the greatest accuracy. Hence, the employment of PCA straight as from original information generated inferior outcomes and might need larger dimensions to enhance the quality and efficiency

**Methodologies**

The procedure included throughout this research is illustrated in Figure 2. The subsequent sections go through all of the 5 stages that describe these strategies in clarity:

Stage 1:
As a result, it prefer to get each of the Data Sets (DS) and records that contain a lot of data in a single location one Data Base (DB) through this scenario. The effectiveness of the merging procedure will be discovered during this moment. Joining numerous DS has been the system that integrates numerous DS into one large DB. The arrangement including its rows from every segment depending on

similar features or columns has been one of the essential, and also highly sophisticated, processes in this operation. It enables the inclusion of multiple sources of data unless a sufficiently big DS has been assembled where the predictions may be derived and generalized with another DS.

Stage 2:
The ETL (Extract, Transform, and Load) mechanism is developed in stages 2 and 3. This preparation enables firms to transport information from diverse origins restructure and filter it, and then upload this into a new Data Warehouse (DW). The information was extracted from the sources of information and loaded further into DW using this procedure.

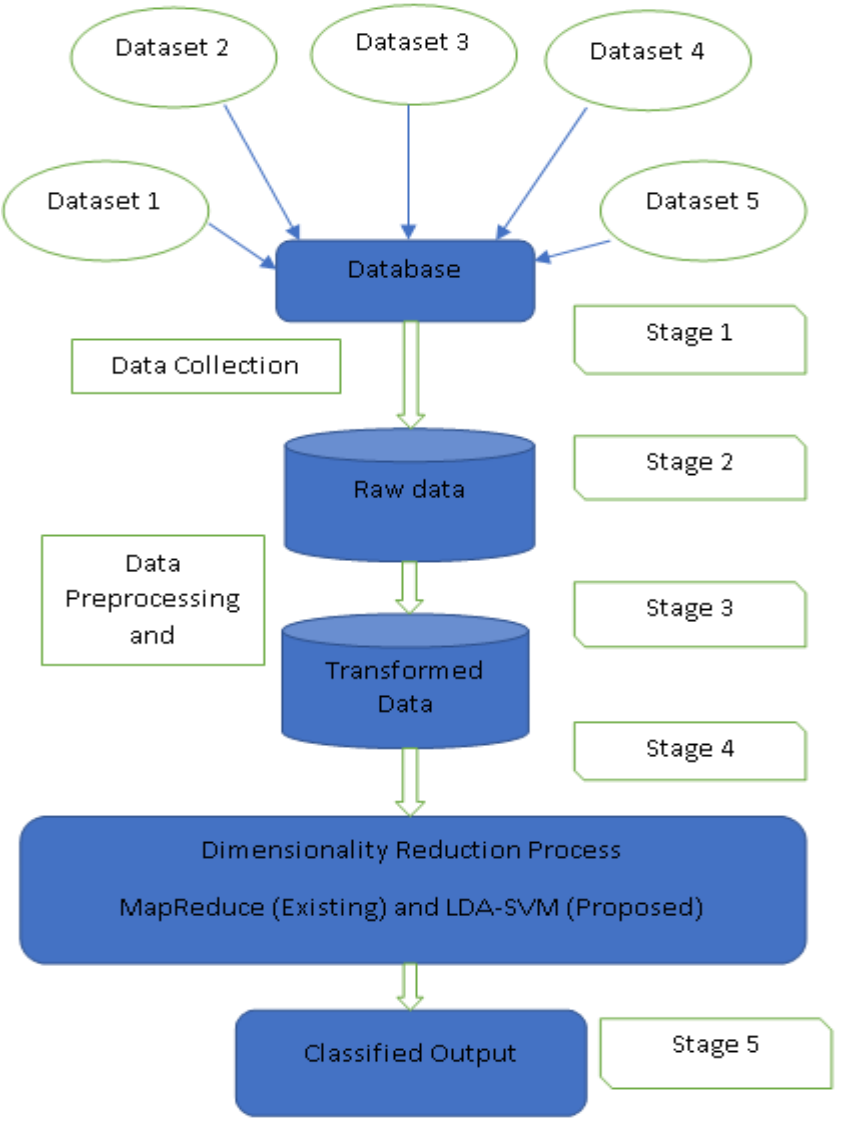


Figure 2: Architecture for the Proposed Methodology

To accommodate the information mostly to DW architecture, it must be evaluated, cleansed, incorporated, and converted. The next sections go through those workflow stages throughout the ETL procedure in further process. The presented approach, specifically in stage 2, conducts the extracting features, which is the initial job in the ETL procedure the information was extracted first from the origin and placed in the scaffolding area during the extraction activity.Most transformations usually performed throughout the scaffolding area so that the origin's performance was not compromised. Moreover, if corrupted information is moved directly from the original into the finalized DB rollback would become impossible. The scaffolding area allows users to double-check the features extracted once it is transformed into the finalized DB.

In addition, there still are different ways to extraction:

- Extraction in its whole.
- Partially extracted data sometimes without or with notice of updates.

This would also perform few checks throughout this procedure, such as:

- Mostly with data sources, keep a checklist of the adaption.
- Ensuring whether no undesirable material, such as spam, has been loaded.
- It’s often a good idea to double-check the type of data.
- Data that is fragmented must be eliminated.
- Lastly, the important area must be examined.

Stage 3:
Extracted features again from the preceding stage is unusable in their actual state. As a result, it must be cleansed, plotted, and converted. In reality, this would be the crucial phase in the ETL procedure since it appends entries and alters information. It executes customizable procedures or functional aggregation on the features extracted during this stage.
There were many numerous types of information that are referred to as passing through information. There is no requirement to transform this data anything further. Since data merging is among the biggest difficult stages, there arise often several issues.

It highlights a few of those here:

- Hardware that is committed.
- Errors due to human mistake and data transport.
- Viruses and bugs.

This should also perform various checkpoints throughout this procedure, such as:

- Filtration, data normalization and encoder management are all things that could be done.
- Change in measurement unit.
- It's important to double-check the information boundary.

- Checking flow of data from the scaffolding to the intermediate tables.
- Fields that are required should never be left blank.
- All columns and rows were cleaned and replaced.
- Undertaking some information verification that is challenging.

The final stage therefore in the ETL procedure is to load the data into the destination DB. The loading procedure must be improved in addition to enhancing it. Recovering procedures should have been allowed to be initiated since the final position while this approach fails but without affecting the integrity of data. Load as initial, Load as incremental, and refresh fully are indeed the 3 forms of loading.

The loading procedure, like the two prior ETL processes, contains validation strategies:

- There should not be blanks or incomplete information in specific areas.
- Modelling perspectives relevant to the destination tables are put to the test.
- It's always a good idea to double-check the combined variables and measuring estimation.
- Lastly, examine the information in the dimensional and historical tables, and also reporting upon this loaded information.

Processing of data and integrating would be performed in accordance to complete a quality assurance procedure which will transform the DS into a sequence of usable information that would be used to employ ML algorithms in the next stage.

Stage 4:
Even though all of the preceding stages (1–3) are component of the entire sequencing throughout the processes of data processing and information retrieval, the relevance and transcending of these stage increases. The work involves using an existing MR methodology on its own, and also be using the presented SVM methodology in conjunction with LDA to reduce the dimension was done in this stage.

Curse of Dimensions seems to be an issue that's been explored and addressed with many techniques all through the literary works, but it is not a problem that has been resolved with better accuracy, therefore working on that is constantly a challenging issue. This research would be involved in acquiring improved accuracy and correlations amongst source and destination variables. It is important to remember that these source variables look to have been modified from their original data. During this stage, existing and proposed approaches are compared and complemented.

**MapReduce (MR) (Existing Model)**

Big data are commonly quantified in Terabytes and Petabytes, indicating that the entire content cannot be kept on a solitary hard drive, because data grows rapidly quickly via continuous accumulation. As a result, numerous discs are required. To evaluate such type of information collected, a revolutionary idea known as

cluster-based computing is gaining traction, wherein the data concurrent calculations are performed in a distributed manner on clusters of workstations.

MR [15] seems to be the pioneering effort in this field. In MR, several machines have been doing similar operations on a vast volume of records. By using the MR, a task's Mapping and Reduction operations must be specified. The task frequently separates the incoming information into tiny, separate DS that the Mapping operations handle concurrently. Regarding the concluding outcome, the Mapping tasks' various outcomes are becoming the Reducing task's sources. By distributing pairs of workloads to multiple computers, the MR could manage breakdowns of computing tasks. Hadoop, a free software MR platform, is renowned mostly among research universities although in a wide range of companies.

**Disadvantages**

Unfortunately, there are two concerns that pose a threat to the effectiveness of the whole MR method:

- Finding a subdomain that effortlessly blends multiple requirements into a unified model is tough.
- The method's adaptability capabilities or resilience of execution while dealing with noise information

**LDA-SVM (Proposed Model)**

**Linear Discriminant Analysis (LDA)**

LDA has been a supervised approach that focuses on distinctions between recognized classes, making it highly convenient in extracting features. Fisher's-Discriminant-Analysis (FDA) is another name for LDA. Unlike the Principal-Component-Analysis (PCA) technique, LDA takes into account both inter-class and intra-class correspondence.
To ensure the effectiveness of extraction of features and dimensionality minimization LDA additionally translates data towards vector space, maximising among class distances and minimizing inside class distances.

**The following are the steps of the LDA technique**

a) Evaluate the scatter matrix inside each class.

Minus the mean from the total. The data input vector was being configured as follows:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \quad \text{Eq} \rightarrow 1$$

Within the matrix

$$\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^p] \quad \text{Eq→2}$$

Here 'N' denotes the sample size and 'p' denotes the number of sample attributes. This deducts every one of its information mostly with the mean result of all other information to every '$x_i$'.

Equations (1) and (2) represent the data input. The data collected has often been categorized as 'C' classes. Next, compute the scattered matrices inside each class as

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathbf{X}_i} (\mathbf{x} - \mathbf{m}_i) \cdot (\mathbf{x} - \mathbf{m}_i)^T \quad \text{Eq→3}$$

in addition

$$\mathbf{S}_w = \sum_{i=1}^{C} \mathbf{S}_i \quad \text{Eq→4}$$

Here, 'mi' has been the data's mean within the similar class as per Equation (3).

b) Compute the scattered matrices across classes.

The scattering matrices across classes is computed as

$$\mathbf{S}_B = \sum_{i=1}^{N} n_i (\mathbf{m}_i - \mathbf{m}) \cdot (\mathbf{m}_i - \mathbf{m})^T \quad \text{Eq→5}$$

The sample size throughout this classes is '$n_i$', and the mean among all samples is 'm'.

c) Identify the eigen values and address the problems.

Once solved the generalised eigen value solution, may go on to the next step.

$$\mathbf{S}_B \mathbf{V} = \Lambda \mathbf{S}_w \mathbf{V} \quad \text{Eq→6}$$

The eigen vectors 'V' and the eigen values are obtained here.
The eigen values are ordered and irrelevant eigen vectors are deleted using similar processes. Merely C-1 eigen values should be no zero, and data input should be restricted to C-1 or maybe less dimensional. The discreet 'V' is sometimes employed to reduce the dimensionality of the testing set. These values are passed to the SVM classifier.

**Support Vector Machine (SVM)**

SVM has always been a binary classification that has been utilized within this research. SVM has been a supervised ML method that is handed with a collection of inputs and responding labels. The data is presented as vector attributes. To obtain maximal distance amongst given classes, SVM creates a hyperplane that divides them. Generalization errors were reduced by splitting the categories according to a considerable interval. While information comes for classifying, the intention of attaining the smallest generalization errors would be to anticipate the proper classes of its data having no or little errors.

**Procedures for DR**

The Eigen decomposing of the covariance matrix is used to execute the DR. The covariance matrix of the actual DB is produced, and also the Eigen values and Eigen vectors on every DS were computed. Every DS scores are computed, and also the associated components are retrieved. The data within those components are arranged in decreasing order of variability. According to the preceding claims, the very initial component retrieved would include the data with the greatest variability, and thus the variability will reduce. As a result, the first components would contain the most data possible. Choosing the optimal quantity of modified data for getting dimension minimized data, on the other hand, is a difficult problem.

**Altered SVM Rules for DR**

Following obtaining the minimized data from LDA's whole database and calculating its variability, the next step is to determine the frequency of features for any subsequent analysis and processing. The inherent dimensionality of such DB refers to the operation of picking optimal features that include significant information about the DS. The set of features is calculated using typical statically techniques by determining the range of features that fall inside a threshold based on the DS's % variability. The variability is usually set between 97 and 99 percent and gets inherent dimensions within the range of 3 to 6 using the approaches. When the range of features was modest, it does provide excellent performance.

Therefore, when a DS with extremely higher dimensions was chosen, such as hyper spectral information the dimensionality produced by the LDA approach is insufficient. Because the dimensions were very large, a lower inherent dimension would have been impractical, and also the odds of detecting all features were limited. As a result, increasing the threshold may enhance the likelihood of appropriate selecting features and recognition from hyper spectral information. Therefore, it is necessary to determine what threshold should be used. It cannot be a random number. To determine if a changed rule was being applied to compute the hyper spectral information's inherent dimensions. The LDA-based DR technique relying on Eigen vector evaluation, automated picking of appropriate converted components, including concurrent classifying of hyper spectral information applying non-linearity SVM are all part of the SVM-based DR methodology proposed in this research.

**The following is a numerical representation of the proposed methodology**

Let "X ($x_i$, $y_i$)" be the data input and "$X^T$ ($y_i$, $x_i$)" transposed matrices of the data input, with "i = 1, …, n ∈ $R^2$" being the primary input.

(a) Compute the data covariance indicated by 'K' as follows:

$$K = \operatorname{cov}(X, X^T) = \sum_{i=1}^{n} \frac{(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)} \quad \text{Eq→7}$$

(b) To extract the Eigen vector and Eigen values, implement Eigen decomposition to K:

$$V^{-1}KV = D \quad \text{Eq→8}$$

The Eigen values matrices are 'V', while the Eigen vector matrices are 'D'.

(c) Arrange the 'V' values from highest to lowest in order of decreasing as:

$$K(p,q) = V(p,q) \quad \text{Eq→9}$$

Here "p=1…m, q=1…l, and 1≤ l ≤ m".

$$\frac{\lambda_j}{\sum_{i=j}^{p} \lambda_i} > b_j \quad \text{Eq→10}$$

While "j=1,2, …, k" and

$$\lambda_{k+1} \leq b_{k+1} \quad \text{Eq→11}$$

Equation (10) represents a reasonable portion of the overall variability expressed as "λj" inside "λj…, λp". The inherent dimensions are given by the term 'j'.

$$C = \frac{(x_i - \bar{X})}{s.h} \quad \text{Eq→12}$$

The normalised value is denoted by the character 'C'.

(d) The data has been dimensionally lessened and presented as follows:

$$Y = W^T.C \quad \text{Eq}\rightarrow 13$$

(e) The following are the steps to deploy SVM to the forecasted data:

$$y_i(w.x_i + b) - 1 + \xi_i \geq 0 \quad \text{Eq}\rightarrow 14$$

With "i= 1,2…, l". The Radial-Basis Φ () Function is used here.

Stage 5:
Predictive model dashboards would be created following the preceding stage to support forecasting, projections, visualizations, and the extraction of patterns and rules among other things. This stage is critical for comprehending most of those connections, which aren't usually simple to come through. In reality, when dealing with enormous datasets, visualization is critical since the data's inherent complexity makes analysis impossible without reference points.

**Results and Discussions**

Evaluation is a vital stage since it provides real analysis of the findings acquired, i.e. the correctness of the categorized information. Evaluation information was gathered from several areas around the research region and categorized DS throughout the whole DB is conducted using these input parameters, often referred to actual content. The DS were a pre-processed and reorganized or transformed form of a widely deployed DS retrieved from the UCI-ML Collection. The primary DB is made up of five separate medical DS, all with 200 documents that comprise a unique subject. The Java platform was used to develop the proposed model.

**Dimensionality Reduction**

The very primary process of this research has been to compute the Information-Ratio (IR), which represents the real characteristics of the source information, utilizing eigen values and eigen vectors with the MR and LDA-SVM methods. The entire DB comprises 5 DS, and when reduced in various dimensions ranging from smaller 10 to higher 90, the IR varies. The findings in Table 1 and Figure 3 revealed that LDA-SVM can achieve a high IR in smaller and higher dimensions, but MR had a lower IR in this implementation.

| DATA DIMENSIONS | MR | LDA-SVM |
|---|---|---|
| 10 | 87.89 | 94.53 |
| 30 | 89.23 | 95.12 |
| 50 | 91.32 | 96.23 |
| 70 | 93.42 | 97.32 |
| 90 | 95.21 | 98.23 |

Table 1:Numerical DR Comparison

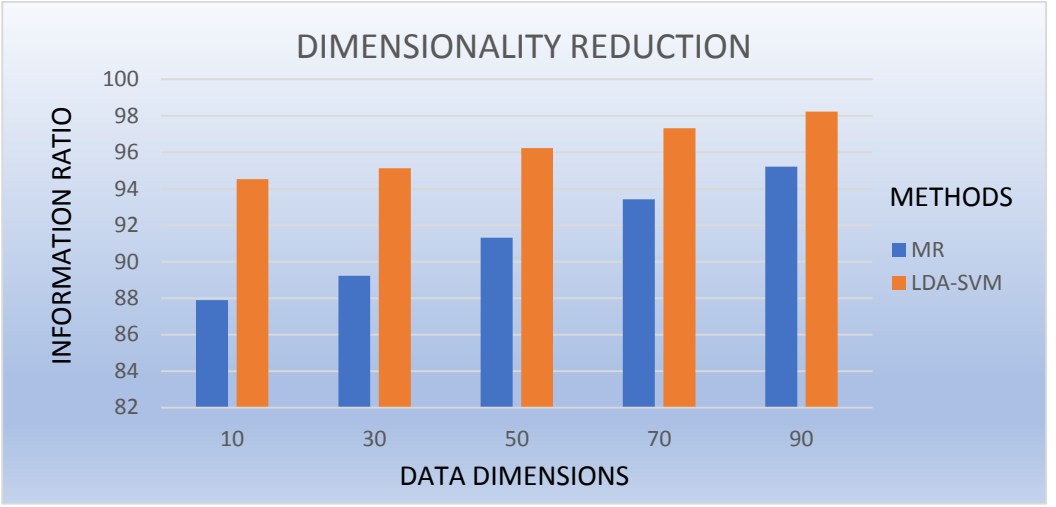


Figure 3:Graphical DR Comparison

**Results of Accuracy**

The confusion-matrix between both the actual information and also the categorized DS is used to measure accuracy. The following formula should be used to determine accuracy:

"Accuracy = (True-Negative + True-Positive) / (True-Negative + True-Positive + False-Negative + False-Positive)**"**

This procedure is carried out on all categorized DS. The findings are produced after evaluation upon categorized DS utilizing actual reality information. The maximum accuracy of the categorized DS was lower without extracting features of DR by existing MR and higher for following extracting features of DR by proposed LDA-SVM. Which might be seen as a favourable DR outcome. This stage illustrates an essential fact: when DR is combined with extracting features on the DS, an accuracy rate of the classifier improves, and so this research effort implies phase is required. Table 2 and Figure 4 demonstrate the accuracy of the classified results comparison for MR and LDA-SVM with various DS in the entire DB.

| NUMBER OF DATA SETS | MR | LDA-SVM |
|---|---|---|
| DATA SET 1 | 89.2 | 94.9 |
| DATA SET 2 | 91.4 | 96.3 |
| DATA SET 3 | 94.2 | 99.1 |
| DATA SET 4 | 93.1 | 98.4 |
| DATA SET 5 | 92.3 | 97.3 |

Table 2: Numerical Accuracy Comparison

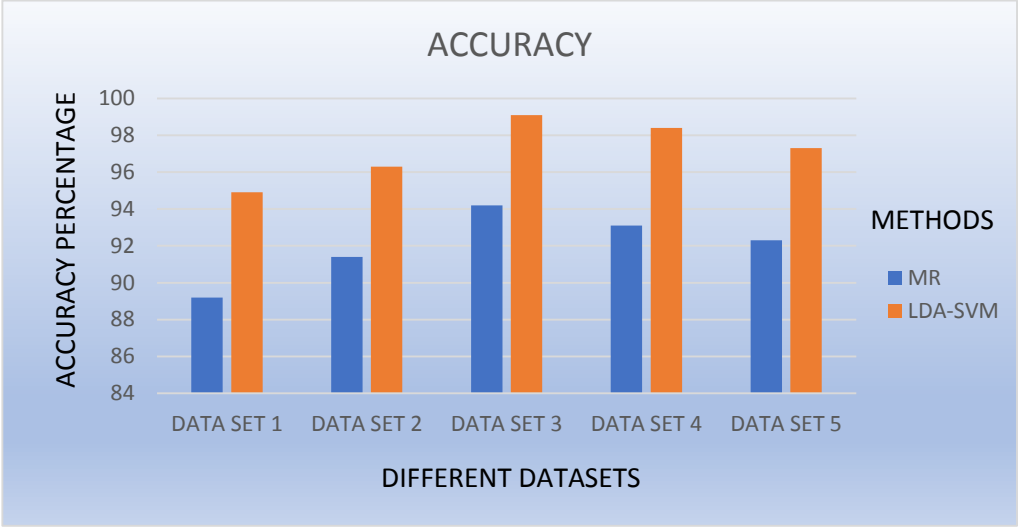


Figure 4: Graphical Accuracy Comparison

**Results of Recall**

The Recall metric assesses the method's capacity to fetch all objects with significant features throughout the whole database. The following formula may be used to determine recall:
"Recall = True-Positive / True-Positive + False-Negative"
Therefore, in this research, recall is defined as the proportion of real data accurately categorized from the whole database. This procedure is carried out with every categorized DS. The findings are generated after evaluating the categorized DS and utilizing actual reality information. The cumulative recall of categorised DS was lower without extracting features of DR by existing MR and higher after extracting features of DR by proposed LDA-SVM. This also might be seen as a favourable DR outcome. Table 3 and Figure 5 demonstrate the categorization recall outcomes comparison for MR and LDA-SVM with various DS in the total DB.

| NUMBER OF DATA SETS | MR | LDA-SVM |
|---|---|---|
| DATA SET 1 | 88.1 | 93.8 |
| DATA SET 2 | 90.3 | 95.2 |
| DATA SET 3 | 93.1 | 98.2 |
| DATA SET4 | 92.2 | 97.3 |
| DATA SET 5 | 91.3 | 96.2 |

Table 3:Numerical Recall Comparison

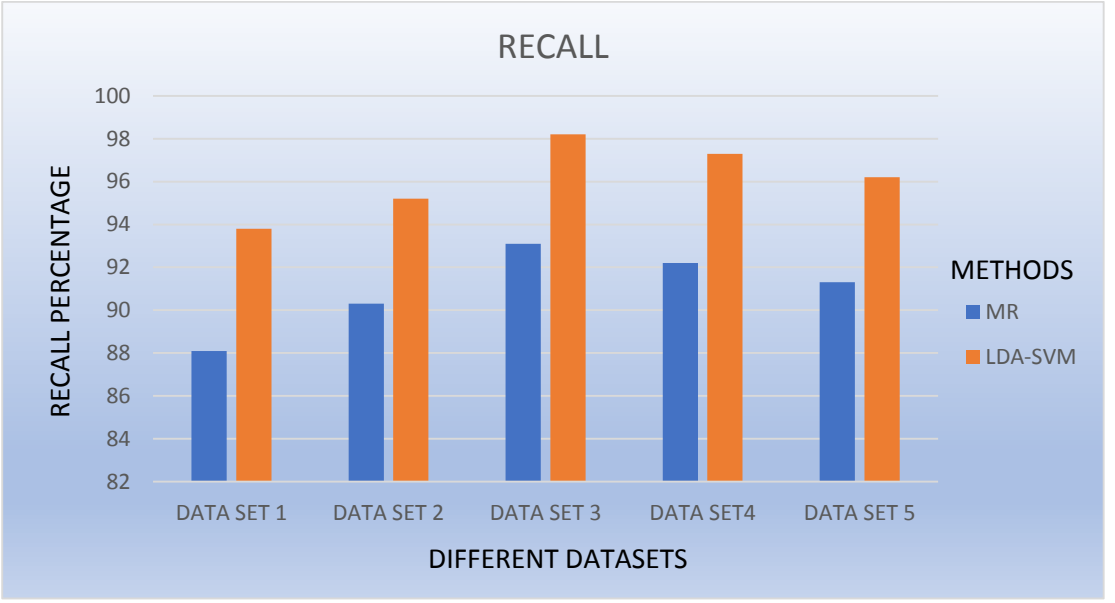


Figure 5:Graphical Recall Comparison

The integrated architecture of DR and classifying is a fundamental component of our research attempt. Here are several benefits to utilizing a fully integrated architecture. The primary is how the operating procedure would be shortened. Rather than a two-step procedure, the task is completed in a single phase, making the operation quicker and better effective as shown by the outcome of this research. The use of SVM throughout the architecture is the application's next benefit. SVM would be a strong and effective classification technique. When SVM is used to normal datasets with smaller dimensions, only simple process is required. The findings of the study suggest that using LDA for DR approaches associated with SVM enhances the accuracy of classification for larger dimensionality datasets.

**Conclusion**

The DR task is already in charge of decreasing the size of a huge set of features together into a consolidated decreased feature subset that forms a massive sphere in large dimensions area. As a result, the DR consideration has benefits such as computing effectiveness and duplication removal, and also downsides such as loss of data and features loss in databases. These have experience in the domains of DR in massive DS. In this research during the initial stage, such ML approaches for minimizing dimensionality in enormous DS integrate all DS into a massive single DB. Following the ETL procedure, it employed LDA for DR by extracting features and SVM for DR classification to tackle the difficulties observed in the MR technique without extracting features. As a result, this innovative methodology focuses on the usage of the Java platform, in which conventional quantitative data presentation and implementations for arithmetic operations at quite a top standard of coding are conducted in an effective context. Furthermore, those algorithms were used in the medical DS released by the UCI-ML Open-Source. The practical findings reveal that LDA-SVM outperforms the MR

approach with this complicated DS, achieving an accurateness close to 99 per cent. Considering the research findings the presented method's functionality should be evaluated with adequate actual reality datasets at a comparable research location in the future before it is declared operational.

**References**


1. Wu X et al (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107.
2. Che D, Safran M, Peng Z (2013) From big data to big data mining: challenges, issues, and opportunities. In: Database systems for advanced applications.
3. Battams K (2014) Stream processing for solar physics: applications and implications for big solar data. arXiv preprint arXiv: 1409.8166.
4. Fan J, Han F, Liu H (2014) Challenges of big data analysis. Nat Sci Rev 1(2):293–314.
5. Chandramouli B, Goldstein J, Duan S (2012) Temporal analytics on big data for web advertising. In: 2012 IEEE 28th international conference on data engineering (ICDE).
6. Ward RM et al (2013) Big data challenges and opportunities in high-throughput sequencing. Syst Biomed 1(1):29–34.
7. Vervliet N et al (2014) Breaking the curse of dimensionality using decompositions of incomplete tensors: tensor-based scientific computing in big data analysis. IEEE Signal Process Mag 31(5):71–79.
8. Fu Y, Jiang H, Xiao N (2012) A scalable inline cluster deduplication framework for big data protection. In: Middleware 2012. Springer, pp 354–373.
9. A. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Newton, MA, USA: O'Reilly Media, 2018.
10. E. L. Mencía and F. Janssen, ``Learning rules for multi-label classification: A stacking and a separate-and-conquer approach,'' Mach. Learn., vol. 105, no. 1, pp. 77-126, Oct. 2016.
11. N. Mustafa, R. A. Memon, J.-P. Li, and M. Z. Omer, ``A classification model for imbalanced medical data based on PCA and farther distance based synthetic minority oversampling technique,'' Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 1, pp. 61-67, 2017.
12. Z. Liu, Q. Pan, J. Dezert, J.-W. Han, and Y. He, ``Classifier fusion with contextual reliability evaluation,'' IEEE Trans. Cybern., vol. 48, no. 5, pp. 1605-1618, May 2018.
13. P. Zhang, G. Liu, and W. Gao, ``Distinguishing two types of labels for multi-label feature selection,'' Pattern Recognit., vol. 95, pp. 72-82, Nov. 2019.
14. A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, Classification Models for Heart Disease Prediction Using Feature Selection and PCA. Amsterdam, The Netherlands: Elsevier, 2020.
15. R. M. Gahar, O. Arfaoui, M. S. Hidri and N. B. Hadj-Alouane, "A Distributed Approach for High-Dimensionality Heterogeneous Data Reduction," in IEEE Access, vol. 7, pp. 151006-151022, 2019, doi: 10.1109/ACCESS.2019.2945889.