International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

An Adaptive Firefly Optimization (AFO) with Multi-Kernel SVM (MKSVM) Classification for Big Data Dimensionality Reduction

Jijo Varghese
Research Scholar,
Department of CS, CA&IT,
Karpagam Academy of Higher Education,
Coimbatore, Tamilnadu, India

Dr. P. Tamil Selvan
Research Supervisor,
Department of CS, CA&IT,
Karpagam Academy of Higher Education,
Coimbatore, Tamilnadu, India

Abstract:

The data's dimensionality had already risen sharply in the last several decades. The "Dimensionality Curse" (DC) is a problem for conventional learning techniques when dealing with "Big Data (BD)" with a higher level of dimensionality. A learning model's performance degrades when there is a numerous range of features present. "Dimensionality Reduction (DR)" approaches are used to solve the DC issue, and the field of "Machine Learning (ML)" research is significant in this regard. It is a prominent procedure to use "Feature Selection (FS)" to reduce dimensions. Improved learning effectiveness such as greater classification precision, cheaper processing costs, and improved model comprehensibility are all typical outcomes of this approach that selects an optimal portion of the original features based on some relevant assessment criteria. An "Adaptive Firefly Optimization (AFO)" technique based on the "Map Reduce (MR)" platform is developed in this research. During the initial phase (mapping stage) the whole large "DataSet (DS)" is first subdivided into blocks of contexts. The AFO technique is then used to choose features from its large DS. In the final phase (reduction stage), every one of the fragmentary findings is combined into a single feature vector. Then the "Multi Kernel Support Vector Machine (MKSVM)" classifier is used as classification in this research to classify the data for appropriate class f"om the optimal features obtained from AFO f"r DR purposes. We found that the suggested algorithm AFO combined with MKSVM (AFO-MKSVM) scales very well to high-dimensional DSs which outperforms the existing approach "Linear Discriminant Analysis-Support Vector Machine (LDA-SVM)" in terms of performance. The evaluation metrics such as Information-Ratio for Dimension-Reduction, Accuracy, and Recall, indicate that the AFO-MKSVM method established a better outcome than the LDA-SVM method.

Keywords: Dimensionality Curse, Dimensionality Reduction, Adaptive Firefly Optimization, Map Reduce, Multi Kernel Support Vector Machine

1. INTRODUCTION

As high-throughput technology has become more widely available, the amount of information that can be mined has grown at an accelerating rate. The BD refers to the overall increase of this enormous quantity of data, which includes both dimensions of the data and sampling volume. This BD management has become more difficult in the modern-day. Conventional methods of management are no longer viable in this situation. It is thus necessary to build "Data Mining (DM)" but also ML algorithms to autonomously comprehend from such BD [1].

Most of the time, those BD are gathered with a lot of noises. Imperfections in data collection systems and the data's original origin seem to be the factors responsible for the noise. When it comes to medicinal imaging, a weakness in a camera might be seen as noisy in the final phase of the task.

The extraction of meaningful information and trends from social networks is also seen as a difficult assignment by many. As a result, retrieving meaningful information from such large and unorganized datasets is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

often hampered by the presence of grammar errors, typographical problems, and poor formatting [2].

Learning from such big DS is one of the major issues for the current ML algorithms. Recent advances in the "Cloud Computing (CC)" environment allow applying standard ML algorithms over massive amounts of data. Redesigning of ML algorithms along with their inclusion in parallel environments becomes necessary in case of BD problems for successful adaptation [3]. One such recent alternative is the MR paradigm, a Google innovation. MR is an efficient and adaptive system for analyzing big DSs. It is an alternate to parallelization scheme known as "Message-Passing-Interface (MPI)". MR has the advantage of fault-tolerant mechanism and simplicity. As a result, parallelization of ML techniques can be focused on using the MR paradigm.

The term "Apache-Spark (AS)" has specifically been coined to describe a unique and highly versatile process that extends the basic MR methodology. To facilitate rapid processing AS seems to be a "Cluster-Computing" technique that runs at lightning speeds. This was based mostly on the "Hadoop-MR" framework.

When making a fresh observation, classification is employed to determine which category it falls into. Data from a training-set of findings wherein categories belonging is confirmed are used to make the categorization [4]. Recognizing and classifying emails into "spam" or "non-spam" groups, dynamically allocating media groupings like "sports," "entertainment," and giving diagnostics based on the observable features of the patients [5] are a few real-world challenges whereby classifications could be employed. Data classification can be divided down into two steps, the "Training Phase" and the "Prediction Phase," as shown in Figure 1.

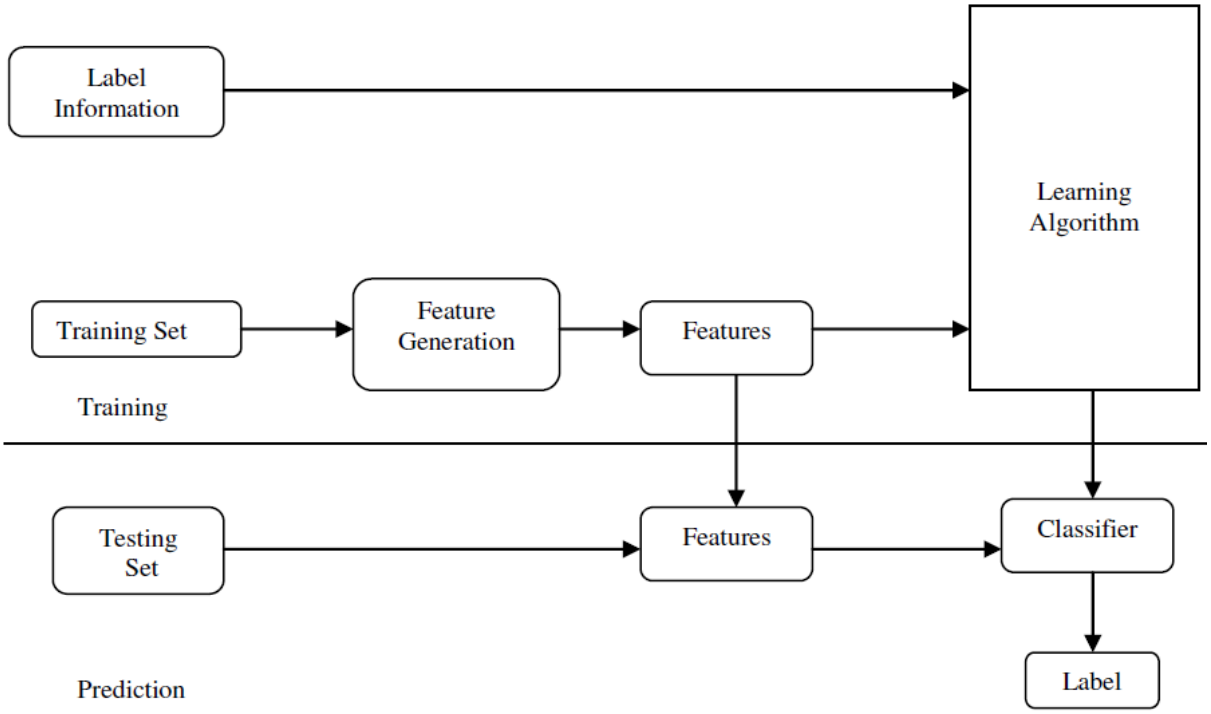

Figure 1: Data-Classification Process

These recent years have seen a dramatic rise in the amount of data available in ever more dimensions. Conventional learning algorithms have difficulties when dealing with data that has a very higher dimension "DC". The effectiveness of a learning algorithm degenerates when there is a huge range of features under it. DR approaches are used to solve the DC issue, which seems to be a significant topic of ML analysis [6].

Reduced dimensions may be achieved using the FS approach. In an attempt to select the most appropriate feature subsets, specific relevancy assessment criteria are used. That typically contributes to superior learning effectiveness such as greater prediction effectiveness, lesser computing costs, and improved modeling comprehensibility [7]. Figure 2 provides a common structure for FS categorization.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
Article Received: 05 May 2022 Revised: 30 May 2022 Accepted: 20 June 2022 Publication: 31 July 2022

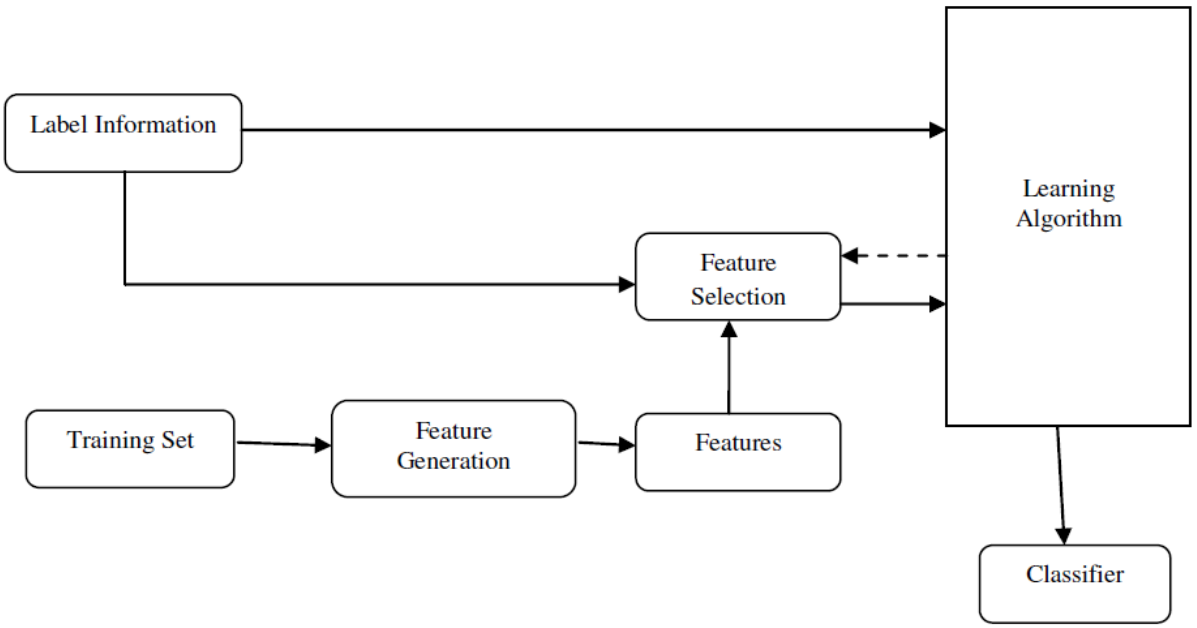

Figure 2: FS for Classification Process

During classification, rather than immediately providing the entire DS to the method, the feature subset will be initially selected by FS before it is fed into a classification model. Afterward, it feeds the relevant information to the training process, which uses the features that were chosen. During the prediction stage, a classification is created using the finalized features that were chosen [8]. The effectiveness of categorization is much improved as a result of this.

According to the research problem statement, much prior literature confines conventional FS to cluster training, which considers the FS work to be performed in a relatively static or queue learning manner. The whole of training scenarios is predetermined in this circumstance. Regarding practical uses, when training samples are presented in continuous order, those assumptions are not necessarily valid. Collecting all of the relevant training data might be costly at periods. Applications involve digital email spam tracking systems, whereby training sets are often sent concurrently, rendering it challenging to implement a conventional batch-based FS approach in a speedy, effective, and flexible way. Likewise, FS in the healthcare industry is believed to be costly because of the large costs associated with performing wet laboratory tests [9]. As a result, a more comprehensive and efficient FS is required to accommodate real-time DS.

An Adaptive FS (AFS) is the objective of this research, which attempts to address the issue by efficiently researching dynamic learning strategies. The objective is to construct adaptive classifications that utilize only a minimal and constant range of features for categorization. Medical DS about heart-rate in "Intensive Care Units (ICU)", provides higher dimensionality, consecutive testing dataset, rendering AFS a must. In the actual environment, this is a far more common circumstance. Whenever the FS process has begun in group learning, every feature must be examined. Because of this, a customizable application is indeed not conceivable in the instance of higher-dimensional DS. Thus, AFS is required, which implies that data instances are provided sequentially through FS upon the entry of every data instance [10].

Using a meta-heuristic method, the contribution of this article proposes a novel AFO-MKSVM hybrid approach that makes use of the MR development platform for selecting features. An "Optimized FireFly" and "Adaptive Annealing" algorithms are used in this AFO to determine the ideal collection of features. The primary goal of this research is to improve the accuracy of classification while minimizing the number of runtimes.

AFO-MKSVM functioning mechanism is as follows:

- The first step is to split down the entire large DS into smaller chunks of sample blocks.
- To facilitate the management of learning, this deconstruction is performed during the mapping stage.
- Next, the AFO method is used to choose features from the large DS instances.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

- Finally, the MKSVM classification evaluates the finalized features that are created after the reduction stage combines all of the fragmented findings.

The 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 detail, Section 4 deals with the experimental comparison results for existing and the proposed model and finally, Section 5 concludes the article with future scope.

2. RELATED WORKS

A "No SQL" data model was developed by the researchers in [11] to enable the superior efficiency of MR by offering two techniques for automatically splitting and parallelizing data. Researchers tried to illustrate a "No SQL" data gathering strategy that was extensible and compatible with the BD-driven development operations. The VMs (Virtual Machines) are used to perform the first technique, which divides the process into numerous clusters. Information segmentation, VM construction, and shortage minimization are all automated using the second technique. The novel DataView for processes of BD platforms, on the other hand, improves the flexibility of processes under the MR approach. Analyzing data gathered from automobiles, the technique was utilized to predict potential dangers linked to a driver's style of driving.

There's an ISOMAP streamed adaption proposed by the developers of [12]. The use of a single projection algorithm to endeavor entire information is a constraint even if these approaches could accommodate data streams. Unless the incoming new information is comparable to the data used to generate the primary purpose does the calculated projections be adequate? Because the function doesn't at all change along with the input, those approaches never describe fresh patterns or structures as they form over a period. Consequently, the layouts created are of poor quality in practice.

Integrated data and OWL ontologies were stored in a "No SQL" dataset, namely the column storage "Cassandra-Dataset (CD)" investigated by the researchers in [13]. A conceptual model was intended to save the ontology categorization in the CD and to promote OWL expansion since interconnected storage of data requires scalability datasets and dispersed stream processing. OWL ontology categorization was stored in an optimum structure using MR techniques on CD. Using the RFD data architecture based on directed graphs, types of data were distributed and linked over the Internet. This paradigm made it easier to connect CDs to OWL files.

When it comes to social media networks, security is a major concern, which is why researchers in [14] presented an "Encrypted Graph database model for online social networking services, termed (Graph SE)". This paradigm was justified by the requirement to store and compute BD in social networking sites, regulate confidentiality of data in the cloud infrastructure, and avoid security breaches using encryption. The "Graph-Data" approaches may be used in social networking sites because of their complicated and rich searching requirements. Scaling issues arise when utilizing techniques such as neighborhood searching or KNN, that restrict query processing. To address this issue, the shared and encryption graph were devised. Data from YouTube was used to examine the data efficiency of the model, operating capacity, scaling, storage capacity, and searching delay.

The "Graph based Meta-Model (GSMM)" suggested by the researchers in [15] was centered on a linear fashion and annotated graph for unstructured and structured data, followed by a "General Semi-Structured Language (GSL)". The objective was to create a platform that allowed for a wide range of feature options. Metamodeling for relational, "OEM, DOEM and XML, TGM, XML, and RDF" may be derived from this meta-model, according to the author's assertion. The "Neo4j" and "Big Table" models might likewise be represented using this meta-model. Grounded GSMM graphs would have been used to describe the data model's attributes. This article was able to do a quantitative, statistical, and customizable assessment of GSL by analyzing the limits and constraints of the two or more data models.

3. METHODOLOGIES

It's standard procedure to utilize DR to eliminate features that are useless or repetitive. The "Feature-Extraction (FE)" and "Feature-Selection (FS)" are the two categories in handling large DS. Though, the FE projects the features onto a lower-dimensional feature space, hence reducing the number of features. The fresh versions are the result of combining existing ones. When it comes to classifying, FS looks for features that are most relevant to the objective, including the number of classes, to reduce repetition and enhance relevancy. Improved learning efficiency, reduced computing overhead, improved generalization models, and reduced storage requirements are all a consequence of using these strategies.

In the instance of FE, the existing feature space is combined with additional features to create a new feature space with smaller dimensions. It gets harder to relate features from the raw data feature area to newer features. The reshaped features that result have no actual tangible

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

significance to them at this point. In this circumstance, continuing the investigation of new features is difficult.

In FS, on the other hand, a subset of the underlying set of features is deliberately selected with no change. In this way, the fundamental features retain their physical significance. As a result, when it comes to understandability and comprehensibility, FS often takes the lead. Because of this quality, FS is useful in a wide variety of real-world situations, such as identifying genes associated with a particular illness or creating a vocabulary of emotions for sentiment classification. FS seeks to pick a subset of features that are strongly discriminating in classification problems. In other words, it looks for features that can distinguish separate samples from distinct types. Owing to the derived available information when using FS for classification, the importance of features is measured as the capacity to differentiate between distinct classes.

FS was already extensively used in a range of contexts, primarily for large DS issues. High-resolution imaging and genetic data processing are just two examples of systems that use massive amounts of data. Other examples include marketing, object recognition, word classification, pattern classification, and vulnerability scanning. Higher-dimensional FS is often considered as a challenge of finding the best appropriate estimation model with the smallest number of features.

In this research, the enhanced FS is employed preliminary to classification to attain a superior DR ratio for the DC issue in medical DS.

3.1 LDA-SVM (EXISTING MODEL)

The main motivation of this existing 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 the DR process in this work, and it was being formulated toward the extensively leveraged ML supervised classifier SVM by employing freely available data sources from the UCI-ML library. This work's initial move is by using LDA on single DSs to identify the most significant variables for DR. Then the collected features are then used for training with the SVM method.

Through 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. It gets inherent dimensions within that range of 3 to 6 using those 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've been impractical, and also the odds of detecting all features would've been 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 can't 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 operates on Eigen vector evaluation, automated picking of appropriate converted components, including concurrent classification of hyper spectral information by applying non-linearity SVM are all part of the LDA-SVM based DR methodology.

3.2 AFO-MKSVM (PROPOSED MODEL)

3.2.1 AFO for Feature Selection

There exists a technique known as the "Fire-Fly (FF)" that draws inspiration from the flashing activity of fireflies. An iterative approach is used to optimize the population. To solve multi-objective optimization issues, FF is commonly employed. FF is widely used, yet it has poor local-searching capabilities besides its superior global-searching capabilities. FF seems to be a population-oriented searching engine, which means it has a high capacity for global-searching but a limited efficiency for local ones.

A further meta-heuristic technique, called "Adaptive-Annealing (AA)," is influenced by the procedure of melting an item. Heating and cooling the thing in such a way that its energy drops to zero are what is meant when the term "melting point" is used to describe a physical process. As a consequence, this method's local-searching performance is superior to its global-searching performance. The AA method mainly incorporates a neighborhood searching strategy. Because of this, the AA procedure contributes to a significant level of local-searching ability.

In this research for FS the hybrid AFO, combines the FF for global-searching and the AA for local-searching, resulting in an improved locally and globally searching to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
Article Received: 05 May 2022 Revised: 30 May 2022 Accepted: 20 June 2022 Publication: 31 July 2022

identify an ideal value. When an advance in a solution space is found, the local-searching moves the solution into the local optimal position.

WORKING PROCESS OF AFO:

Model Assumption:

When modeling the AFO process, the fundamental factors are taken into account. Representation for AA variables to identify an optimized subset of the features includes "Map" for the huge DS sample-matrix, "TempMap" for the temporary mapped-DS matrix, "Temp" for the annealing-temperature, "B" for the Boltzmann-constant, and "tempcool" for the cooling-rate.

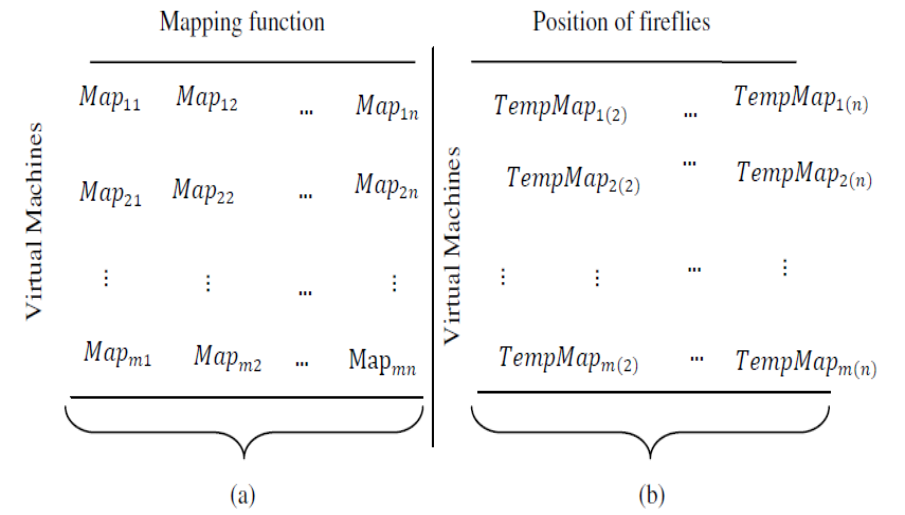

Figure 3: Encoding method to demonstrate how the actual huge DS matrices are mapped

Metaheuristic optimization techniques' effectiveness is strongly influenced by the use of an appropriate encoding system that accurately reflects the properties of a large DS sample-matrix. There are two components to this encoding system, that would be employed to show how the initial huge DS sample-matrix is mapped. Figure 3 depicts almost the same concept. A "Map_{ij}" matrix is shown on the diagonal of the "Map" sample-matrix DS, which is shown in Figure 3(a). Temporary huge DS sample-matrix "TempMap" is denoted by the "TempMapij" matrix as shown in Figure 3(b). DS samples are shown in this matrix per row, with each row representing a certain position on the machine. Also, every column with this matrix represents the total number of features used in the categorization or forecasting of the large DS.

(i) Initialization

Large DS sample-matrix "Map" has the initial value of "Mapij = 100" for every cell on the diagonal of the matrix. This means that all of a data center's FS is originally stored inside that location.

(ii) Selection of the best feature set with a neighborhood searching

According to Figure 3, two-alternative neighborhood searching architectures such as smaller and larger neighborhood architectures are employed there to locate the near-optimal set of features relying on the encoding strategy. The "*move*" operators of the FF have been used to choose the next optimum set of features in sequence to create the neighborhood of the initial searching architecture. Towards the FS issue, the 3 fundamental operators "*swapping*", "*inserting*", and "*inverting*" are taken into account in an attempt to build the neighborhood of the second searching architecture. The left-shift operators are used to transfer the temporary features from Selected-FS "*Sftemp*" to Current-Selected-FS "*Sfcurren*".

With FF, the underlying Equation (1) determines the migration of the ideal feature "*i*" towards the maximum attracting "*brighter*" alternative feature "*j*":

$$x_i = x_i + \beta_0 e^{-\gamma d_{ij}^2}(X_j - X_i) + \alpha\left(rand - \frac{1}{2}\right)$$

Eq→1

The "$\beta_0 e^{-\gamma d^2 ij}$" is the attraction function its value decreases as the spacing across the consecutive FFs "d_{ij}" increases. It is "d^2_{ij}=0" where " the "β_0" is the attraction, and "γ" is the constant light-absorbing coefficient in the surroundings. The expression "α *rand*" is used to randomize movement, where "α" has been the randomized parameter and "*rand*" has been a function that produces randomized integers with uniformly distributed in the [0,1] range. The

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

value of fitness is thus produced at randomness and has a range of [0,1]. From Equation (2) it yields the distance across the 2-FFs "*i*" and "*j*" through using Cartesian-Distance:

$$d_{ij} = \left|\left|x_i - x_j\right|\right| = \sqrt{\sum_{k=1}^{d}\left(x_{ik} - x_{jk}\right)^2}$$

Eq→2

The "k^{th}" portion of the "i^{th}" firefly is represented by the symbol "x_{ik}". The "*TempMap*" is discovered by computing the distance among features matrix. To acquire a fresh fitness level, the accuracy of the classifier is used to produce the temporary DS sample-matrix "*TempMap*".

Fresh placement is allowed if the innovative fitness value "$F \in [0,1]$" is smaller than the existing fitness; the beginning "*Sfcurrent*" is changed to match the current "*Sftemp*" feature matrix. Conversely, a randomized number "*r*" is created in the range [0,1] if "$SFcurrent \geq SFtemp$ and $F \in [0,1]$". While "$r < F$" means that the new position is acceptable, and the existing feature matrix is modified appropriately.

(iii) Swapping, Inserting and Inverting

The traditional inserting-operator and inverting-operator are adjusted to construct the neighborhood of the second searching architecture. Every one of those operators starts with a randomly chosen feature matrix and DS matrix. A random selection of DS samples is then taken into consideration for the implementation of the operator. Throughout the instance of the swapping-operator, 2 random locations are selected, and also the locations of the 2 current DS sample-feature matrices are exchanged among each other and concerning all DS samples.

According to the inserting-operator, single feature and single location are selected randomly, and the chosen features have been placed in the designated location throughout the full DS matrix. According to the inverting-operator, the DS matrix's data center placements among 2 randomly picked places in the sequence are inversed. The result is ignored if "$r \geq F$" is not present. At the beginning of each iteration, the method picks a column and row at randomness from the "*Map*" DS sample-matrix and temporally assigns the cell to "0" if it isn't on the diagonal.

Simulating the global optimal is done using AA, a probability approach. Especially in the vast searching field, it would be a meta-heuristic optimization approach for global optimizing estimations. Whenever the searching area is discrete, is therefore commonly used. Beginning at "s_0", the searching proceeds unless and until "k_{max} *(steps)*" is reached or searching with the lowest possible energy consumption, known as "$eneg_{min}$ (percentage accuracy)" is found.

Neighbors are randomly chosen and returned using the "Call-Random (0,1)", which constantly returns a randomized number in the [0,1] range in this approach. Through an annealing schedule, the temperature "*Temp*" determines the temperature to employ, given that the proportion "*r*" of the time limit has been completed. The rate of cooling "*tempcool*", "i.e., $Temp \leftarrow temp_{cool} \times Temp$", updates the temperature, and the approach returns to the perturbation phase to create a fresh move.

Two conditions must be met for this to happen: (a) Either the temperature "*Temp*" doesn't get better over time or the temperature is equal to or lower than the "$Temp \leq Temp_{ground}$" annealing temperature, in other words, it's in the frozen state.

(iv) MR based parallelization of the AFO method

An MR computing concept is used in this part to prove the parallel processing of the AFO method by acquiring a vector of weight. Considering "*T*" being a training-set with "*m*" to have been the count of mapping processes that are included. MR divides "T" into "m" separate subsets, each with a different collection of instances. All subsets contain nearly an equal proportion of instances, and also the randomness of the "T" file ensures a sufficient proportion of the classes as a result of this sequential partitioning process. Every "T_i" is mapped using the FS procedure, though in this instance is dependent on AFO.

This means that every mapping operation produces a binary-vector "$fv_i = \{fv_{i1},\ldots,v_{iD}\}$", where "*D*" is the number of features, which identifies that the features are selected by the AFO method. To get a vector "*x*" as described in Equation (3), the reduction process takes the average among all binary-vectors. This vector "x_j" represents the percentage of FS systems that contain the feature "*j*" in their results. To generate the DS which will be required for later ML concepts, this vector will be the result of entire FS learning.

$$x_{ij} = \frac{1}{m}\sum_{i=1}^{m} fv_{ij}, j \in \{1,2,\ldots,D\}$$

Eq→3

Once the feature-vector "*x*" has been calculated, the purpose is to remove the less probable features from the input DS as quickly as possible. Further MR paradigms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

were created in an attempt to do this more efficiently. To begin with, a threshold "θ" is used to binarize the vector "*x*".

$$SF = \{sf_1, \ldots, sf_D\}$$

Eq→4

$$sf_j = \begin{cases} 1 \ \ if \ x_j \geq \theta \\ 0 \ \ otherwise \end{cases}$$

Eq→5

The "*SF*" vector indicates which features are to be used in the decreased DS. High thresholds may be used to restrict the range of features picked "(D' = $\Sigma^{D}_{j=1}SF_j$)" that can be selected, while lower thresholds enable a larger range of features can be picked.

Three threshold levels "θ = {0,0.6,0.75}" are employed in this research. The initial DS will be zero (0), and the FS will be "0.6", derived from these threshold values. Once this threshold value has been "0.75", it's deemed to be the most accurate one. For the DS reductions, the MR hypothesis is as follows: A fresh map is generated every time an instance is processed, and it only contains the features selected in SF. Using the reduction process, the generated instances are lastly combined to create a complete reduced DS.

3.2.2 MKSVM FOR CLASSIFICATION

Large amounts of information are too much for conventional classification techniques to manage. The AFO technique is used for FS, and afterward, MKSVM classification is used to achieve DR performance in this research. One of the most used kernels approaches in ML is the SVM. By calculating just the linear combinations of all sets of values in the feature set, kernel algorithms may work in a higher-dimensional, implicitly vector space even without calculating the coordinates of any values in that space. Explicitly calculating the coordinates is more expensive in terms of computing time than doing this process.

MKSVM is used to classify large DS sample data as the first step in the classifier's development process. If a system relies entirely on the dot-product among two vectors, this kernel technique of SVM offers a gateway from linearity to non linearity. Wherein a dot-product has been utilized, a kernel function replaces it when employing the kernel trick.

The class labels of "Positive (+1)" and "Negative (-1)" is denoted by the notation: "$\{(yi, sf)\}^{N}i=1$" with "$yi \in \{-1, +1\}$" and the notation "$sfi \in Rn$" for applying an MKSVM formulation to the labeled training data.

The "$k(sf_i, sf_j)$: $R^n \times R^n \rightarrow R$" has been the inner product of two "$\Phi(sf_i)$". "$\Phi(sf_j)$", for a kernel on data points. According to Equation (6), the technique discovers a hyperplane that optimally separates data in an unrealized, potentially higher-dimensional feature space.

$$\tau(w, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i$$

Eq→6

Subjecting towards "yi((w. sfi) + b) ≥ 1 – ξi and $\xi i \geq 0$", where "C > 0" becomes a balancing act among both regularization and boundary violation. To optimize in the dual formulation, maximize the following by Equation (7):

$$W(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j\, y_i y_j k(sf_i, sf_j)$$

Eq→7

Subjecting towards "$0 \leq \alpha i \leq C$" and "$\Sigma ij \ \alpha i \ yj = 0$". The "*(h(sf))*" is the sign of the decision function.

$$h(sf) = \sum_{l=1}^{m}\alpha_l\, y_l k(sf_i, sf_j) + b$$

Eq→8

Support-vector (SV) is used to indicate the features "sfj" and "$j \in \{1, 2, \ldots, m\}$" in an attempt to make their meanings more clear. To categorize a point with an MKSVM, "*m*" kernel computations are required, and all "*m*" of SV should always be kept. Classification of MKSVMs will most likely be accelerated as a result. The intersection-kernel is "$k(sf_i, sf_j)$" as provided in Equation (9) for feature-vectors "$sf \in R^n{}_+$":

$$k(sf_i, sf_j) = \sum_{i=1}^{n} min(sf(i), sf(j))$$

Eq→9

The classification is based on evaluating as in Equations (10) and (11):

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
Article Received: 05 May 2022 Revised: 30 May 2022 Accepted: 20 June 2022 Publication: 31 July 2022

$$h(x) = \sum_{j=1}^{m} \alpha_j y_j k(sf_i, sf_j) + b$$

Eq→10

$$\sum_{j=1}^{m} \alpha_j y_j \left(\sum_{i=1}^{n} min\left(sf_i, sf_j\right) \right) + b$$

Eq→11

Where "b" is the bias vector, "α_j" is the Lagrange multiplier. Through this, the higher dimensional DS is composed into smaller dimensional DS with a good accuracy level.

4. 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. Evaluation of the categorized DS throughout the whole DB is conducted using these input parameters, often referred to as actual content. The experiments are carried out by employing Java Programming Tool upon "Cardiotocogram-Tracing (CT)" DS from the "UCI-ML repository", which has been publicly accessible. This simulation has been carried out using a Windows-10 notebook with 8-GB of RAM.

The following is a brief description of the dataset: Mostly during the third quarter of pregnancy, most women are concerned about their health and well-being. The supply of oxygen to a fetus's heart rate is often compromised at this period. The heart rate of an unborn child may be traced using a CT. Fetal heartbeat and uterus contractions may be monitored with this device. The CT-DS is taken from the UCI-ML repository, which consists of "2126-occurrences" and "23-features" in total. Undergoes two main transformations: from "Uterine Contractions per second (UCs)" to "Fetal Movements per second (FMs)". There are a few more features that may be used to identify the fetal heartbeat. Table 1 summarizes the DS's most critical features.

Table 1: Most important DS features

LB	FHR baseline (beats per minute)
AC	Accelerations per second
DL	Light decelerations per second
DS	Severe decelerations per second
DP	Prolongued decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability

(i) 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 with the LDA-SVM and AFO-MKSVM methods. The entire DB when reduced in various dimensions ranging from smaller 10 to higher 90, the IR varies. The findings in Table 2 and Figure 4 revealed that AFO-MKSVM can achieve a high IR in smaller and higher dimensions, but LDA-SVM had a lower IR in this implementation.

Table 2: Numerical DR Comparison

DATA DIMENSIONS	LDA-SVM	AFO-MKSVM
10	98.23	99.35
30	97.32	98.45
50	96.23	97.32
70	95.12	96.47
90	94.53	95.21

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

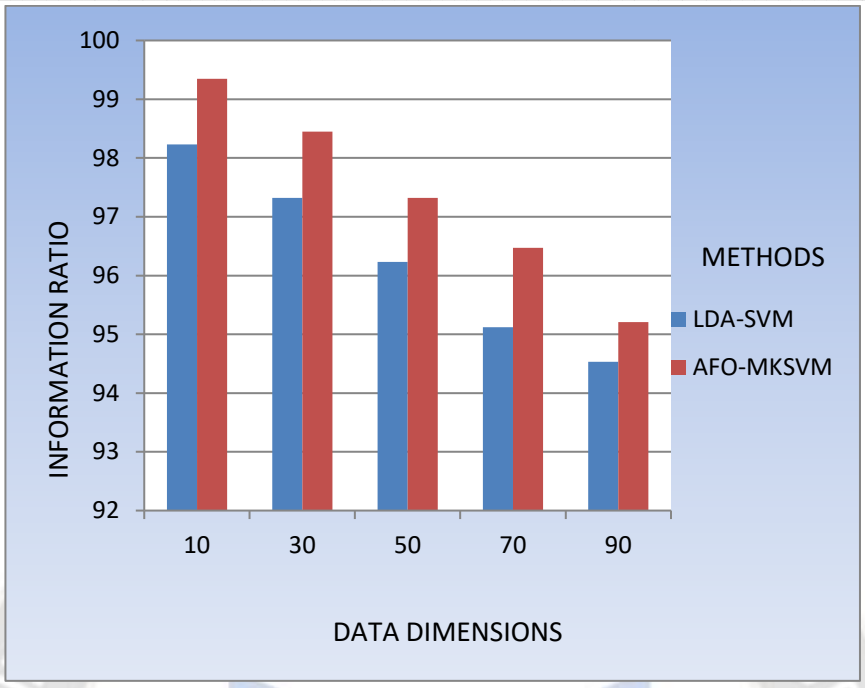

Figure 4: Graphical DR Comparison

(ii) 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 with extracting features by existing LDA-SVM and higher for the following selection of optimal features by proposed AFO-MKSVM. Which might be seen as a favorable DR outcome. This stage illustrates an essential fact: when DR is combined with selecting only optimal features on the DS, an accuracy rate of the classifier improves, and so this research effort implies that this phase is required. Table 3 and Figure 5 demonstrate the accuracy of the comparison of the classified results for AFO-MKSVM and LDA-SVM with various DS in the entire DB.

Table 3: Numerical Accuracy Comparison

DATASETS	LDA-SVM	AFO-MKSVM
DATASET 1	94.9	96.7
DATASET 2	96.3	98.3
DATASET 3	99.1	99.7
DATASET 4	98.4	99.2
DATASET 5	97.3	98.7

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
Article Received: 05 May 2022 Revised: 30 May 2022 Accepted: 20 June 2022 Publication: 31 July 2022

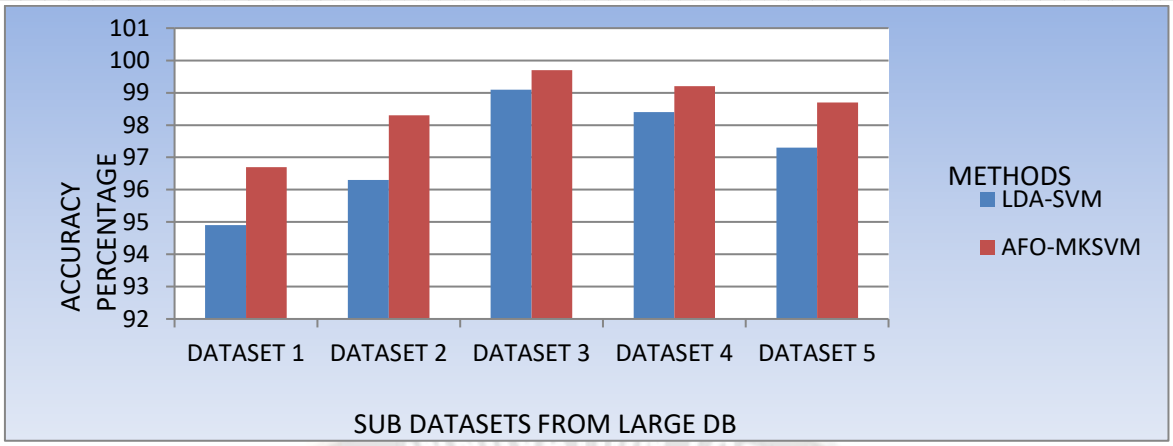

Figure 5: Graphical Accuracy Comparison

(iii) 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 evaluation on the categorized DS utilizing actual reality information. The cumulative recall of its categorized DS was lower with extracting features only by existing LDA-SVM and higher after selecting optimal features by proposed AFO-MKSVM. This also might be seen as a favorable DR outcome. Table 4 and Figure 6 demonstrate the categorization recall outcomes comparison for LDA-SVM and AFO-MKSVM with various DS in the total DB.

Table 4: Numerical Recall Comparison

DATASETS	LDA-SVM	AFO-MKSVM
DATASET 1	93.8	94.9
DATASET 2	95.2	96.4
DATASET 3	98.2	99.3
DATASET 4	97.3	98.4
DATASET 5	96.2	97.5

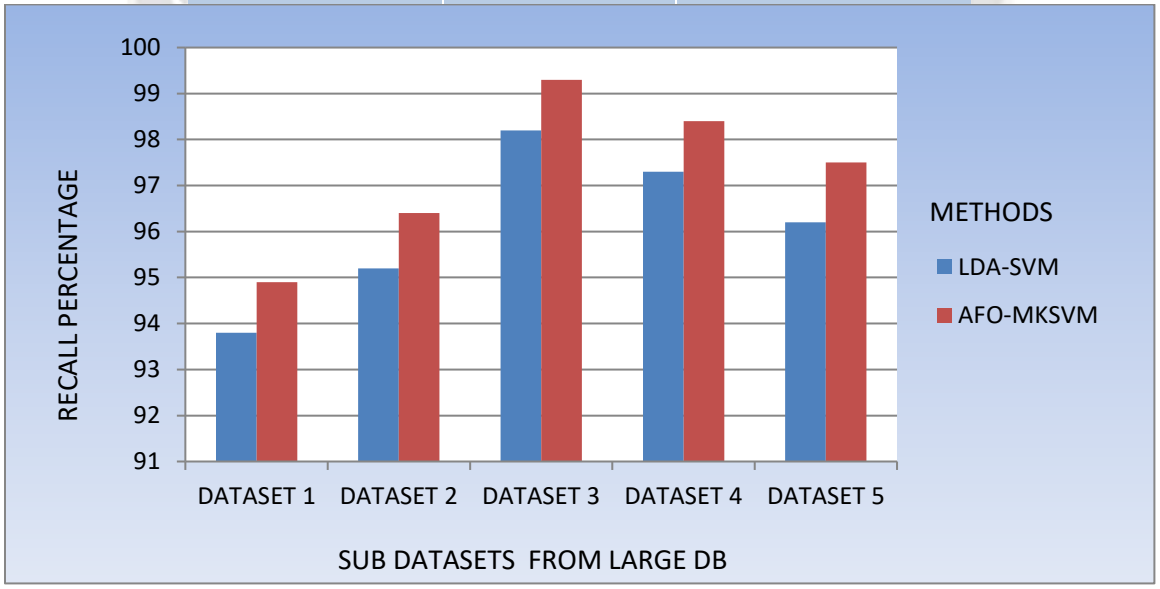

Figure 6: Graphical Recall Comparison

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v10i7.5595
***Article Received**: 05 May 2022 **Revised**: 30 May 2022 **Accepted**: 20 June 2022 **Publication**: 31 July 2022*

5. CONCLUSION

The BD's categorization capability may be vastly improved by including ML in the process. When applied to BD, the ML has shifted from conceptually learning to complicated cognitive learning and handling, and this trend is expected to continue. It's a critical step in enhancing and optimizing knowledge discovery. The primary purpose of ML is to improve system efficiency. The only way to do this is to keep learning new algorithms all the time. For knowledge acquisition to become more efficient, it must accumulate at a specific level in quantity. A novel AFO approach for FS that uses the MR design pattern to classify BD for DR is presented in this paper. The process begins with a mapping step wherein the source huge DS is broken down into blocks of samples that may be used as training data. The AFO approach is then used to choose the desirable features. The reducing phase integrates fragmented findings into a finalized vector of features, which would then be assessed using the MKSVM classification for achieving better DR. The developed AFO-MKSVM classification has progressively greater run time than the LDA-SVM classification, according to the results of the experiments. As a result, using MKSVM for BD classifications should lead to improved accuracy (i.e., lower errors). In the future, we try to implement some advanced bio inspiring approaches to enhance the DR with higher accuracy.

REFERENCES

[1]. Mishra D, Sharma S (2021) Performance analysis of dimensionality reduction techniques: a comprehensive review. Adv Mech Eng:639–651

[2]. Silva R, Melo-Pinto P (2021) A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries. Appl Soft Comput 113:107889

[3]. Zhou J, Gandomi A, Chen F, Holzinger A, Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 2021, 10, 593.

[4]. Shamsolmoali P, Kumar Jain D, Zareapoor M, Yang J, and Afshar Alam M. (2019) High-Dimensional Multimedia Classification Using Deep CNN and Extended Residual Units. Multimedia Tools and Applications, 78, 23867-23882.https://doi.org/10.1007/s11042-018-6146-7

[5]. Tău ţan A.M, Rossi A.C, de Francisco R, Ionescu B Dimensionality reduction for EEG-based sleep stage detection: Comparison of autoencoders, principal component analysis, and factor analysis. Biomed. Tech./Biomed. Eng. 2020, 66, 125–136.

[6]. Brodinová Š, Filzmoser P, Ortner T, Breiteneder C, and Rohm M. (2019) Robust and Sparse K-Means Clustering for High-Dimensional Data. Advances in Data Analysis and Classification, 13, 905-932. https://doi.org/10.1007/s11634-019-00356-9

[7]. Wang Xiang, Hu Xuegang. Overview on feature selection in high-dimensional and small-sample-size classification [J]. Journal of Computer Application. 2017, 37(9):2433-2438, 2448.

[8]. Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063

[9]. Rostami M et al (2021) Review of swarm intelligence-based feature selection methods. Eng Appl Artif Intell 100:104210

[10]. Zebari R et al (2020) A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends 1(2):56–70

[11]. Abualigah L, Gandomi A.H, Elaziz M.A, Al Hamad H, Omari M, Alshinwan M, Khasawneh A.M, Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering. Electronics 2021, 10, 101.

[12]. J. . Hermina, N. S. . Karpagam, P. . Deepika, D. S. . Jeslet, and D. Komarasamy, "A Novel Approach to Detect Social Distancing Among People in College Campus", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 153–158, May 2022.

[13]. A. Mohan, M. Ebrahimi, S. Lu, and A. Kotov, ''A NoSQL data model for scalable big data workflow execution,'' in Proc. IEEE Int. Congr. Big Data, BigData Congr., Oct. 2016, pp. 52–59, 2016, doi: 10.1109/BigData Congress.2016.15.

[14]. Mahapatra S, Chandola V, "Multi manifold learning from streaming data". In: 2017 IEEE international conference on big data (big data). IEEE; 2017, p. 716–25.

[15]. L. Reyes-Álvarez, M. D. M. Roldán-García, and J. F. Aldana-Montes,''Tool for materializing OWL ontologies in a column-oriented database,'' Softw., Pract. Exper., vol. 49, no. 1, pp. 100–119, Jan. 2019, doi: 10.1002/spe.2645.

[16]. S. Lai, X. Yuan, S.-F. Sun, J. K. Liu, Y. Liu, and D. Liu, ''GraphSE: An encrypted graph database for privacy-preserving social search,'' in Proc. ACM Asia Conf. Comput. Commun. Secur., Jul. 2019, pp. 41–54, DOI: 10.1145/3321705.3329803.

[17]. E. Damiani, B. Oliboni, E. Quintarelli, and L. Tanca, ''A graph-based meta-model for heterogeneous data management,'' Knowl. Inf. Syst., vol. 61, no. 1, pp. 107–136, Oct. 2019, DOI: 10.1007/s10115-018-1305-8.

[18]. Nouby M. Ghazaly, M. M. A. . (2022). A Review on Engine Fault Diagnosis through Vibration Analysis . International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 01–06. https://doi.org/10.17762/ijrmee.v9i2.364