Australian Journal of Basic and Applied Sciences

2020 April; 14(4): pages 1-14
DOI: 10.22587/ajbas.2020.14.4.1

Original paper AENSI Publications

Parallel SVM Based Classification Technique on big data: HPC center in Sudan

Iatimad Satti Abd Elkarim[1], Johnson Agbinya[2], Atika Hussein[3]

[1]Sudan University of Science & Technology, Faculty of Computer Science &Information Technology, computer science department, Box 407, Khartoum, Sudan. eatimadsatti@hotmail.com.
[2]School of Information Technology and Engineering, Melbourne Institute of Technology, box3000, Melbourne, Australia, jagbinya@mit.edu.au
[3]Sudan University of Science & Technology, Faculty of Computer Science &Information Technology, computer science department, Box 407, Khartoum, Sudan. atikahussien@gmail.com

Correspondence Author: Iatimad Mohamed Satti: 1Sudan University of Science & Technology, Faculty of Computer Science &Information Technology, computer science department, Box 407, Khartoum, Sudan.
E-mail:eatimadsatti@hotmail.com
Phone number:00249902664358

Received date: 20 January 2020 , Accepted date: 24 February 2020, Online date: 29 March 2020

Abstract

The fast growth of data in different data domains and a large amount of data generated by digital technologies is a significant challenge for standard data management systems for managing and processing big data. It requires parallel computing for machine learning methods. This study aims to deal with big data analysis, using parallel computing through the SVM algorithm and Parallel SVM based MapReduce in Hadoop cluster. Support Vector Machines are an excellent example of the commonly used methods for producing classification problems. It is a suitable classifier in the area of machine learning because of its generalization ability and expertise to classify big data accurately. However, the traditional SVM is not appropriate for huge datasets due to its high computational complexity.This paper is a study an amended form of SVM algorithms, and Parallel Support Vector Machine (PSVMs), and their applications in different big data fields. PSVM is used in the Hadoop cluster, which is running in the HPC center in Sudan, the article applies PSVM on realistic data.
These two models are implemented in four datasets for classification. Real water quality dataset from different water stations and the Ministry of Health in Sudan. The Adult dataset, the diabetes dataset, and the cover type dataset from UCI repository. The numerical experiment applying the PSVM is compared with SVM frameworks. The results showed that the applying of the parallel support vector machine gives the highest accuracy and has a good impact on reducing computation time. The performance is compared using time-consuming accuracy.

Keywords: Parallel Support Vector Machine, Support Vector Machine, Big Data, Hadoop, and Classification

INTRODUCTION

Knowledge is never abstract; it should be acknowledged by reality and applied research. With three dimensional, followed by the multidimensional prevalence of events, the experience will no longer accommodate a theoretical approach. Knowledge has to be involved in real-time and technological systems. Knowledge in technological researches should be accurate and specific. The features of Big Data have significantly appealed to many organizations, such as health care institutions, Ministries, researchers, academicians, etc. (Thanigaivasan, Vivekanandan *et al*., 2018).

There is a massive challenge to High-Performance Computer, and parallel processing tools for enhancing the performance of machine-learning algorithms in terms of time, especially for big data problems. The Support Vector Machines, as a supervised machine-learning technique, can take advantage of HPC. SVM is used as a widespread technique because of its excellent generalization performance on many real-life data. Big data faces many problems, especially in computational time and storage, and this is where we are. Four real datasets are applied to two models to cope with the new era of computational time problems.

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

Big Data manipulating has acquired an interest in the expansion of two critical fields of Information Technology, which are Big Data and High-Performance Computer processing. (Thanigaivasan, Vivekanandan *et al*., 2018).

There are many implementations of parallel support vector machines viewed in the related works, their benefits, and their limitations. Many factors impacting the performance of implementations, like algorithm optimization, the size, and dimensions of a problem, kernel functions, parallel programming stack flow, and hardware architecture. It is needed to balance between execution time and classification accuracy. Multiple classification algorithms are suggested for big data. Many of these algorithms have restrictions and weaknesses, such as weak performance in a large dataset, low run-time execution when the training set is large, and high execution cost. To cover these weaknesses, multiple researchers use classification algorithms (Pakize *et al*., 2014).

This paper is a study of many applications and performance of PSVMs on large scale data sets and parallel computing using different algorithms. Machine-learning algorithms have different sections, supervised, unsupervised, and semi-supervised learning. SVM is an algorithm which is a supervised learning algorithm, that has been known for its efficient use in many problems because of its high-performance classification ability. The SVM is a great machine learning technique that has shown the perfect result in different application areas such as regression, power system, hydrology, power system, and medical fields. SVM can easily be used to reduce the generalization error only by maximizing the margin. (Rezvani, Salim et al.,2019).

The challenge due the big data is the development regarding execution time, efficiency, scalability, and memory size (Tavara and Shirin, 2019). The essential goal of this paper is to study and implement the support vector machine algorithm and parallel support vector machine algorithm in big data classification. They are implemented in four different data sets. The results show that the PSVM gives the best efficiency than the sequential SVM algorithm.

1. MATERIALS AND METHODS

2.

1.1. SUPPORT VECTOR MACHINE (SVM)

A support vector machine is a classification algorithm that is a supervised machine learning algorithm. to provide the mapping between the feature space and the target labels (He, Taiping et al., 2019). Support Vector Machine provides for classifying labeled data and designing a model for a classification job. Fig1 shows the SVM classifier. The Support Vector Machine method was also used in multi-class classification problems, which is shown in Fig2. The parameters *C* and γ defines the performance quality of the classification by SVM. The parameter γ represents the impact of a single training example. Low values mean wide, and high values mean close. If *C* is small, it smoothens the decision surface, and if *C* is high it classifies all training examples correctly (Jessica et al., 2019). Equation (1) is a constrained optimization problem that gives the standard SVM formulation.

$$\min imize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (1)$$

$$subject \quad to \quad y_i\,(w.\phi(w_i)-b) \geq 1-\xi_i \qquad \xi_i \geq 0$$

The maximum margin with hyperplane f(x) = $w\cdot\varphi(x) + b$ is determined by explaining and obtaining the parameters *w* and *b*. ξi are slack variables, $\xi_i = max\left(0,1-Y_i\left(w\varphi\left(xi\right)-b\right)\right)$ are used to expand the SVMs where datasets are not linearly separable (Schlag et al., 2019). The curved quadratic programming problems in equation (2) is solved by optimizing the dual function.

$$\max_{\alpha} G(\alpha) = \sum_{i=1}^{N}\alpha_i y_i - \frac{1}{2}\sum_{i.j}\alpha_i\alpha_j K(x_i.x_j). \qquad (2)$$

$$subject \quad to \begin{cases} \sum_i \alpha_i \\ A_i \leq \alpha_i \leq B_i \\ A_i = \min(0, cy_i) \\ B_i = \max(0, cy_i) \end{cases}$$

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

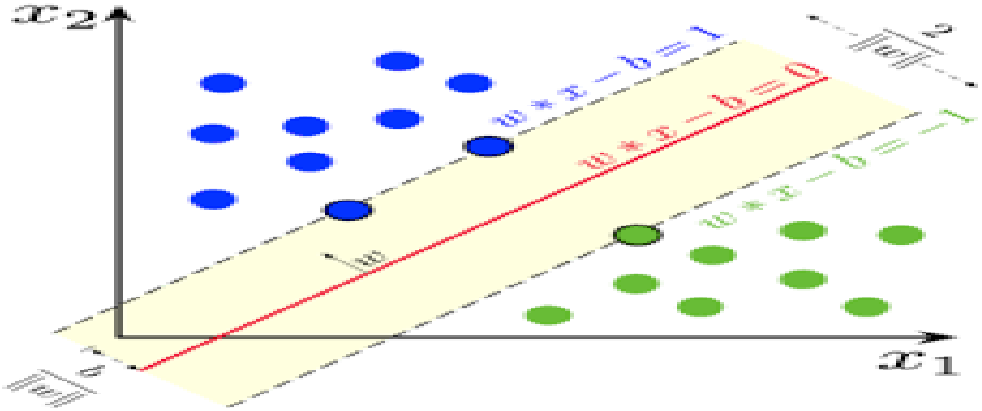

Fig1: Support vector machine classifier

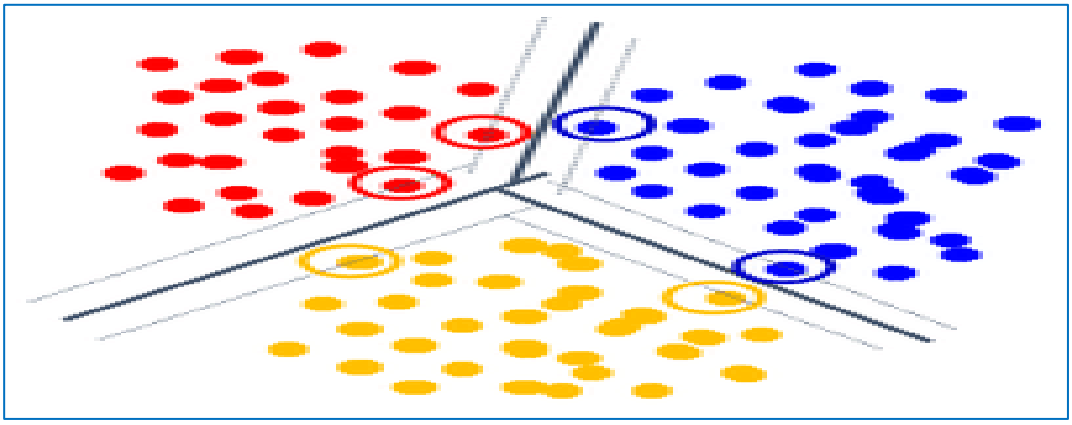

Fig2: Multi-class classification

SVM is taken as the most useful classification and regression model. The execution cost of SVM is balancing to the number of training data squared. The standard SVM model is challenging for analyzing big data problems. Parallel SVM can develop computation speed significantly. The parallel SVM based on twister can decrease the computation time, but it does not mean that it is sufficient to partition the dataset into many parts (Sun et al., 2012). To select the suitable kernel function is an essential job. This kernel function must be meet the Mercer condition. It represents a symmetric positive function on a square as a sum of a convergent sequence of product functions. Many kernel functions models have been developed. Commonly used kernel functions are included in Table1 (Tavara and Shirin, 2019).

Table1: Some examples of Kernel Functions

Kernel Function	Inner Product	Kernel Type
Linear kernel	$K(x_i.x_j) = x_iTx_j$	Linear
Gaussian/Radial-Basis Function (RBF)	$K(x_i.x_j) = \exp(-\gamma\|x_i - x_j\|^2)^d, \gamma > 0$	Non-linear
Polynomial	$K(x_i.x_j) = (\gamma x_iTx_j + r)^d, \gamma > 0$	Non-linear
Sigmoid or Laplacian	$K(x_i.x_j) = \exp(-\gamma\|x_i - x_j\|^2)^d, \gamma > 0$, Here, $\gamma, r, and\ d$ are kernel parameters.	Non-linear

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

1.2. Parallel Support Vector Machine (PSVM)

The Parallel computing of SVMs has become needed for developing the effectiveness of SVMs for big data and has already promised for enhancing results of large data problems (Tavara and Shirin, 2019). In a parallel Support Vector Machine model, the training samples obtained through partial SVMs. It is driven the partial solutions towards the global optimum solutions. Each subSVM is used as a filter, the output support vectors of the first subSVM are used as the input of the next subSVMs. Through the PSVM model, big data problems can be partitioned into small problems, the subSVM are combined into one SVM hierarchically. (Sun et al., 2012). Fig3 shows the training flow of parallel SVM.

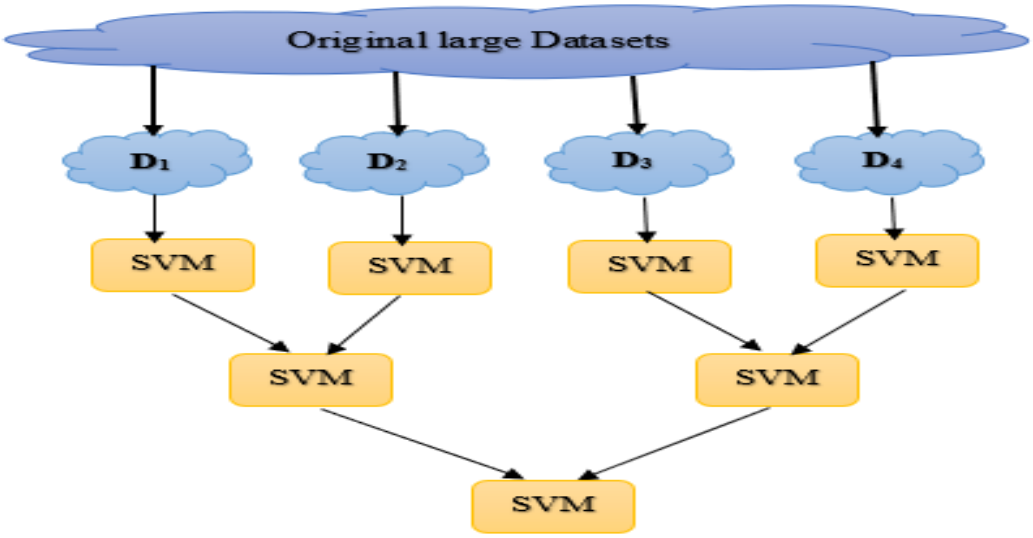

Fig3: The training flow of parallel SVM

In Fig3, the structure of PSVM shows that the new input set of SVs of the next SVM is collected from the output of the previous level. This process is done until only one set of vectors is left. A single SVM does not have to deal with all training sets. The training set of the subproblem is smaller than the whole problem, this means the last level is to handle a few vectors than the actual number support vectors. It becomes optimum when the filters in the first layers are useful in extracting the support vectors. (Sun et al., 2012).

A method that divides the problem into smaller tasks is proposed in reference (Zhao, Hai, et al., 2011). In each task, there are selected parts of α called the working set used to be optimized, and the rest α constant. This process repeats until optimal global conditions are satisfied. The α, y, and Q can be written as in equation (3), where B is the working set with n variables, N is the non-working set with $(l - n)$ variables.

$$\alpha = \begin{vmatrix} \alpha_B \\ \alpha_N \end{vmatrix}, y = \begin{vmatrix} y_B \\ y_N \end{vmatrix}, Q = \begin{vmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{vmatrix} \quad (3)$$

the small task written as in equation (4)

$$\min \frac{1}{2}\alpha_{\mathbf{B}}^T Q_{BB\alpha B} - \alpha_{\mathbf{B}}^T (1 - Q_{BN\alpha N} + \frac{1}{2}\alpha_{\mathbf{B}}^T Q_{NN\alpha N} - \alpha_{\mathbf{B}}^T \quad (4)$$

Subject to

$$\alpha_{ByB}^T + \alpha_{NyN}^T = 0$$

$$\mathbf{O} \leq \alpha_B \leq C$$

1.3. Hadoop Framework

A Hadoop is one of the open-source frameworks that help in distributed applications. The user application can work with several computer nodes and petabytes of data. The essential Hadoop feature is to partition the data into multiple machines and executed them in a parallel way. The Hadoop cluster can be installed using specialty hardware to process large scale data efficiently. The two components of Hadoop are Hadoop Distributed File System (HDFS) and the MapReduce distributed programming model (Priyadarshini, Anushree, 2015).

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

1.4. MapReduce

MapReduce is a method used to run parallel applications for big datasets. MapReduce is taken from the map and reduce function, which is combined from functional programming. MapReduce uses key/value pair data type in a Map and Reduce tasks. Fig4 shows the MapReduce system. MapReduce method distributes the dataset over cloud computing systems. The proposed model of MapReduce based on parallel distribution SVM is used for binary classification. Every data set has to find the binary classifier function at its node. The algorithm collects all Support Vectors from all nodes and save a global one. (Çatak et al., 2016).

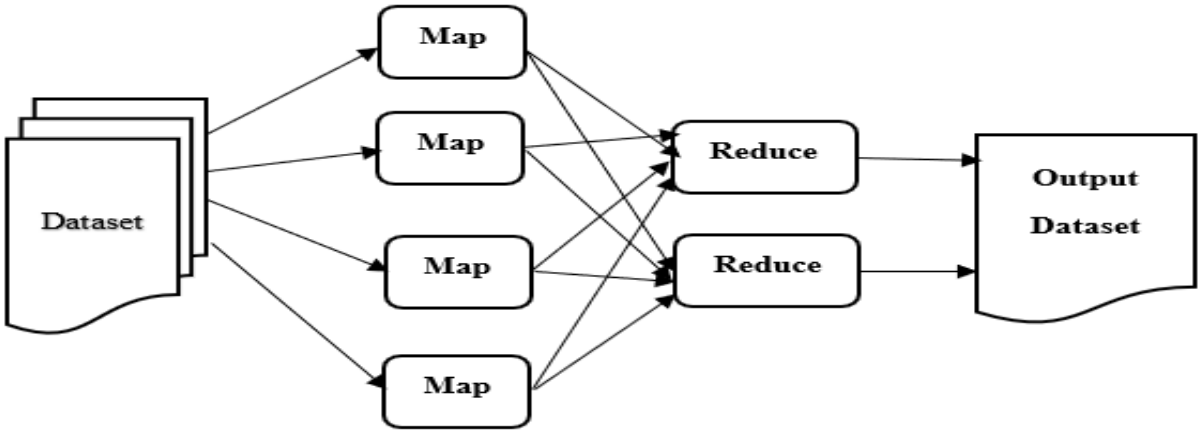

Fig4: Overview of MapReduce system.

1.5. Hadoop Distributed File System (HDFS)

The Hadoop framework used the HDFS to distribute the file systems. Hadoop Distributed File System is a movable filesystem written in Java. It stores all its metadata to its devoted server, known as NameNode, also called a master node. The user used the NameNode to communicates and perform any input and output to the Hadoop cluster. (Priyadarshini, Anushree, 2015). The NameNode daemon is running in HDFS by the NameNode. The JobTracker controls the process and organizing the execution of MapReduce functions. JobTracker is running by the job submission node, which contacts the client to execute a MapReduce function. These services run on four separate machines in smaller clusters. The size of a Hadoop cluster consists of slave nodes that run both a TaskTracker, which is efficient for actually running user code and a DataNode, for serving HDFS data.

2. Review of parallel Support Vector Machines

A parallel Support Vector Machine with MapReduce (PSMR) method for email classification is suggested in (Jadhav, Kishor, 2013). The proposed algorithm gives the best performance than the Naive Bayes (NB) classifier and SVM model. First, SVM was used to classify the folder of the email based on a specific field of data on the email. Next, SVM was used on each email applied as a separate bag of words. Naive Bayes is a simple method for structuring classifiers, which attribute the instances of the class with the highest posterior probability. A limitation of Naive Bayes is the assumption of independent predictors, a complicated condition to be fulfilled in real life (Jadhav, Kishor, 2013).

Cloud computing is developing as a new computational model is Proposed. Hadoop-MapReduce is a great computation model for treating big data on distributed hardware clusters, such as clouds. In all Hadoop applications, the tasks are scheduled in a default FIFO order with help for other priority schedulers. (Rao, Thirumala et al., 2012).

A Hadoop MapReduce models are observed in all Standalone, distributed, and Fully distributed form. This Hadoop cluster contains four nodes, and one Master called NameNode and three Slaves called DataNode. (Shvachko, Konstantin et al., 2010).
Many parallel applications for support vector machines are studied. But still, there is no clear plan for all application situation. Many factors are impacting the efficiency of implementations, such as optimization methods, the size, and dimension of the problem, kernel function, parallel programming stack, and hardware structure. The user is one who was balanced between execution time and classification efficiency. The Parallel model of SVMs is promising results for enhancing big data problems and increase performance percentage. (Tavara and Shirin, 2019).

In another application by (Liu, Peng, et al., 2019) a parallel naive Bayes Algorithm (PNBA) is proposed and implemented to solve the problem of the Chinese text data, because this data is increasing on the internet, making it challenging to classify data by using spark platform for big data. The authors used parallel computing in the entire training and prediction of naive Bayes

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

classifier using resilient distributed datasets (RDD). PNBA was implemented in Hadoop. The result was compared with the SPARK. It was found that the Spark PNBA gives more accuracy than the Hadoop PNBA, especially in terms of speed and scalability (Liu, Peng, et al., 2019).

The suitable hardware implementation of cascade support vector machines (CSVM) is enhanced to handle data of two classes problems efficiently and speedup execution time than single SVM classifiers. Such as image classification problems. However, SVM classification is a computationally challenging job, and the current hardware designs for SVMs observe only unified classifiers. This model is used to design low-cost PSVM processors and intelligent embedded systems for online real-time classification applications to allow SVM designs to solve big data problems. (Kyrkou, Christos, et al., 2015).

The speedup of CSVMs through a compound processing hardware configuration is enhanced for the CSVM classification. Two methods follow it, the first one to decrease the needed of hardware resources for its implementation and another method to develop the classification speed by employing cascade information to canceled datasets samples. (Kyrkou, Christos, et al., 2015). A parallel method of support vector machine, called kSVM, is proposed for the useful nonlinear applications of big datasets. The k- means clustering model is used to partitions the data into k clusters and uses a nonlinear SVM in each cluster to classify the data in a parallel way on multicore machines. The kSVM model is speedup the execution time than the standard SVM in the nonlinear analysis of big datasets (Do, Thanh, et al., 2015). A Coarse Grained Parallel Genetic Algorithm (CGPGA) is employed to enhance the optimization of the feature subset and attributes for SVM concurrently. The transport system of the CGPGA model and the topology help to find optimal feature subset and attributes for SVM in a short time. A new function that merges the classification accuracy is achieved from the bootstrap method, a quantity of features, and some support vectors are proposed to manage the search of CGPGA to the way of optimal generalization error. However, two problems must be marked for SVM, feature choice, and parameter optimization. (Chen, Zhi, et al., 2016).

The topology and movement system of CGPGA allows the search for the solution space with different search approaches in a parallel fashion, thereby producing powerful search capability and high performance. The method not only enhances the SVMs' model parameters but also gets the feature subset. The ratio of SVs in the design presented by the technique was preserved at a low level. So, the classification is much faster on the unseen new model's applications, where the classification has done at high speed (Chen, Zhi, et al., 2016). A Distribution Preserving of the kernel SVM (DiP-SVM) model is introduced. The first and second orders statistic of the whole dataset is held in each one of the partitions separately. The DiP-SVM reaches the smallest loss in classification efficiency, among other distributed SVM methods on many benchmark datasets (Singh, Dinesh et al., 2016).

The purpose of a training SVM model for big datasets has been performed by splitting the dataset into controllable sized, then training a sequential SVM on each of these partitions separately to get local SVs. The classified SVMs have confirmed to be much faster than sequential SVMs on big datasets; however, this process usually leads to weak classification accuracy. The global SVs have not been chosen as local SVs in their separate partitions. (Singh, Dinesh et al., 2016).

The SVM method with Hadoop based on MapReduce for English document classification in the Cloudera parallel network system environment is proposed. The new model is examined on the English data set, and it performs 63.7% accuracy of sentiment classification. This model can be applied to many other languages, although these data sets are small. However, if the new model can be applied to the big data set with millions of English documents, it gives less performance. The use of the new model showed that the average time of the classification of the SVM algorithm in the sequential environment is higher than the average time of the classification of the SVM in the Cloudera parallel network environment (Phu, Ngoc et al., 2017**)**.

A parallel computing framework is used to accelerate the SVM-based classification. The characteristics of multi-threads and powerful parallel processing capability in have been obtained by graphics processing unit (GPU). The general-purpose computing with GPU (GPGPU is developed. It is a new area due to the highly parallel nature of GPU. With this GPU, parallel computing can be produced with low cost and less power consumption. The given GPU has achieved maximum speed up with high accuracy. This speedup can be increased for a given number of training samples by using GPUs producing more computation ability. This approach can be expanded for multiclass classification by using parallelism related to both CPUs and GPUs. This method can be more helpful with more complex datasets (Kshirsagar, et al., 2018).

A projection-SVM is proposed. A distributed application of kernel SVM for big datasets was implemented using subspace partitioning. A decision tree is built on the projection of data along the direction of the maximum variety to get smaller partitions of the dataset. On each partition, a kernel SVM is trained independently over a cluster, thereby reducing the overall training time and reducing the prediction time significantly (Singh et al., 2018).

The distributed SVM is trained in the model much faster, and it requires a shorter time in prediction for new data points. The dominant eigenvector and decision tree for the splitting of the dataset are less expensive computation costs when it compares with the kernel k-means method with complexity are proposed. (Hsieh, Cho, et al., 2014) (You, Demmel et al., 2017). So, the proposed model is achieved excellent classification performance with a small change efficiency. A Parallel Support Vector Machine for big data analysis is adapted due to its limitation in handling big data. The study is taken from the heart Disease dataset, and the performance of the classification method is compared with other frameworks, it found that the PSVM exceeds different

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

algorithms. In the case of Big Data, SVM is specified to suffer from slow processing time. Hence, Parallel SVM based classification is preferred to classify the big dataset. It has enormously reduced the execution time and classifies the data accurately (Vivekanandan, Swathi, et al., 2018).

A newly mixed solver of parallel and approx SVM for Big Data classification using the extended model of the SVMs is proposed. This compound is called Parallel Support Vector Machines (PSVM). The main disadvantage of a PSVM model is that the feature can be deleted over time, so the accuracy is decreased. To solve this problem, they used an approximates SVM model with the Radial Basis Function (RBF) kernel, which is called the Approx SVM. This new model helped to overcome two main problems, first, the inability to handle big datasets. Second, the exchange of attributes numbers over time. The PSVM has the advantage of decreasing the execution time in the classification model. So, the researchers in this paper had achieved excellent results in terms of efficiency compared with the regular SVM. The parallel approx SVM considerably reduced the time in building a new model for newly generated datasets over time. (Ksiaâ, Walid, et al., 2018).

A new parallel method and classification are proposed, which consists of the preprocessing of data, data selection, or feature extraction. The feature selection methods have been analyzed for the removal of datasets. These models are the SVM with recursive feature elimination (SVM-RFE), minimum redundancy maximum relevance (MRMR), principal component analysis (PCA), successive feature selection (SFS), and independent component analysis (ICA) (Sadasivam, Sudha, et al., 2018).

The mixture of general-purpose graphics processing unit (GPGPU) computing and MapReduce method on the Hadoop framework is introduced to deal with computational complexity and the big datasets. The experiments show that the development of time accuracy in feature extraction and classification performance is excellent. The parallel model of the proposed methods using CPU+GPU clusters enhances the time efficiency for feature extraction. By using Hadoop clusters, the training time of an SVM is decreased. In SVM-RFE, the parallelization is achieved only in the SVM training phase. The process of feature elimination is recursive, so it cannot be parallelized. The parallelism is produced by splitting data into blocks because the MapReduce processes require data to be independent. (Sadasivam, Sudha, et al., 2018).

A robust method to define kinship similarities between a given pair of facial images using feature descriptors to train the SVM classifier is proposed. The feature descriptors are used to extract the salient facial features. These extracted facial features are then combined to create a new high-dimensional feature vector. SVM trains these high-dimensional feature vectors to classify facial images based on feature similarities. The KinFaceW-I dataset is used to confirm the Kinship verification accuracy. A positive kinship pair is matched the real or own parent-child pair. While negative kinship pair is matched the pair of one's parent with another's child (Goyal et al., 2019).

A Fuzzy Twin Support Vector Machine (FTSVM) model for binary classification problems is proposed, which merges the idea of intuitionistic fuzzy number with TSVM. A sufficient fuzzy group is employed to lessen the noise created by the inputs, which is a useful machine learning method that can overcome the negative impact of noise and outliers in tackling data classification problems. Linear and nonlinear functions are used to formulate two nonparallel hyperplanes. An FTSVM not only reduces the influence of noises, but it also separates it from the support vectors. Further, this update can minimize a new formulated structural risk and enhance the classification accuracy. The result shows that an IFTSVM can produce promising results as compared with the original support vector machine, fuzzy support vector machine. However, it is sensitive to C, if it is not appropriately chosen, the IFTSVM produces inferior results. Our future work is focused on enhancing the structure of the IFTSVM to solve the imbalance classification problems. (Rezvani, Wang, et al., 2019).

3. METHODOLOGY

Apache Hadoop is a popular parallel processing platform for big data. Many data mining algorithms are relocating towards Hadoop. In this paper, the SVM and PSVM algorithms are studied. Once the problems are identified as big data problems, the Hadoop platform can be obtained. The performance of PSVM algorithms is increased on the Hadoop, and it has attracted more data mining processes to be moved to the Hadoop cluster platform (Nandakumar and Yambem, 2014).

The classification is one of the data mining techniques that is used to classify and organize the unstructured data into the structured class. In this paper, two models are designed and compared. The first model is a single SVM trained using different kernels implementation. The second one is the PSVM model based on the Hadoop framework. At last, the results of both models are compared. The structure of these models is shown in Fig5.

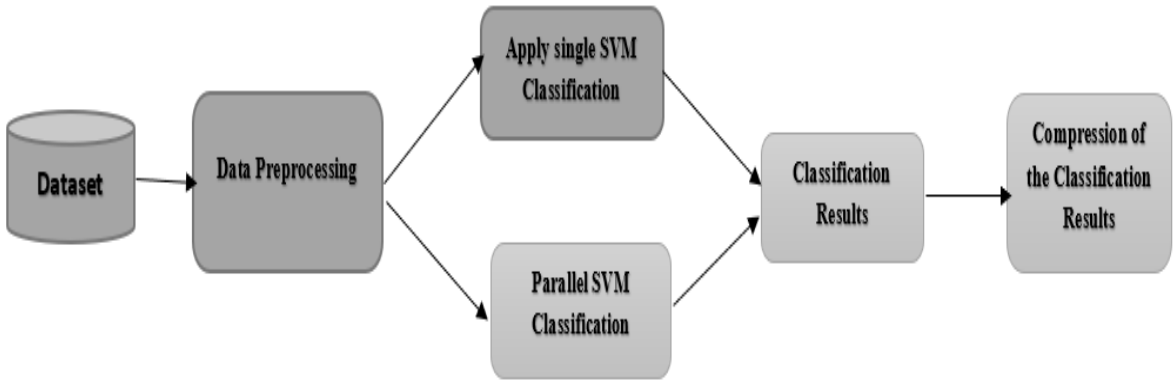

FIG5:

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

METHODOLOGY FRAMEWORK.

3.1. Data preprocessing

The necessary step in the classification task is data preprocessing, especially when the values of the attributes have varied in ranges and distributions (Hofmann and Klinkenberg, 2014). Handling big data is not easy and needs to work in an appropriate environment. Four datasets are used in this paper varied in sample size, and dimension. To improve the performance, hence, preprocessing was required. Missing values can be removed or replaced by the minimum, maximum, zero, or average cost of that attribute. The type of attribute is changed from non-numeric type to a numeric type because the SVM does not deal with non-numeric data.

3.2 Hadoop Implementation with PSVM

The essential characteristics of the Hadoop framework are partitioning the data into thousands of machines and executing it in a parallel manner. The framework in Fig6 shows the flow steps of the classification task of the PSVM implementation on the RapidMiner tool using Radoop extension on the Hadoop cluster. The organization of these steps as follows:

- Upload dataset on HDFS in Hadoop cluster.
- Retrieve the data from HDFS.
- Preprocess the datasets
- Train the PSVM
- Classification result.

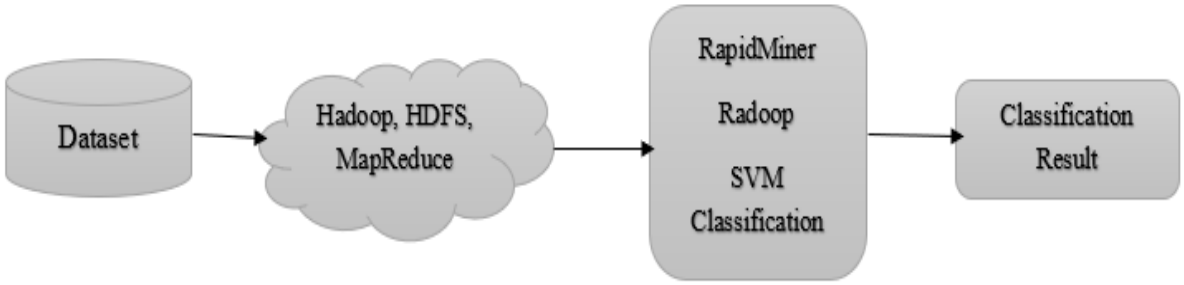

Fig6: PSVM with Hadoop cluster using Radoop extinction.

3.2. Parallel SVM Based Classification

Big data environment has been simulated by importing the four data sets into RapidMiner (Chisholm and Andrew, 2013). RapidMiner platform is data mining software tool produced for maintaining the machine learning process, including data preprocessing, model effectiveness, validation, and optimization. The dataset is grouped horizontally by creating 4 CPU core nodes in the HPC cluster in Sudan. The parallel implementation is performed using PSVM, Fig7 shows the parallel execution. All the simulations works are done using the "Intel Core i7, 2.90 GHz" system,8GB memory. The produced results are combined and represented.

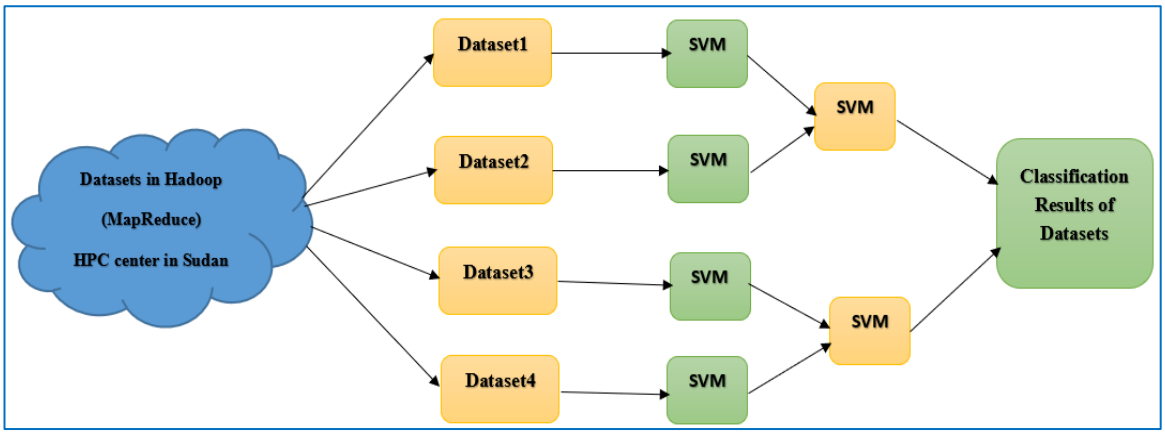

Fig7: Parallel SVM Based classification algorithms.

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

4. RESULTS AND STATISTICAL ANALYSIS

This section is partitioned into two parts. The first part presents the implementation of a single SVM. The second part presents the implementation of PSVM in the Hadoop cluster. Then the results were compared. The statistical analysis is done by using Excel 2016 for analyzing the results. The accuracy is used to measure the performance of the models.

4.1 The implementation of SVM

The four datasets are implemented using the SVM algorithm. There are three types of kernels used., the Radial kernel has a higher accuracy for all datasets. The parameters C, gamma, and epsilon have a fixed value, which are 0.0, 1.0, 0.01, respectively, as shown in Table2, Fig8, Fig9 the execution time increase as the data size increase. The four datasets shown are imported to the RapidMiner platform as excel or CSV files. Fig10 shows the scatter before the binary classification of an adult dataset, while Fig11shows the scatter after binary classification of an adult dataset.

Table2. The accuracy of four datasets with deferent kernel types

Kernel type	Polynomial	Dot	Radial	Execution Time (in sec)
Water quality	69.33%	69.11%	69.79%	56m 30s
Diabetes	67.23%	57.01%	96.40%	4h,54m,60s
Adult	78.01%	76.66%	93.40%	1 h and 40m
cover type	69.90%	75.08%	74.52%	1 days and 22 h

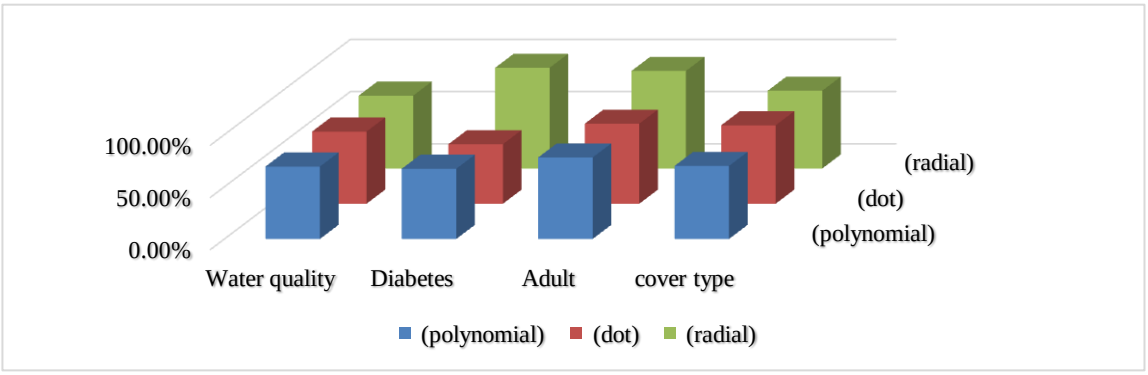

Fig8: Results accuracy of SVM

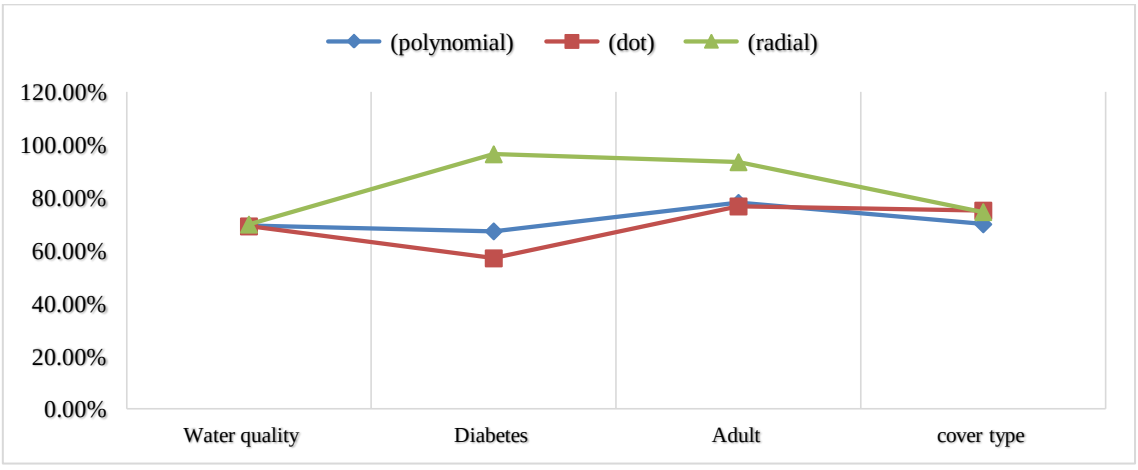

Fig9: Results accuracy curve of SVM

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

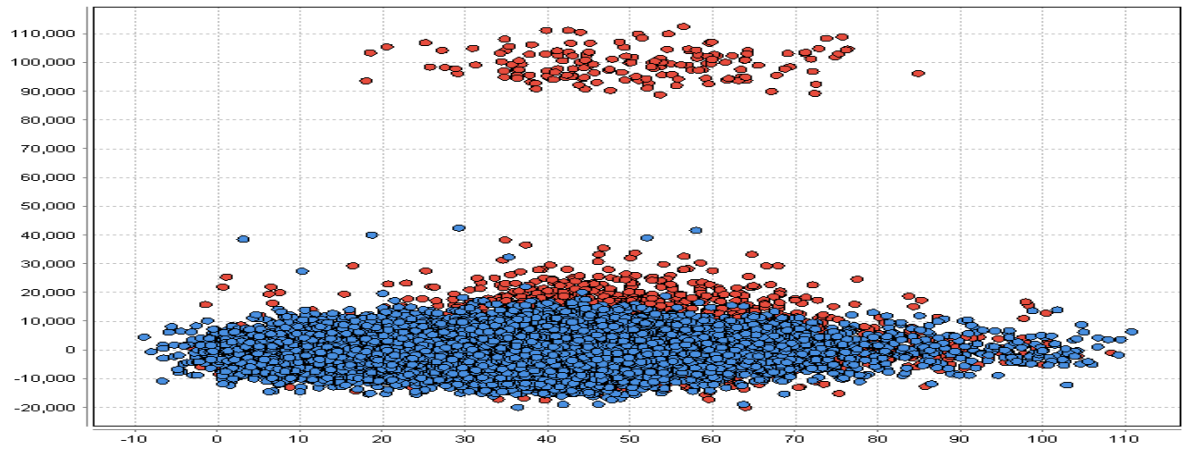

Fig10: Adult data before classification.

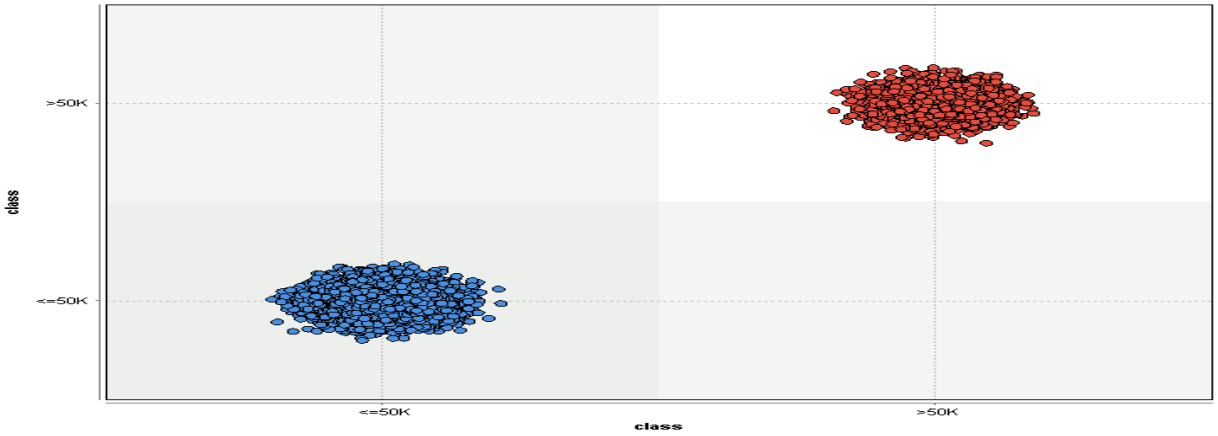

Fig11: Adult data after classification.

4.2 The implementation of PSVM on Hadoop Cluster

The experiments were created using the Hadoop structure because one of its central components is MapReduce has the same characteristics as the PSVM strategy. The Hadoop software has been set up in HPC in Sudan using Hadoop cluster hardware and software configuration, which are described in Table 3. The Hadoop configuration is done in RapidMiner using a mange Radoop connection setting. The efficiency of the Parallel SVM is done as follows: The SVM algorithm was implemented for classification measuring on a MapReduce Hadoop cluster. It is provided in RapidMiner parallelized and configured as a MapReduce task. The arrangement of the Hadoop cluster with the resources is described in Table 3.

Table3.Hadoop cluster configuration resources.

Hardware Resources		
	CPU	**RAM**
Node 1, 2, 3, & 4	Intel Core i5	8 GB
Client	Intel Core i7(64-bit OS)	8 GB
Software Environment		
SVM	RapidMiner 9.5	
OS	Ubuntu 16.04	
Hadoop	Apache Hadoop 2.2+	

The experiment is carried out by using the RBF kernel function, parameter C=1, and gamma =0.01. An investigation is carried out by four nodes on the Hadoop cluster in the HPC center in Sudan. The results of four datasets are shown in Table 4. Fig12 (a) and (b) shows the scatter plot of the cover type dataset before classification is done, while Fig13 shows the scatter plot of cover type dataset after classification.

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

Table 4. Hadoop Cluster Results

Dataset	Accuracy	Computation Time (in sec)	Number of nodes
Water quality	90.05%	15.05	4
adult	100%	530.74	4
Diabetes	95.47%	630.74	4
Cover type	100%	690.91	4

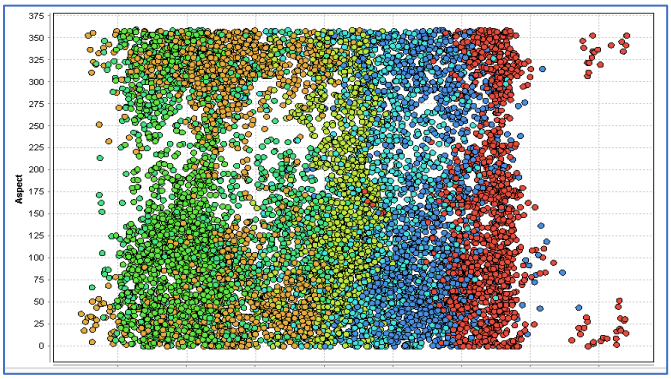

Fig12:(a)

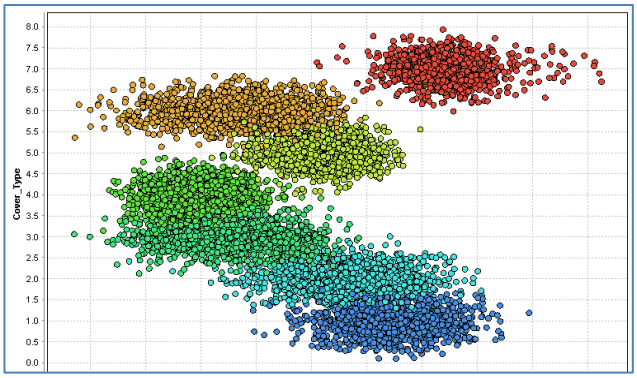

Fig12:(b)

Fig12: (a)Cover type dataset before classification. (b) after preprocessing

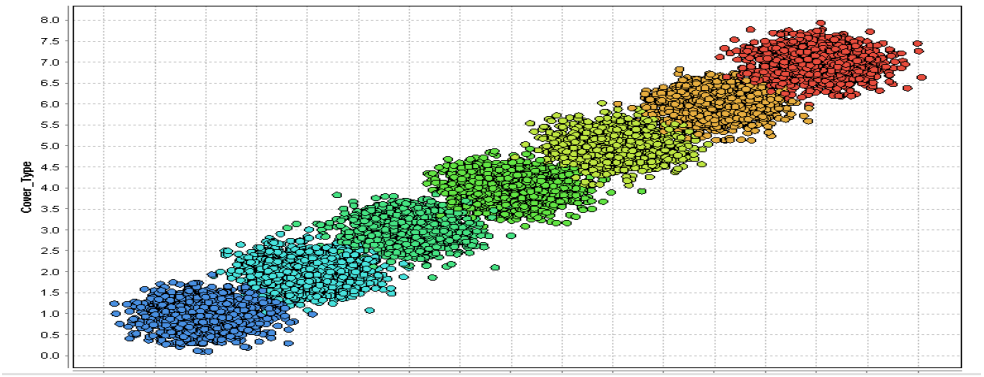

Fig13: Cover type dataset after classification

4.3 Comparison of Both Models

- By using a single SVM model, it is found that when the number of examples increases, the execution time also increases, leading to weak performance. So, PSVM is used to reduce and split the number of these instances producing better accuracy and time-consuming.
- Using regular SVM is challenging and complex to work with big data set. The MapReduce based parallel SVM works efficiently on large datasets as compared with the sequential SVM. The advantage of using MapReduce based SVM is the core components of the Hadoop framework HDFS and MapReduce distributed programming model provides data awareness between the NameNode and DataNode. The number of nodes that are used on the Hadoop cluster is four. It is worked in a parallel manner.
- The RBF shows the best result over the other two kernels with the best accuracy. For this reason, the compression between the two models is done in terms of the RBF. Table5, Fig14, Fig15 show the compression results.
-

Table 5: Accuracy comparison of SVM and PSVM.

Datasets	Water quality	Adult	Diabetes	Cover type
SVM (using RBF)	69.79%	94.29%	96.40%	74.52%
PSVM (using RBF)	90.05%	100%	100%	95.47%

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

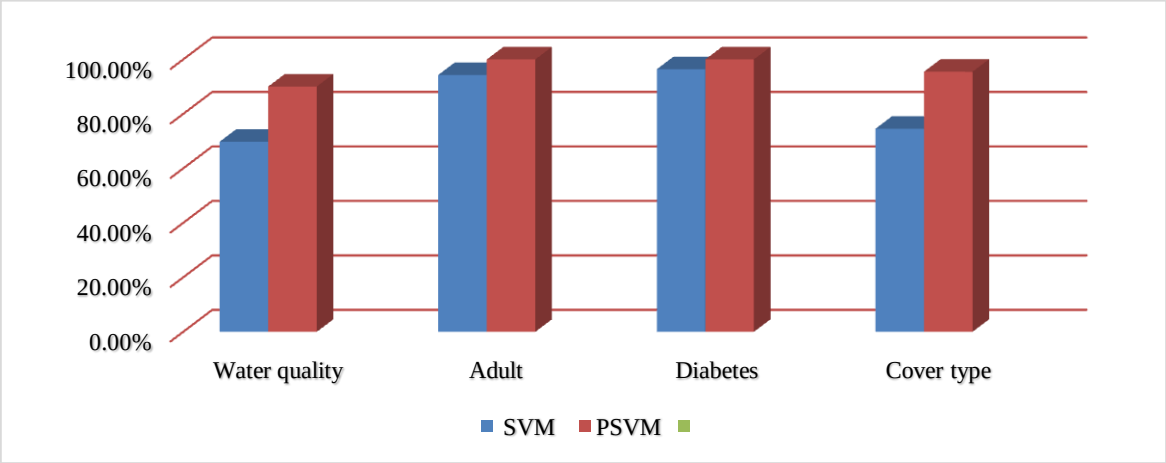

Fig14: Results comparison accuracy of PSVM.

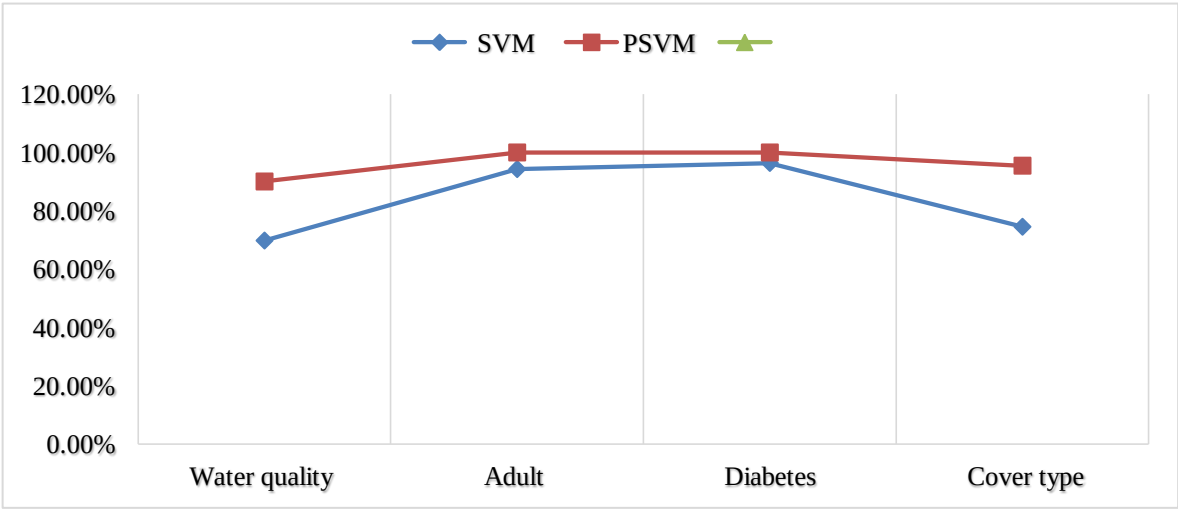

Fig15: Accuracy curve comparison of PSVM.

Table 6 and Fig16 Predicted that if the data size grows, then the processing time of SVM increases. On the other hand, the processing time of PSVM is comparably low concerning that of SVM. Hence, to improve the processing speed of SVM applied to big data for classification, the parallel SVM is used.

Table 6. Comparison of Execution Time (sec) for SVM and PSVM algorithms

Execution Time in (Sec)		Data size (in bytes)			
		888(116 KB)	48842(2.32 MB)	100000(18.2 MB)	581012(109 MB)
Algorithm	SVM	3000.39	17000.7	6000.001	165000.6
	PSVM	15.05	530.74	630.74	690.91

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

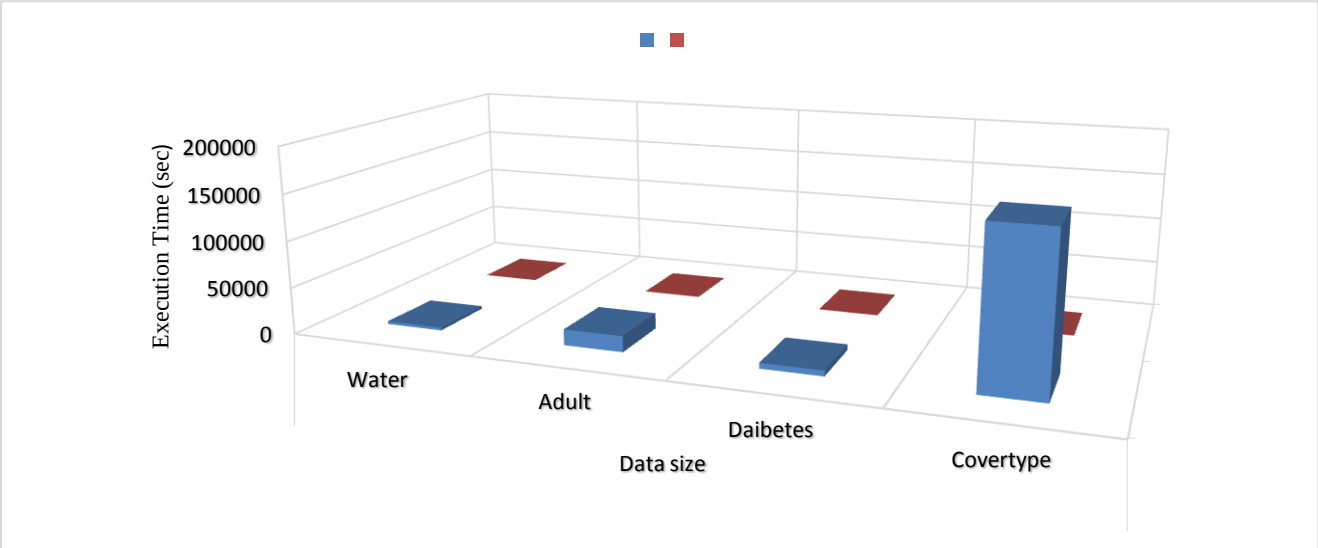

Fig16: The execution time of SVM and PSVM on different sizes of dataset

4. CONCLUSION & RECOMMENDATIONS

Big data faces many problems when dealing with machine learning techniques. SVM is a robust method with a great execution value when data size increases. Parallel computation is found to be ideal with big data analysis, for it divides the data into smaller partitions to increase the efficiency and performance and decrease the computation time. The results in this paper had shown that developing a PSVM model applied to large datasets produces better efficiency than traditional SVM.

It is recommended to use parallel SVM model with other algorithms and compare its results to models in other recent research papers. Also, the PSVM model can be extended for multi-classification by utilizing parallelism related to both CPUs and PSVM. So, this approach can be more beneficial with more complex datasets.

ACKNOWLEDGMENT

We are greatly thankful to our supervisor and all professors who helped us in technical writing, data collection, analysis of this research. Thanks to everyone who were involved to this in the validation of this paper.

REFERENCES

Birzhandi, P., Kim, K.T., Lee, B. and Youn, H.Y., 2019. Reduction of training data using parallel hyperplane for support vector machine. Applied Artificial Intelligence. 33.6 (2019): 497-516.

Çatak, Ferhat Özgür, and Mehmet Erdal Balaban. "A MapReduce-based distributed SVM algorithm for binary classification." Turkish Journal of Electrical Engineering & Computer Sciences 24.3 (2016): 863-873.

Chen, Z., Lin, T., Tang, N. and Xia, X., 2016. A parallel genetic algorithm based feature selection and parameter optimization for support vector machine. Scientific Programming, 2016. Chisholm, Andrew. Exploring data with RapidMiner. Packt Publishing Ltd, 2013.

Do, Thanh-Nghi. "Non-linear classification of massive datasets with a parallel algorithm of local support vector machines." Advanced Computational Methods for Knowledge Engineering. Springer, Cham, (2015). 231-241.

Gola, Jessica, et al. "Objective microstructure classification by support vector machine (SVM) using a combination of morphological parameters and textural features for low carbon steels." Computational Materials Science 160 (2019): 186-196.

Goyal, Aarti, and T. Meenpal. "Kinship verification from facial images using feature descriptors." Cognitive Informatics and Soft Computing. Springer, Singapore, 2019. 371-380.

He, T., Wang, T., Abbey, R. and Griffin, J., 2019. "High-Performance Support Vector Machines and Its Applications." arXiv preprint arXiv:1905.00331 (2019).

Hofmann, Markus, and Ralf Klinkenberg, eds. RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2014.

Hsieh, Cho-Jui, Si Si, and Inderjit Dhillon. "A divide-and-conquer solver for kernel support vector machines." International conference on machine learning. 2014.

Jadhav, Deepali Kishor. "Big data: the new challenges in data mining." International Journal of Innovative Research in Computer Science & Technology 1.2 (2013): 39-42.

Kshirsagar, Nabha, and N. Z. Tarapore. "GPU Parallel Computing of Support Vector Machines as applied to Intrusion Detection

Citation: Iatimad Satti Abd Elkarim1et al., 2020. Parallel SVM Based Classification Technique on big data: HPC center in Sudan. Australian Journal of Basic and Applied Sciences, 14(4): 1-14. DOI: 10.22587/ajbas.2020.14.4.1

System." International Journal of Computer Science and Information Security (IJCSIS) 16.6 (2018).

Ksiaâ, Walid, Fahmi Ben Rejab, and Kaouther Nouira. "Big Data Classification: A Combined Approach Based on Parallel and Approx SVM." International Conference on Intelligent Interactive Multimedia Systems and Services. Springer, Cham, 2018.

Kyrkou, C., Bouganis, C.S., Theocharides, T. and Polycarpou, M.M., 2015. Embedded hardware-efficient real-time classification with cascade support vector machines. IEEE transactions on neural networks and learning systems, 27.1 (2015): 99-112.

Liu, P., Zhao, H.H., Teng, J.Y., Yang, Y.Y., Liu, Y.F. and Zhu, Z.W., 2019. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark. Journal of Central South University, 26.1 (2019): 1-12.

Nandakumar, A. N., and Nandita Yambem. "A survey on data mining algorithms on apache Hadoop platform." International Journal of Emerging Technology and Advanced Engineering 4.1 (2014): 563-565.

Pakize, Seyed Reza, and Abolfazl Gandomi. "Comparative study of classification algorithms based on MapReduce model." International Journal of Innovative Research in Advanced Engineering (IJIRAE) 1.7 (2014): 251-254

Phu, Vo Ngoc, Vo Thi Ngoc Chau, and Vo Thi Ngoc Tran. "SVM for English semantic classification in parallel environment." International Journal of Speech Technology 20.3 (2017): 487-508.

Priyadarshini, Anushree. "A map reduce based support vector machine for big data classification." International Journal of Database Theory and Application 8.5 (2015): 77-98.

Rao, B. Thirumala, and L. S. S. Reddy. "Survey on improved scheduling in Hadoop MapReduce in cloud environments." arXiv preprint arXiv:1207.0780 (2012).

Rezvani, Salim, Xizhao Wang, and Farhad Pourpanah. "Intuitionistic Fuzzy Twin Support Vector Machines." IEEE Transactions on Fuzzy Systems (2019).

Sadasivam, G.S., Madhesu, S., Mumthas, O.Y. and Dharani, K., 2018. Crop Disease Protection Using Parallel Machine Learning Approaches. In Classification in BioApps (pp. 227-259). Springer, Cham..

Schlag, Sebastian, Matthias Schmitt, and Christian Schulz. "Faster Support Vector Machines." 2019 Proceedings of the Twenty-First Workshop on Algorithm Engineering and Experiments (ALENEX). Society for Industrial and Applied Mathematics, 2019.

hvachko, K., Kuang, H., Radia, S. and Chansler, R., 2010, "The Hadoop distributed file system." MSST. Vol. 10..

Singh, Dinesh, and C. Krishna Mohan. "Projection-SVM: Distributed Kernel Support Vector Machine for Big Data using Subspace Partitioning." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

Singh, Dinesh, Debaditya Roy, and C. Krishna Mohan. "DiP-SVM: distribution preserving kernel support vector machine for big data." IEEE Transactions on Big Data 3.1 (2016): 79-90.

Sun, Zhanquan, and Geoffrey Fox. "Study on parallel SVM based on MapReduce." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012

Tavara, Shirin. "Parallel computing of support vector machines: A survey." ACM Computing Surveys (CSUR) 51.6 (2019): 123.

Thanigaivasan, Vivekanandan, Swathi J. Narayanan, and N. Ch Sriman Narayana Iyengar. "Analysis of Parallel SVM Based Classification Technique on Healthcare using Big Data Management in Cloud Storage." Recent Patents on Computer Science 11.3 (2018): 169-178.

Y. You, J. Demmel, K. Czechowski, L. Song, and R. Vuduc, “Design and Implementation of a Communication-Optimal Classifier for Distributed Kernel Support Vector Machines,” IEEE Trans. Parallel and Distributed Systems, vol. 28, no. 4, pp. 974–988, 2017.

Zhao, Hai-xiang, and Frédéric Magoules. "Parallel support vector machines on multi-core and multiprocessor systems." 11th International Conference on Artificial Intelligence and Applications (AIA 2011). IASTED, 2011.