# Data Science : exploration of machine learning, data mining and big data into image recognition pattern

Soniarimamy Nantenaina Serge Rochel, Razafindramintsa Jean Luc, Mahatody Thomas and Manantsoa Victor
Doctoral School of Computer Modelization
Software Engineering and Emergency Research (CLoRe)
Fianarantsoa, Madagascar
soniarimamys@gmail.com, razafindramintsa.jeanluc@yahoo.fr, tsmahatody@gmail.com, vmanantsoa@moov.mg

*Abstract*— **This paper creates accurate image recognition pattern that allows to treat heterogeneous, unstructured and humongous image data. Because, we have noticed that the current image recognition algorithms don't permit to achieve a relevant result and a best accuracy. This inaccuracy result is caused by the data enhancement on social networks, the enhancement of firms' requirement and the ill-treatment of the data into the algorithms. Thus, in order to reach this goal, we have used different data science techniques such as machine learning algorithms like Convolutional Neural Network (CNN), K-Nearest Neighbors (KNN), K-Means, Support Vector Machine (SVM) and Gaussian Mixture (GM). We have also used data mining techniques like Dependency Tree (DT) and Random Forest (RF). These different techniques allow us to achieve as result a robust architecture of image recognition. This architecture supports data science techniques for unstructured and humongous data processing which come from different data storage systems (Mongo DB, CSV file and Raw Data). Then, to validate our strategy, we have conducted image recognition as study case. And after this validation, we have deduced that the use of dimension-reduction techniques like Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are needed to treat huge datasets in case of KNN algorithm, K-Means algorithm, SVM algorithm and Gaussian Mixture algorithm whatever the nature of data. We have also deduced that CNN algorithm, KNN algorithm and RF algorithm produce more accuracy than other algorithms either for heterogeneous, unstructured and voluminous data or for homogenous, structured, and less voluminous data.**

***Keywords- data science; image recognition; machine learning; big data; data mining***

## I. INTRODUCTION

Nowadays, heterogeneous and unstructured data is accumulating due to the enhancement of massive data used in social networks, and in firms [1], [2]. Faced with this enhancement of these data, the accuracy and the effectiveness of the algorithms decrease less and less [3], [4]. As for the image recognition case, the algorithms poorly recognize an object [5]. This poor image recognition result is caused by the treatment of voluminous data, the velocity of data treatment, heterogeneousness of data and the weakness of algorithms [6]. For that, the problem treated in this paper deals with the lack of performance of image recognition pattern faced with the enhancement of heterogeneous, unstructured and humongous data. According to that, we are going to achieve efficient model trained with heterogeneous data, massive data and unstructured data. This attempt is conceivable provided that we know how to manipulate this data and we know how to treat this data with the appropriate algorithms. Therefore, in this paper, we shall confirm this assumption by training the machine learning algorithms and the data mining techniques with unstructured and huge datasets which come from different data storage systems. So, our goal is, one hand, to achieve accurate image recognition pattern trained with heterogeneous data, unstructured data and humongous data and other hand to classify the algorithms according to their accuracies in order to show that Convolutional Neural Network algorithm (CNN), Random Forest algorithm (RF) and K-Nearest Neighbors algorithm (KNN) are more productive than other algorithms with around 99% of accuracy. The rest of paper is divided as the following: in section II, we will analyze the different related works which linked closely into our research. In section III, we will explain the proposed image recognition pattern. In section IV, we will validate our strategy by describing the used datasets and the obtained results. And finally, in section V, we will conclude this work by giving research perspective.

## II. RELATED WORK

Algorithms 'optimization is not a novel problem in image recognition domain. This optimization problem has been declared since its apparition, persists nowadays and have been accentuated since the arrival of massive data treatment.

The hybridization of perform algorithms with structured and homogeneous data is amongst of the solutions to resolve this problem [7]. Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016) have measured the performance of KNN algorithm through the homogenous data classification [8]. While authors have been combined KNN algorithm with clustering algorithm like K-Means, they have achieved just 83 percent of performance as accuracy result. This lack of performance is caused by the act of focusing on algorithm technique not on data treatment nor on data pretreatment. AI-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K.,

& Taha, K. (2015) have also given a glance theory about the use of classifier algorithm with Big Data [9].

Jarrah et al have tried to explain the solution to achieve a perform image recognition model by using Naïve Bayes and by using linear-SVM to treat voluminous data. Unfortunately, this theory is not validated with pragmatic manner in their works.

In distinction with related works that have been shown before, the dimension reduction technique is also amongst of the solutions to resolve the algorithms 'optimization problem. For instance, Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J., A., & Plaza, A. (2015) have attempted to obtain an accurate model by training SVM algorithm with huge datasets [10]. Unfortunately, authors have achieved just 77.9% accuracy even they have combined principle component analysis (PCA) with SVM. This lack of performance is caused by ill-treatment of data during data the treatment phase. Besides, the training data is not voluminous nor various when they have used Big Data term which is based with 03 concepts such as volume of data, velocity of data treatments and variety of data. Almotiri, J., Elleithy, K., & Elleithy, A. (2017, May) have also used dimension reduction technique to compare the result of PCA with Auto-Encoder in domain of handwritten digit recognition via Mnist dataset [11]. As result, they have achieved a good accuracy even the data used are homogeneous, structures and less voluminous (98.1%).

According to these results, we can deduce that the classic approaches which are used in less voluminous, structured and homogeneous data treatment are not valid in heterogeneous, unstructured and massive data treatment [12], [13], [14], [15]. Besides, it's primary to change the manner of how to treat these data during the data preprocessing phase by adding data aggregation function and by adding data dimension reduction techniques. Hence, we propose to introduce a perform image recognition pattern adaptable in scalability of heterogeneous, unstructured and humongous data treatment. First, this model will receive, aggregate, formalize and reduce dimension of data which are voluminous and miscellaneous. Second, this model will treat these data in order to recognize data and in order to give high accuracy.

## III. IMAGE RECOGNITION PATTERN

This section III concerns about the description of our model. This description of model consists to underline its architecture and to explain in detail its each treatment part.

### *A. Data Science Applied in Image Preprocessing Phase*

The achievement of relevant image recognition model dealing with heterogeneous, unstructured and humongous data is conceivable. Then, in this work, we introduce 03 computers which contain different data storage systems and many algorithm to train those data. To do that, the architecture of our model is presented in fig. 1. This fig.1 shows us the importance of data science concept utilization during the data preprocessing phase.

Firstly, to illustrate the use of humongous, unstructured and voluminous data, we have stored 70000 pixels of image in csv files and in Mongo DB none-relational database. Each image's pixels stored in these two different data storage systems is shaped as (70000, 784), (10000, 784) with their corresponding labels which are shaped like (70000,) and (10000,). We have also used 949 raw images formed by (600,784) shape and formed by (349,784) shape with their corresponding labels which are shaped as (600, 10) and (349, 10). Then these data which are stored in different spaces are standardized during the preprocessing phase. This data standardization consists about loading of data, restructuration of data, reshaping of data and aggregation of data.

Secondly, the data standardization is followed by data dimension reduction techniques. In other words, we have to reduce these humongous data which are obtained from heterogeneous and unstructured data in order to train them easily with the machine learning algorithms and with the data mining techniques. Consequently, the dimension reduction techniques which are used in this research are Latent Discriminant Analysis (LDA) and PCA. Briefly, PCA is unsupervised technique used to detect the correlation between variables. For this, the more the correlation between variables are highest, the more the dimension reduction techniques 'outcome is appropriate. Hence, the equation (1) is a basic PCA equation which is presented as the matrix annotation.

$$Y = W'X \quad (1)$$

- W means matrix of coefficients that is determined by PCA.
- X means data matrix which consists of n observations (rows) on p variables (column).
- Y is the matrix of scores.

LDA is also amongst of dimension reduction techniques. Otherwise, the difference between LDA and PCA is that LDA is supervised technique based on Bayesian theory. Hence, the mathematical equation of LDA is presented as in (2).

$$\delta k(x) = x \frac{\mu\kappa}{\sigma^2} - x \frac{\mu\kappa^2}{2\sigma^2} + \log(\pi\kappa) \quad (2)$$

- $\delta k(x)$ presents Linear equation called discriminant.
- $k$ means whole number higher than or equal 2 (number of classes).
- $\sigma 2$ means weight average of sample variances for each classes.
- $\mu k$ means average of all observations in the training data.

### *B. Data Science Applied in pattern building phase*

The In this section, we are going to detail the used algorithms to train and test the standard data that we have reduced during data preprocessing phase. Note that, the aim of using of many algorithms in this work is to find the appropriate algorithm which fits well for the image recognition dealing with humongous, unstructured and miscellaneous data.

First, we have used unsupervised and supervised machine learning algorithms to train the humongous and miscellaneous data which come from different data storage systems. In the case of supervised machine learning algorithms, we have used CNN algorithm. Then, in this work, CNN algorithm is composed with 03 convolutional layers (CONV) and 03 pooling layers (POOL) in hidden layer. Each CONV is followed by dropout function and RELU activation function in order to fight against the over fitting concept. Besides, at the end of training, we can say that CNN is amongst of the 03 algorithms the most appropriate for image recognition algorithm dealing with various, humongous and unstructured data (table.3). We have also used KNN as supervised machine learning algorithm. KNN algorithm is based on act of finding K (whole number) dots which is nearest of unknown dot. Therefore, KNN leans on calculation of the distance between of unknown point and other points defined in training data. The basic equation of Euclidian distance is presented in (3):

$$d(q,p) = d(p,q) = \sqrt{(q1-p1)^2 + (q2-p2)^2 + (qn-pm)^2} \quad (3)$$

- q1 to q2 means the attributed values for the one observation.
- p1 to p2 means the attributed values for the other observation.

In this research, we have 10 observations in our datasets and we have chosen k value to 10. Choosing value of K to 10 means that we have tried to find the classification of one unknown point from the classification of 10 nearest points of this unknown point. After training our model with KNN algorithm, we have achieved a good result (table.1). Moreover, SVM belongs of machine learning algorithms that we have used but its result is not appropriate in relation to other algorithms and techniques (table.2). This SVM inappropriate result is caused by the adjustment of its hyper plan to separate the data into 02 parts when our data are separated by default into 10 classes. In the case of unsupervised machine learning algorithms, we have used K-Means and Gaussian Mixture to test the efficiency of these clustering algorithms into the image recognition dealing with humongous and miscellaneous data. K-Means is used to gather data into k (whole number) clusters from the distance of these data towards of cluster's centroid. For this, the basic mathematical equation is based on Euclidian distance which have shown in formula.3. At the end, we have chosen k value to 10 and we have deduced that K-Means and Gaussian Mixture produce low performance in relation to other techniques (table.1). This low performance is caused by the in appropriation of clustering algorithms dealing with labeled data treatment. Otherwise, when we combined K-Means and Gaussian Mixture with dimension reduction techniques such as PCA and LDA, we have visualized that its results are little improved (table.4).

Second, we have used data mining techniques such as Dependency Tree (DT) and RF. DT is amongst of algorithm to formalize the way of making decision. Then RF is just deep collection of several DT that uses heuristic functions such impurity and Gini impurity for data treatment. The basic mathematical equation of Gini impurity and entropy is presented in (4) and (5).

$$f(x) = \sum_{i=1}^{N} -fi(1 - fi) \quad (4)$$

$$f(x) = \sum_{i=1}^{N} -fi(1 - fi) \quad (5)$$

- fi means the proportion between the number of elements in the separated groups in relation to number of elements in group before separation.

In short, RF is considered as improved version of DT. Consequently, in our model, we have put the minimum number of leaf to 1 (mean_samples_leaf=1) and estimator number value to 10 (n_estmators=10). And after training and validating our model, we have achieved a good accuracy even its accuracy is little decreased in relation to the accuracy of KNN algorithm and CNN algorithm (table. 2).

## IV. EXPERIMENTS

In this section, we are going to describe the datasets that we have used to validate our strategy. We are also going to show and discuss the results of our model by using image recognition domain as study case.

### *A. Datasets*

In this approach, we have chosen two datasets (FashionMnist, Mnist) and image recognition domain as case of study. FashionMnist is a dataset of cloths which gather 70000 images. Then, each image in this dataset is made with black and white color and also shaped with 784 dimensions (28*28). FashionMnist dataset is divided into groups: the training data which has 60000 images and the test data which contains 10000 images. Moreover, all of the images in this dataset are classified into 10 categories such as: T-Shirt/Top, Trouser/pants, Pullover shirt, Dress, Coat, Sandal, Shirt, Sneaker, Bag and Ankle boat. Likewise, Mnist dataset has the same structure as FashionMnist dataset. Unfortunately, the little difference is that Mnist contains 70000 images of digit numbers which starts with 0 and ends to 9. Thus, one hand, Mnist dataset and FashionMnist dataset can be downloaded on kaggle web site[a] and other hand, the created models are pushed in github repository[b].

a. https://www.kaggle.com/rochel/mnist_and_fashionmnist_dataset_csv_raw
b. http://www.github.com/rochel05

TABLE I. ACCURACY OF OUR MODEL IN RELATION TO THE VARIETY OF DATA

| **Variety** | **Model accuracy in relation to the variety of data (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | ***RF*** | ***DT*** | ***KNN*** | ***K-Means*** | ***CNN*** | ***SVM*** | ***GM*** |
| HoSD Mnist | 90.88 | 72.7 | 94.1 | 12.07 | 98.79 | 12.6 | 21.7 |
| HoSD Fashion | 73.2 | 64.0 | 73.0 | 15.02 | 78.09 | 13.3 | 19.0 |
| HeUD Mnist | 90.47 | 64.51 | 93.51 | 14.89 | 98.88 | 12.79 | 16.42 |
| HeUD Fashion | 76.87 | 64.13 | 69.23 | 20.69 | 78.20 | 29.72 | 11.11 |

TABLE II. ACCURACY OF OUR MODEL IN RELATION TO THE VOLUME OF DATA

| **Volume** | **Model accuracy in relation to the volume of data (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | ***RF*** | ***DT*** | ***KNN*** | ***K-Means*** | ***CNN*** | ***SVM*** | ***GM*** |
| [10600,784] | 37.24 | 59.02 | 78.79 | 09.45 | 92.55 | 12.79 | 10.05 |
| [60600,784] | 77.65 | 66.85 | 93.12 | 11.91 | 98.99 | 12.79 | 16.45 |
| [70000,784] | 75.87 | 76.57 | 93.65 | 08.24 | 98.88 | 11.38 | 11.663 |
| [70600,784] | 71.20 | 76.16 | 93.02 | 15.56 | 98.72 | 11.32 | 13.53 |

TABLE III. ACCURACY OF OUR MODEL IN RELATION TO THE EVALUATION METRICS

| **Metrics** | **Model accuracy in relation to evaluation metrics (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | ***RF*** | ***DT*** | ***KNN*** | ***K-Means*** | ***CNN*** | ***SVM*** | ***GM*** |
| F1-score | 0.87 | 0.67 | 0.94 | 0.11 | 0.99 | 0.12 | 0.16 |
| Recall | 0.78 | 0.67 | 0.93 | 0.11 | 0.99 | 0.12 | 0.16 |
| Precision | 0.99 | 0.67 | 0.95 | 0.11 | 0.99 | 0.12 | 0.16 |

TABLE IV. ACCURACY OF OUR MODEL IN RELATION TO PCA AND LDA UTILIZATION

| **Model** | **Model accuracy in relation to PCA and LDA (%)** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***RF*** | | ***DT*** | | ***KNN*** | | ***K-Means*** | | ***CNN*** | | ***SVM*** | | ***GM*** | |
| | Pca | Lda | Pca | Lda | Pca | Lda | Pca | Lda | Pca | Lda | Pca | Lda | Pca | lda |
| HoSD Mnist | 45.31 | 71.36 | 81.5 | 85.42 | 94.26 | 91.82 | 12.91 | 10.29 | 92.66 | - | 12.5 | 74.55 | 20.1 | 22.0 |
| HoSD Fashion | 30.9 | 35.3 | 50.8 | 38.7 | 73.0 | 55.2 | 22.14 | 14.29 | 78.90 | - | 9.1 | 9.1 | 13.3 | 17.2 |
| HeUD Mnist | 71.20 | 89.84 | 59.21 | 85.36 | 93.86 | 91.40 | 14.94 | 17.05 | 79.42 | - | 12.7 | 88.90 | 12.4 | 11.3 |
| HeUD Fashion | 52.85 | 56.83 | 55.44 | 53.48 | 69.53 | 63.82 | 31.20 | 28.89 | 71.53 | - | 29.7 | 29.72 | 30.0 | 30.0 |

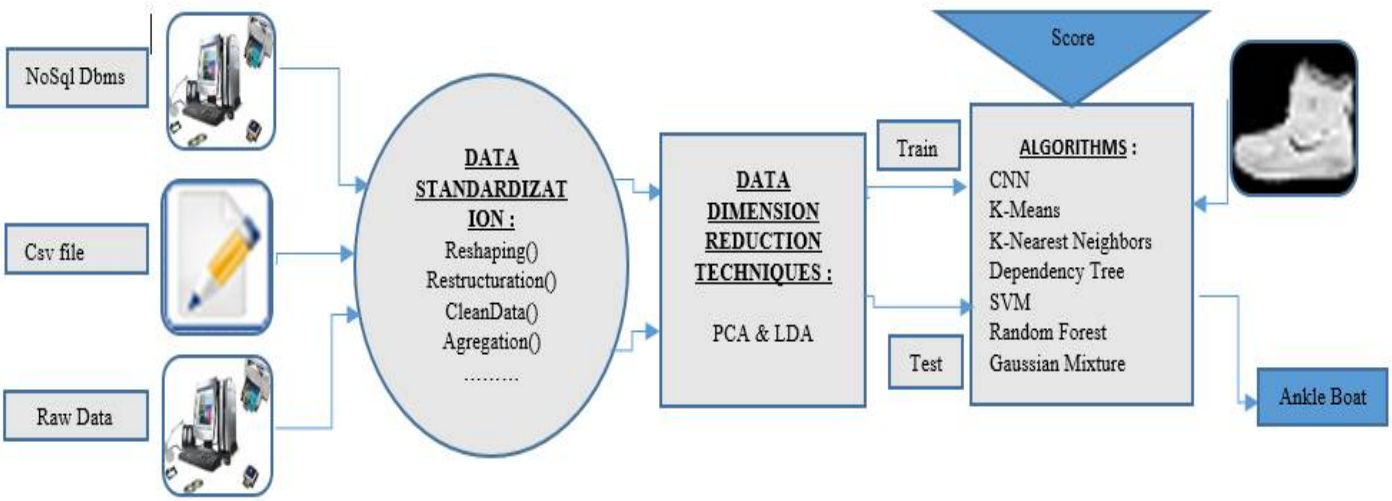


Fig. 1: Architecture of our Model

Then, to orientate the use of these datasets into our aim such the obtainment of perform image recognition model in the case of using heterogeneous, unstructured and humongous data, we have treated these datasets into 03 parts. First, we have trained our model by using csv files which contain 60000 images (60000, 784).

Then, each image in this dataset has 28 width pixels and 28 length pixels (table.1). Second, we have also trained our model with heterogeneous and unstructured data which is obtained by hybridization of data come from Mongo DB non-relational database, csv file and Raw Data (table.3). This, we have enhanced the volume of trained data progressively during the training phase. Therefore, the result of this data volume enhancement can be visualized in table 2.

### *B. Results and Discussions*

In this section, we will show the results of the proposed approach. These results consist about the accuracy of our model in different manners by using unstructured and voluminous data which come from different data storage systems (Mongo DB -Csv, Csv-Raw data, Mongo DB-Raw Data, Mongo DB-Csv-Raw Data).

*1) Result in relation to the variety of data:* In relation to the variety of data, CNN algorithm, KNN algorithm and RF algorithm are the three algorithms which produce more accuracy (table.1).

This table. 1 shows that CNN algorithm, KNN algorithm and RF algorithm produce more accuracy than other algorithms either in homogenous and structured data treatment (HoSD) or heterogeneous and unstructured data treatment (HeUD)

(CNN: 98.88%). Likewise, within utilization of Mnist dataset of FashionMnist dataset, the results which are given by CNN algorithm, KNN algorithm and RF algorithm are profitable. Hence, these results prove that we can achieve accurate image recognition model even data are heterogeneous, unstructured and voluminous. We can also deduce that the results of our model by using FashionMnist dataset is less accurate than Mnist dataset (KNN Mnist: 94.1%), (KNN fashion: 73%). For those, discussion is open with the selection of datasets in the case of selection domain dealing with heterogeneous and voluminous data.

*2) Result in relation to the volume of data:* In relation to the volume of data, the accuracy of CNN algorithm, KNN algorithm and RF algorithm are always higher than other algorithms.

After having trained and tested our model with heterogeneous, unstructured and humongous data, table.2 shows that CNN algorithm, KNN algorithm and RF algorithm produce more accuracy than other algorithms and other data mining technique (CNN: 98.88%). This table.2 shows also that CNN algorithm accuracy is always the best even we have enhanced the volume of data progressively. Otherwise, K-Means algorithm, Gaussian Mixture algorithm and SVM algorithm produce less accuracy in this work (Gaussian Mixture: 16.45%). So, the achievement of best accuracy given by CNN shows us that our assumption is approved and validated. But, the discussion is opened about the pertinence of CNN algorithm if we shall enhance progressively our data more than we have done.

*3) Result in relation to the evaluation metrics:* In relation to the evaluation metrics (F1-score, Recall, and Precision), the result given by our model is presented in table.3.

This table. 3 shows that the use of CNN algorithm is worthwhile to recognize image dealing with heterogeneous, informal and voluminous data (CNN: 0.99%). This best CNN algorithm is achieved thanks to the optimization of algorithm such as the use of dropout function in hidden layer and the formalization of data like data normalization, data aggregation and data dimension reduction. Thus, we can say that our approach is verified. Verified approach means that the achievement of perform image recognition model dealing with heterogeneous, unstructured and humongous data is conceivable if we know how to optimize the data and how to choose the relevant algorithm. Then, in this research, the result shown in table. 3 is achieved thanks to the mixing of data which come from different data storage system such as Mongo DB non-relational database, csv files and Datasets of Raw image. From this fact, the discussion is concerned about the choice between none-relational database and relational database.

*4) Result in relation to the use of PCA and LDA :* In relation to the use of PCA and LDA as data dimension reduction techniques, the result is displayed in table. 4.

One hand, according to table. 1 and table. 4, we can deduce that the use of dimension reduction techniques (PCA and LDA) with heterogeneous and voluminous data is really appropriate for KNN algorithm, K-Means algorithm, SVM algorithm and Gaussian algorithm either by HoSD or HeUD (KNN with PCA: 93.86%), (KNN without PCA and LDA: 93.51%). Other hand, we can visualize that the result from CNN algorithm, DT algorithm and RF algorithm is decreased little a bit but is not induced serious impact in image recognition test processing (RF with PCA: 89.84%), (RF without PCA and LDA: 90.47%). Moreover, we can also show that LDA accuracies are higher than PCA accuracies. According to these results, the discussion is about the LDA accuracies in relation to PCA accuracies in case of dataset modification.

*5) Result in relation to other searcher results:* In relation to other searcher results, the result of our model is always pertinent (table. 5).

From the table.5 and the other tables, we can announce that our hypothesis is verified and validated. The good processing of heterogeneous, unstructured and voluminous data during the preprocessing phase and the choice of a good algorithm can lead to the best accuracy and the best result in image recognition domain dealing with data science concept.

TABLE V. ACCURACY OF OUR MODEL IN RELATION TO OTHER SEARCHER RESULTS

| **Model** | **Model accuracy in relation to searcher results (%)** | |
|---|---|---|
| | ***Dataset*** | ***Accuracy*** |
| RC-KNN[8] | MNIST | 0.72 |
| LC-KNN[8] | MNIST | 0.83 |
| Neural + No Loss[7] | MNIST | 0.975 |
| Auto Encoder + PCA[11] | MNIST | 0.981 |
| Our Model | MNIST | 0.99 |

## V. CONCLUSION

In this paper, we have dealt with the problem of deficiency performance of image recognition algorithms with enhancement extensively of data. Reminding that the aim of proposed approach is the achievement of accurate image recognition pattern trained with unstructured and humongous data which come from different data storage systems. To do that, we have used Mongo DB non-relational database, csv file and datasets of several raw images. We have also trained the heterogeneous, unstructured and humongous data with machine learning algorithms (CNN, KNN, SVM, and Gaussian Mixture) and data mining techniques (RF and DT). After having trained and tested these data with different machine learning algorithms and with data mining techniques, we have achieved accurate image recognition model which recognize image with 99 percent. Moreover, our result shows that the use of dimension reduction techniques (PCA, LDA) dealing with heterogeneous, unstructured and voluminous data is really efficient for the KNN algorithm, K-Means algorithm, SVM algorithm and Gaussian Mixture algorithm. We have also deduced that CNN algorithm, KNN algorithm and RF algorithm produce more accuracy than other algorithms either with various and heterogeneous data treatment or voluminous data treatment. Finally, the approach treated in this paper is different by the use of heterogeneous and humongous data which are gathered from different data storage systems, by the comparison of machine learning algorithm accuracies and data mining accuracies to treat the various data and by use of data dimension reduction techniques. Not only this approach can be used in many image recognition subdomain and text recognition subdomain such vehicle number plate recognition system. It can also be applied in data treatment on the cluster of remote servers. This product is also tested throughout different machine learning algorithms and data mining techniques before filtering the best algorithm which fit with massive data treatment. Then, our future work will be consisting about the widening of this work by using Big Data Framework.

## REFERENCES

[1] Liu, X., Shyn, H., & Burns, A. C. (2019). Examining the impact of luxury brand's social media marketing on customer engagement : Using big data analytics and natural language processing. Journal of Business Research.

[2] Solanki, J., Jadeja, V., Patel, C., Parmar, S., Gadhiya, S., Solanki, J., … & Gadhiya, S. (2019). Security and Challenges in Big Data : A Survey. International Journal, 5, 61-63.

[3] Chang, R.M., Kauffman, R. J.,& Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. Decision Support Systems, 63, 67-80.

[4] Roh, Y., Heo, G., & Whang, S. E. (2018). A Survey on Data Collection for Machine Learning:a Big Data-AI Integreation Perspective. arXiv preprint arXiv:1811.03402.

[5] L, C. H. ,& Yoon, H. J. (2017). Medical big data:promise and challenges.Kidney research and clinical practice,36(1),3.

[6] Bhatnagar, R. (2018, February). Machine Learning and Big Data Processing : A Technological Perspective and Review. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 468-478). Springer, Charm.

[7] Demidova, L. A., Klyueva, I. A., & Pylkin, A. N. (2019). Hybrid Approach to Improving the Results of the SVM Classification Using the Random Forest Algorithm. Procedia Computer Science, 150, 455-461 .

[8] Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient KNN classification algorithm for big data. Neurocomputing, 195, 143-148.

[9] AI-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87-93.

[10] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J., A., & Plaza, A. (2015). On understanding big data impacts in remotely sensed image classification using support vector machine methods. IEEE journal of selected topics in applied earth observations and remote sensing, 8(10), 4634-4646.

[11] Almotiri, J., Elleithy, K., & Elleithy, A. (2017, May). Comparizon of autoencoder and Principal Component Analysis followed by neural network for e-learning using handwritten recognition. In 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-5). IEEE.

[12] Agarap, A. F. (2017). An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification. 2017. arXiv preprint arXiv:1712.03541.

[13] Jia, S., Cristianini, N. (2015). Learning to classify gender from four million images. Pattern Recognition Letters, 58, 35-41.

[14] Aparna P K, Dr. Rajashree Shettar, "Hybrid Decision Tree using K-Means for Classifying Continuous Data", International Journal of Innovative Research In Computer and Communication Engineering, October 2015.

[15] Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learnig. arXiv preprint arXiv:1712.04621.