ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse327952020.pdf
https://doi.org/10.30534/ijatcse/2020/327952020

Empirical Evaluation of the Classification of Deep Learning under Big Data Processing Platforms

Dima Suleiman[1,*], Malek Al-Zewairi[2], Wael Etaiwi[3], Ghazi Al-Naymat[4,5]
[1] Information Technology Department, King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan
Dima.suleiman@ju.edu.jo
[2] Jordan Information Security and Digital Forensics Research Group, Amman, Jordan
[3] Business Information Technology Department, King Talal School of Business Technology, Princess Sumaya University for Technology, Amman, Jordan
[4] Department of IT, College of Engineering and Information Technology, Ajman University, UAE
[5] Computer Science Department, King Hussein Faculty of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan

ABSTRACT

A significant amount of data produced by industries must be processed by machine learning algorithms to facilitates the decision-making process. Traditional machine learning platforms cannot handle the data characterized by their volume, variety, and velocity. Several machine learning toolkits have recently been developed to manage big data. This paper examines the efficiency of H2O and SparklingWater machine learning platforms, dedicated to the processing of big data, in terms of training time and error metrics. Experiments are conducted on four datasets which are the Santander Bank, Credit Card Fraud Detection Dataset, p53 Mutants, and Poker Hand Datasets. The evaluation results show that the H2O platform significantly outperforms the SparklingWater platform in terms of model training time by almost fifty percent thus achieving convergent results.

Key words: Big Data, H2O, SparklingWater, Spark, Prediction.

1. INTRODUCTION

In general, when data is complicated and huge, we can call it Big Data. However, when the problem characterizes by at least one of the characteristics of the big data 3Vs, which are Variety, Velocity, and Volume, then it can be called a big data problem. Variety means that the data can have a range of formats such as structured, semi-structured, and unstructured format. An Example of structured data is the one stored in tables, whereas semi-structured data such as XML, and finally multimedia files are considered as unstructured data [1]. On the other hand, the velocity is the speed of collecting data, while the last V refers to the volume of data which is related to the size of big data, which is huge, concluded that commodity machines and basic tools cannot be used to process it.

* Corresponding author. Tel.: +962-6-5359949; fax: +962-6-5347295.
E-mail address: d.suleiman@psut.edu.jo

Machine learning tools can be used to predict future knowledge by learning from experience, such as in Clinical Anesthesia [2]. In general, machine learning algorithms such as deep learning can be used to make predictions [3,4,5] by building models where machine learning workflows begin with building the prediction model.

After that, the parameters of the prediction model would be modified as a result of the evaluation of the model in order to produce significant results. Besides, deep learning is much related to Artificial Intelligence, both of which have acted like human brains [6] to make decisions by learning and analyzing complex problems. However, traditional machine learning algorithms cannot be used to make big data decisions, while dealing with big data is crucial since many industries produce a huge amount of data (e.g. IoT sensors, and user clickstreams) that need to be processed.

New machine learning toolkits have been produced to process big data, such as Spark, Hadoop, Steam, H2O, and SparklingWater, since traditional toolkits, such as Weka and R that cannot handle a huge amount of data. The presence of big data has opened the door to research in several domains such as bioinformatics, healthcare, and business analytics, making the advent of a new generation of machine learning toolkits important.

Hadoop (Apache Hadoop) and Spark are both an open-source implementation of MapReduce and both operate in cluster environments and share the task of processing big data among commodity computers[7] within the cluster. However, Spark can process the data faster as it is in-memory, and stores data in memory in Resilient Distributed Datasets (RDD) [2]. On the other hand, MapReduce is used to divide large jobs into sub-tasks to perform all sub-tasks in parallel [8].

Besides, Hadoop and Flame, H2O, and SparklingWater are open-source platforms. On the other hand, the H2O platform offers built-in machine learning libraries, parallel processing engines, math, and data analytics; it also facilitates the processing of data using its tools. SparklingWater, though, is a combination of Spark and H2O; it takes advantage of both. SparklingWater, therefore, enables developers to use scalable and fast machine learning algorithms supported by H2O in their applications [9,10].

One of the interesting research topics that gain competitive advantages is the development of a product recommendation system that is highly dependent on customer preferences that must be studied, analyzed, recorded, and collected through the tracking of customer activities, behavior, and habits. However creating a successful online product or service prediction model is one of the challenges and has many advantages, such as increasing the customer's option to buy additional items. This prediction model can be applied by processing massive customer behavior data using an intelligent machine learning algorithm such as deep learning.

Several machine learning algorithms and associated toolkits have been explicitly developed to deal with big data problems such as personalized product recommendation system, predicting the outcome of a game, finding protein anomalies for biomedical applications, and credit card fraud detection system. However, their performance is yet to be evaluated. In this paper, an empirical evaluation study of two machine learning platforms under big data processing systems (namely; H2O and SparklingWater) is provided. Four different datasets were used to evaluate the performance of the two platforms in solving supervised classification problems using a deep learning classifier in a cluster environment.

The rest of this paper is organized as follows: Section 2 presents the literature review. Section 3 describes the dataset and its preparation process. The evaluation results are discussed in section 4. Finally, the conclusion is covered in section 5.

2. RELATED WORKS

In this section, comparisons were made between several studies that using advanced big data machine learning platforms in terms of the datasets used, advantages, evaluation criteria, conclusion, technique, and machine learning algorithms. These comparisons can be found in Table 1.

In conclusion, deep learning algorithms are the most machine learning techniques used to make predictions about big data. Furthermore, since the H2O platform has deep learning algorithms that are implemented on it, it is used in many areas of research in addition to the significant result it has obtained in predicting. In this paper, experiments were performed on four different datasets to measure the effects of using deep learning in prediction by using H2O and SparklingWater ML platforms.

3. DATASETS

To evaluate the two systems' ability to handle big data problems, four separate datasets were used in the evaluation process. Each dataset has its unique characteristics in terms of the number of instances, the number of features, and type of data. The following subsections address the various datasets used in the evaluation of H2O and SparklingWater.

3.1 Santander Bank Dataset

The dataset includes personal and financial information about the clients of the Santander Bank. The dataset was published on the Kaggle website [2] as part of a public competition to come up with a recommendation system to predict-product(s) existing customers of the Santander Bank could buy. The dataset consists of forty-eight features and 13,647,310 instances; the first twenty-four features are personal data; while the last twenty-four features are financial products provided by the bank [21].

3.2 Credit Card Fraud Detection Dataset

The dataset contains two days of credit card transactions made by European cardholders in September 2013 and classified as either fraudulent or genuine transactions. Data were collected as part of big data mining and fraud detection project between Worldline and the Machine Learning Group at The Université libre de Bruxelles (ULB). The dataset has a total of thirty features; twenty-eight features were anonymized using the Principal Component Analysis (PCA) and are referred to V1 to V28; while, the remaining two features (i.e. Time and Amount) were not anonymized. The dataset consists of 284,807 transactions and only 492 instances are listed as fraud [22].

3.3 P53 Mutants Dataset

The dataset was released in 2010 by the Institute for Genomics and Bioinformatics at the University of California to detect mutations in the tumor suppressor protein (i.e. p53) as either active or inactive. It should be noted that a newer edition was released in 2012. It consists of 16,772 instances and 5408 numerical features, where the first 4,827 features represent 2D electrostatic and surface-based features, while, the remaining 582 features represent 3D distance-based features. The class is labeled as "active" or "inactive" p53 protein [23].

3.4 Poker Hand Dataset

The Poker Hand dataset constitutes a multivariate classification problem. The dataset was created by the Intelligent Systems Research Unit at Carleton University to predict a poker hand (a poker is a form of card games) in 2002. It contains 1,025,010 instances and 10 numerical features [24].

4. RESULTS AND EVALUATION

In this section, the results of the experimental evaluation between H2O and SparklingWater are presented and discussed for each of the four datasets. The configuration of the evaluation environment is similar to the one presented in [21]. For each dataset, the default deep learning classifier for the platform was used, where the number of hidden layers is set to the number of classes of each dataset (i.e. two hidden layers for the Santander, Credit Card Fraud, and p53 datasets

[2] https://www.kaggle.com/competitions

Table 1: Comparisons between several studies of different big data machine learning platforms

Ref	Data Set	Main Advantages	Criteria / Conclusion	Technique/ Machine Learning Algorithms
(Kejela, Esteves, & Rong, 2014) [11]	The dataset is gathered from sensors that produce real-life data and Equipment in companies of oil and gas, these sensors monitor the drilling process. Training-set is (504,389 data points) and test-set is (10,351 data points).	Ease of use, Machine learning algorithms are strong and scalable, and processing is in-memory.	A criterion is Prediction errors. The conclusion is that the Prediction accuracy of GBM is better than the prediction accuracy of GLM. Criteria are classification performance. The conclusion is that Both GBM and GLM produce the same accuracy in terms of classification.	1. Comparison between: 2. The generalized linear model (GLM) which is a built-in machine learning in H2O. It was used in cases where linear distribution does not provide good summarization results. It is used for gamma, Gaussian, binomial, and Poisson. 3. A gradient boosted model (GBM) is a prediction model in a decision tree form. For classification, GBM used multinomial loss function and Gaussian for regression.
(Grolinger, Capretz, & Seewald, 2016) [12]	Data is obtained from Budweiser Gardens. The capacity of data was of over 10000 seats located in Canada, London, and Ontario. Green Button (GB) Connect My Data was used for obtaining Electricity consumption. It is 15 minutes reading of electricity consumption. 43,680 data points.	Reducing the training and computation time and increasing the prediction accuracy.	The first criterion is the mean absolute percentage of error (MAPE) and the second one is the coefficient of variance (CV). The conclusion is that in terms of both accuracy and training time, Local SVR outperformed H2O.	Comparison between: 1. Deep learning - H2O deep neural networks 2. Local SVR and local learning (local SVR) where SVR is Support Vector Regression which is a form of SVM Support Vector Machine.
(Ha & Nguyen, 2016) [13]	For validating their approach, they used two datasets that are publicly available for credit approval in German and Australian.	The reasons for using deep learning is reducing running time and dimensionality and increasing classifier accuracy.	Higher prediction rate aver-age accuracy increases to 74.68% for German credit approval and 86.24% in Australian.	Comparisons between: 1. Random Forest. 2. H2O Deep Learning using the R language.
(Miškuf & Zolotová, 2016) [14]	Dataset from UCI repository, which contains 20000 instances of 26 classes. The model is used to make prediction on the last 4000 items and training on the first 16000.	This idea can be used in industrial systems, data analytics and cloud.	Precision and Recall where H2O is the best among the others and according to error rate, H2O is lower two times than the best class in Azure.	Comparisons between 1. Azure multi class identifier: Multiclass Neural Network, Multiclass Decision Jungle, Multiclass Logistic Regression, Multiclass Decision Forest. 2. H2O Deep Learning using R language.
(Wakita, Oku, & Kawagoe, 2016) [15]	Fashion SNS "WEAR" in Japan is the data set used for learning. Data set consists of 115 profile items.	Improve the recommendation accuracy.	The accuracy of Deep learning was very high, it is three times faster than RF and seven times faster than SVM.	Comparisons between 1. SVM (Support Vector Machine). 2. RF (Random Forest). 3. H2O Deep Learning using R language.

(Uppu, Krishna, & Gopalan, 2016b) [16]	The training sample consists of 320000.	Prediction Accuracy	Improvement of prediction accuracy over other algorithms.	Deep Learning in H2O with one input layer and three hidden layers with 50 neurons for each layer. Comparisons between 1. Deep learning. 2. LR, RF GBM.
(Zhang et al., 2016) [17]	To train the model 970 samples were taken, and 30 to verify the model.	Deep Learning is more effective than Back Propagation.	Root Mean Square Error (RMSE). Mean Relative Error (MRE). selection of model parameters: (learning rate and number of hidden nodes, which was 12 in this study)	Comparisons between 1. Deep Learning. 2. Back Propagation (BP) neural network.
(Uppu, Krishna, & Gopalan, 2016a) [18]	There are two scenarios; the first one consists of 2400 datasets and the second one consists of 25200 datasets.	Prediction accuracy, cross-validation consistency (CVC), and classification error.	The accuracy of deep learning is the highest prediction of 97.01%.	Deep Learning with one input layer and three hidden layers with 50 neurons for each layer. The distribution is cross-entropy along with multinomial. Comparisons between 1. Deep learning. 2. SVM, NN GBM, and many others.
(Gupta, 2016) [3]	For analysis purposes, the dataset that was used is hosted from the UCSD-FICO Data mining contest 2009. The transactions in the testing data set are 50000 while training contains 100,000 transactions.	High Recall, lowest loss in terms of finance.	True Positive (TP), False Positive (FP), True Negative (TN), False Positive (FP), Precision, Recall. Recall of the Deep learning model is very high, however; it does not improve the performance of identifying fraudulent transactions, which results in customer dissatisfaction.	Comparisons between: 1. Deep learning H2O. 2. Sampling methods: Over-sampling SMOTE Under-sampling ROSE and hybrid. 3. Ensemble methods: RF, GLM, and GBM.
(Etaiwi et al., 2017) [19]	customer's personal information and behavior dataset from the Santander Bank contains around 14 million records	multi-class classifiers are more efficient than binary classifiers for prediction problems	The results showed that the NB prediction approach is more efficient than SVM in term of precision, recall, and f-measure	Then Naïve Bayes and Support Vector Machine classifiers of the Apache Spark MLlib were applied.
(Baldominos et al., 2019) [20]	The data of the dataset consists of 13 different human activities such as biking, jogging, walking, sitting, typing, standing, and others. Data collected from healthy persons whose ages range from 23-35	The best accuracy was achieved using the ensembles of decision trees	Accuracy and F1-measure	1. Traditional machine learning (NB, KNN, logical regression, ensemble and multi-layer perceptron, random forest) 2. Convolutional neural networks

and nine hidden layers for the Poker dataset). Whereas, the number of neurons is set to the number of features of each dataset for each hidden layer as follows:

- Santander dataset: 48 neurons
- Credit Card Fraud dataset: 30 neurons
- P53 dataset: 5408 neurons
- Poker dataset: 10 neurons

Two sets of experiments were conducted; in the first experiment, each dataset was randomly split into three subsets: the training set with 60% ratio, the testing set with a 30% ratio, and the validation set with 10% ratio using the same seed value (i.e. 793555). Then, each classifier was trained using 3-fold cross-validation. However, for the Santander dataset twenty-four different classifiers were generated as described in [21] to accommodate the twenty-four products. For each dataset, the average value of Accuracy, F1-score, Precision, Recall, and Time were reported as shown in figures Figure1 (a-h). In the second experiment, the datasets were randomly split into two subsets: the training set with a 60% ratio, the testing set with a 40% ratio using the same seed value (i.e. 190235). Similarly, the average value of the aforementioned measures was reported as shown in figures Figure2 (a-d). It is worth mentioning that because the testing results vary according to the threshold value, only the results of 0.5 threshold values are reported. Also, the time measure was normalized to a value between zero and one on the interval (0, 1].

In Figure1 (a), the results of training the model using a 3-folds cross-validation technique on the Santander Bank dataset are shown; while in (b) the results for testing the same model on 30% unseen data are shown. Both platforms have scored convergent results. For the error generalization measures, the results were close up to 0.02. On the other hand, SparklingWater has scored slightly better training time with 0.02 faster than H2O. However, for testing on unseen data H2O has scored better results with 0.07 faster testing time.

Similarly, the error generalization results of training and testing the model on the Credit Card Fraud Detection dataset were very close up to 0.03 as shown in Figure1 (c) and Figure1 (d). Nonetheless, the H2O platform was significantly slower in training the model than SparklingWater (i.e. 0.57 slower). However, it was significantly faster in testing the model on unseen data (i.e. 0.71 faster).

For the P53 Mutants dataset, the error generalization measures were identical up to 0.001%, but the training and testing time varied significantly between the two platforms as shown in Figure1 (e) and Figure1 (f). Training the model on H2O was 0.35 faster than SparklingWater. However, the testing time was 0.42 faster on SparklingWater.

When considering the Poker Hand dataset, the error generalization results were also close up to 0.06 where H2O has scored the higher score as shown in Figure1 (g) and Figure1 (h). The timing results between the two platforms were also different, in which SparklingWater has scored the better training and testing time (i.e. 0.59 and 0.27 faster time respectively).

For the second experiment, each model was trained on 60% of seen data without using cross-validation and tested on 40% unseen data. As shown in Figure2, the results of error generalization measures for Santander, P53 Mutants and Poker Hand datasets were almost identical between the two platforms with a difference of up to 0.02, but the for Credit Card dataset the recall and f1-score were noticeably lower (up to 0.4) on the H2O platform. For the model testing time, the Santander dataset has a convergent testing time of up to 0.06 difference where H2O has scored the lower time as shown in Figure2 (a). The model testing time for the credit card fraud detection dataset was 0.15 faster on SparklingWater as shown in Figure2 (b). While for the P53 Mutants dataset, it was greatly faster on SparklingWater of up to 0.8 with identical error generalization results of up to 0.001 as shown in Figure2 (c). On the contrary, the testing time for the Poker Hand dataset was noticeably faster on the H2O platform of up to 0.63 as shown in Figure2 (c). Interestingly, using cross-validation did not significantly improve the error generalization measures for Santander, Credit Card, and Poker Hand datasets. However, for the P53 Mutants dataset, utilizing the cross-validation technique has resulted in a better recall measure of up to 0.45 on the H2O platform and 0.09 on SparklingWater. Moreover, the F1-Score improved by 0.34 on H2O platform.

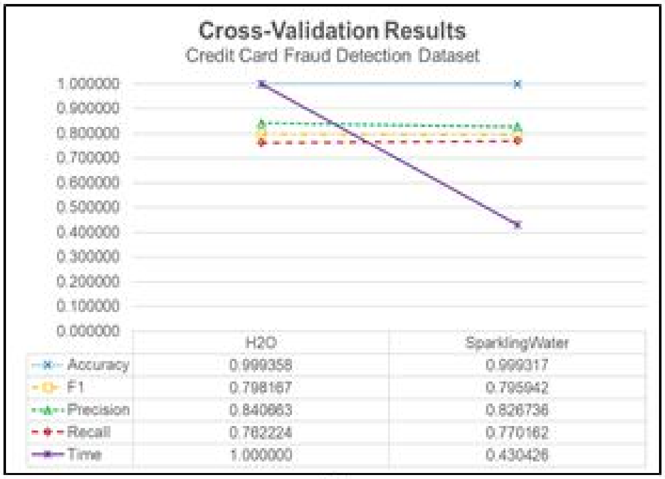
On the contrary of our findings in [21], each platform has performed uniquely on each dataset and both experiments without any consistent separating advantage that can allow one to conclude the superiority of one platform over the other.

Noting that the experiments were conducted on a controlled environment as follow:

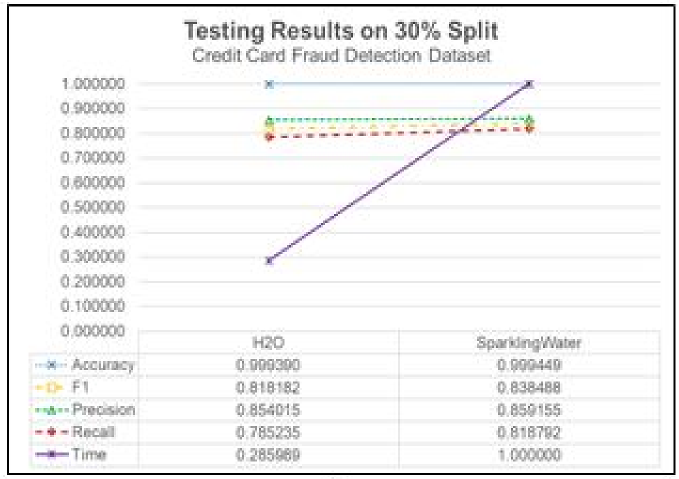
- Running on a hypervisor where network latency is minimal to none.
- The firewall was disabled on all cluster nodes.
- All scheduled tasks were disabled prior running the experiment to avoid CPU delay.
- Only the two platforms and their dependencies were installed on the nodes.
-

Which leaves the question open of why did each platform behave differently while running under the same conditions? The source-code built models, evaluation results and related files are all available for interested researchers at the project repository on GitHub[3].

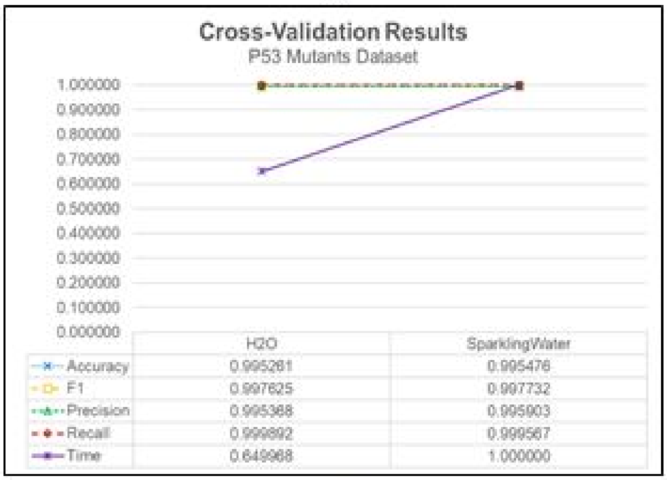
[3] https://github.com/alzewairi/H2O_vs_SparklingWater



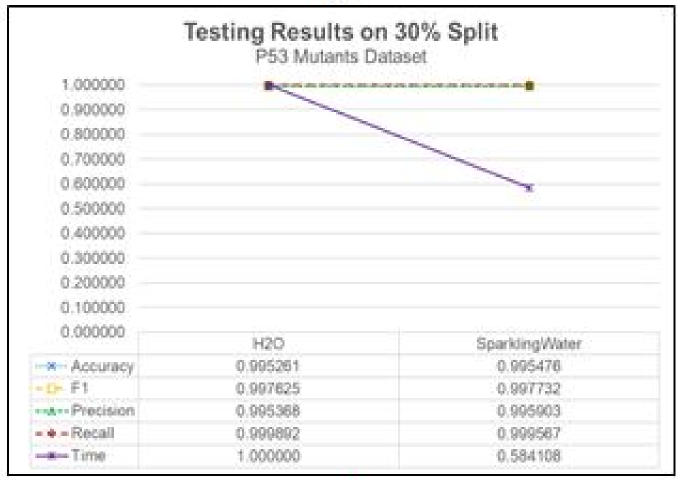
(c)



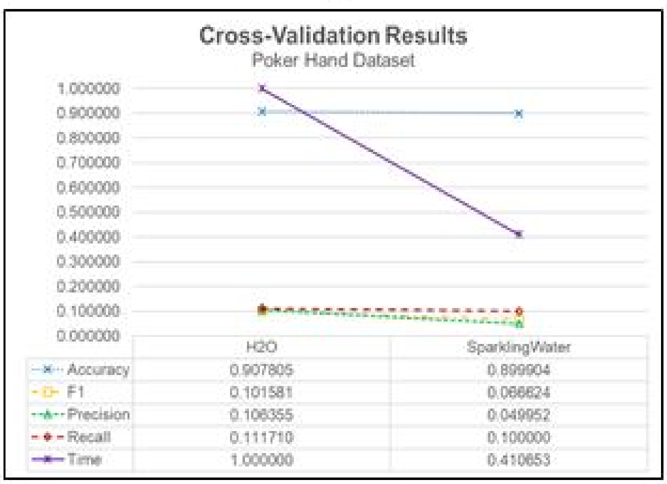
(d)



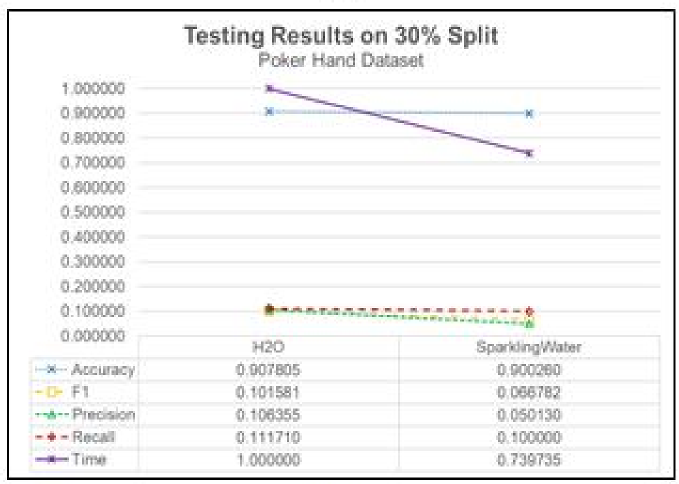
(e)



(f)



(g)



(h)

Figure 1. The evaluation results for the first experiment where the model was trained using a 3-folds cross-validation technique and tested on 30% unseen data. (a) Cross-Validation Results on Santander Bank Dataset. (b) Results of Testing on 30% Unseen Data on Santander Bank Dataset. (c) Cross-Validation Results on Credit Card Fraud Detect Dataset. (d) Results of Testing on 30% Unseen Data on Credit Card Fraud Detect Dataset. (e) Cross-Validation Results on P53 Mutants Dataset. (f) Results of Testing on 30% Unseen Data on P53 Mutants Dataset. (g) Cross-Validation Results on Poker Hand Dataset. (h) Results of Testing on 30% Unseen Data on Poker Hand Dataset.

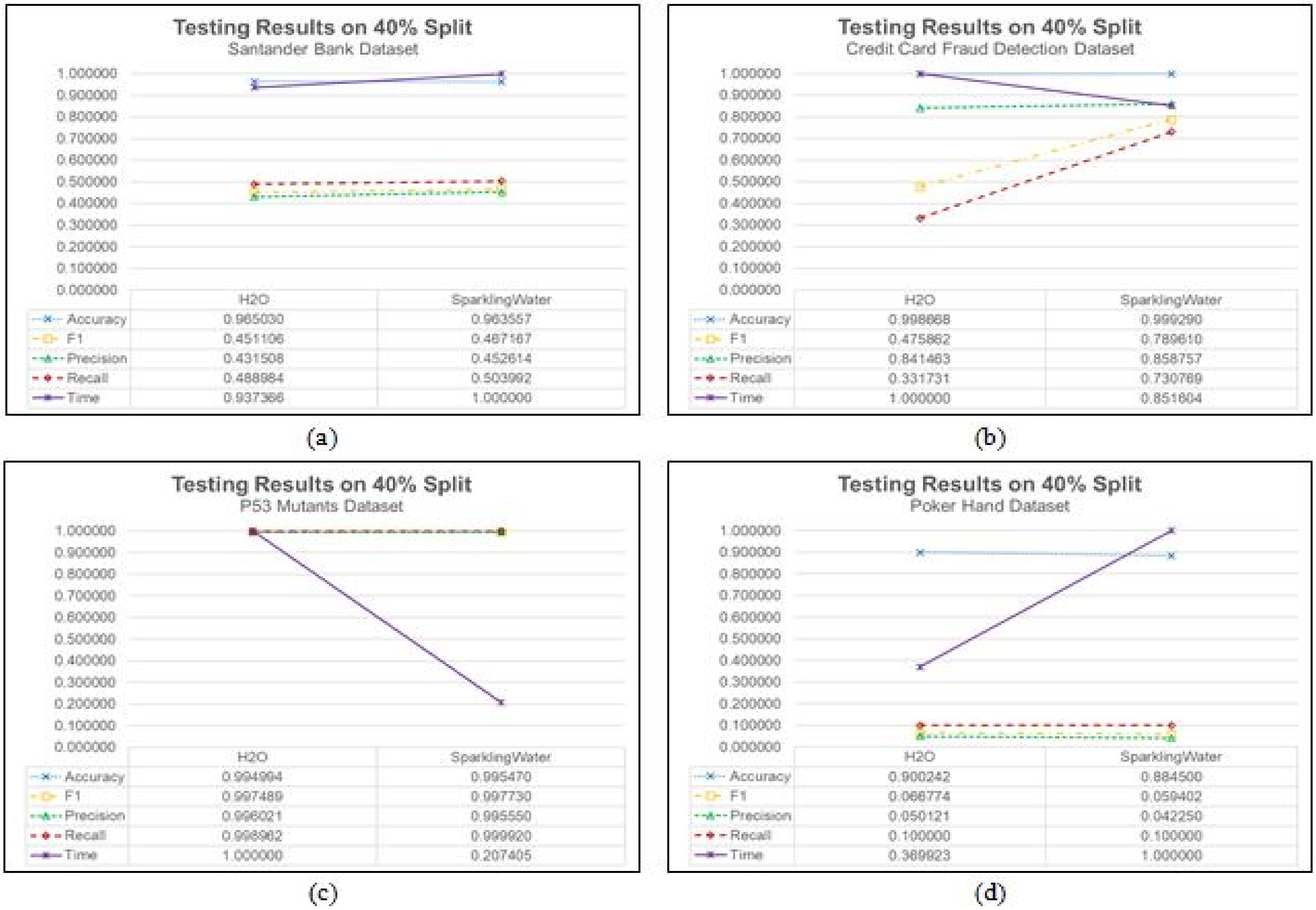


Figure 2. The evaluation results for the second experiment where the model was trained on 60% of the data and tested on 40% unseen data. (a) Results of Testing on 40% Unseen Data on Santander Bank Dataset. (b) Results of Testing on 40% Unseen Data on Credit Card Fraud Detect Dataset. (c) Results of Testing on 40% Unseen Data on P53 Mutants Dataset. (d) Results of Testing on 40% Unseen Data on Poker Hand Dataset.

5. CONCLUSION

Several machine learning platforms and toolkits are built to analyze big data used in various domains such as business, scientific and medical applications. In this paper, we evaluated the efficiency of two big data analysis machine learning platforms called H2O and SparklingWater.

. In this paper, experimental comparisons were made between the two platforms by comparing their accuracy, AUC, f1-score, precision, recall, specificity, and training time to solve the public prediction challenge. Experiments were performed using four different datasets: the Santander Bank dataset, Credit Card Fraud Detection Dataset, P53 Mutants Dataset, and Poker Hand dataset. The experimental results showed that the two platforms obtained convergent results in terms of accuracy, f1-score, precision, recall, and specificity with the SparklingWater platform marginally exceeded the H2O platform in terms of model accuracy. However, in terms of model training time, H2O has obtained a significant performance. For future work, further experiments will be performed using a variety of datasets to reduce bias.

REFERENCES

1. Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). **A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data**, 2(1), p.p. 24. https://doi.org/10.1186/s40537-015-0032-1
2. D. S. Char and A. Burgart, "**Machine-Learning Implementation in Clinical Anesthesia: Opportunities and Challenges**," Anesthesia & Analgesia, vol. 130, no. 6, pp. 1709–1712, Jun. 2020, doi: 10.1213/ANE.0000000000004656.

3. Gupta, S. (2016, August 22). **Deep Learning vs. traditional Machine Learning algorithms used in Credit Card Fraud Detection (masters)**. Dublin, National College of Ireland. Retrieved from http://trap.ncirl.ie/2495/
4. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). **Deep learning applications and challenges in big data analytics. Journal of Big Data**, 2(1), p.p. 1. https://doi.org/10.1186/s40537-014-0007-7.
5. D. Suleiman and G. Naymat, "Sms spam detection using h2o **framework**, " Procedia Computer Science, Vol. 113, (2017), 154-161, doi:10.1016/ j.procs.2017.08.335.
6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). **Deep learning. Nature**, 521(7553), p.p. 436–444. https://doi.org/10.1038/nature14539.
7. White, T. (2015). **Hadoop: The Definitive Guid**e (4th Edition). Retrieved from http://shop.oreilly.com/product/0636920033448.do
8. Abhishek, S. (2015). **Big Data and Hadoop. Java COE** www.marlabs.com. Retrieved from http://www.marlabs.com/sites/default/files/Marlabs-White Paper-BigData-Hadoop.pdf.
9. Arora, A., Candel, A., Lanford, J., LeDel, E., & Parmar, V. (2015, August). **Deep Learning with H2O. H2O.ai, Inc**. Retrieved from https://h2o-release.s3.amazonaws.com/h2o/master/3190/docs-website/h2o-docs/booklets/DeepLearning_Vignette.pdf
10. G. Nguyen et al., "**Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey**," Artif Intell Rev, vol. 52, no. 1, pp. 77–124, Jun. 2019, doi: 10.1007/s10462-018-09679-z
11. Kejela, G., Esteves, R.M., Rong, C., 2014. **Predictive Analytics of Sensor Data Using Distributed Machine Learning Techniques**, in: 2014 IEEE 6th International Conference on Cloud Computing Technology and Science. Presented at the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 626631. doi:10.1109/CloudCom.2014.44
12. Grolinger, K., Capretz, M.A.M., Seewald, L., 2016. **Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources. Electrical and Computer Engineering Publications,** 2016.
13. Ha, V.-S., Nguyen, H.-N., 2016. **Credit scoring with a feature selection approach based deep learning**. MATEC Web of Conferences 54, 5004. doi:10.1051/matecconf/20165405004.
14. MiÅkuf, M., Zolotov, I., 2016. **Comparison between multi-class classifiers and deep learning with focus on industry 4.0**, in: 2016 Cybernetics Informatics (KI). Presented at the 2016 Cybernetics Informatics (K I), pp. 15. doi:10.1109/CYBERI.2016.7438633.
15. Wakita, Y., Oku, K., Kawagoe, K., 2016. **Toward Fashion-Brand Recommendation Systems Using Deep-Learning: Preliminary Analysis**. ResearchGate 2, 128131. doi:10.18178/ijke.2016.2.3.066.
16. Uppu, S., Krishna, A., Gopalan, R.P., 2016b. **A Deep Learning Approach to Detect SNP Interactions**. ResearchGate 11, 965975. doi:10.17706/jsw.11.10.965-975.
17. Zhang, F., Wang, Y., Cao, M., Sun, X., Du, Z., Liu, R., Ye, X., 2016. **Deep-Learning-Based Approach for Prediction of Algal Blooms**. Sustainability 8, 1060. DOI:10.3390/su8101060.
18. Uppu, S., Krishna, A., Gopalan, R., 2016a. **Towards deep learning in genome-wide association interaction studies**. PACIS 2016 Proceedings.
19. W. Etaiwi, M. Biltawi, and G. Naymat, "**Evaluation of classification algorithms for banking customer's behavior under Apache Spark Data Processing System,**" Procedia Computer Science, vol. 113, pp. 559–564, 2017, doi: 10.1016/j.procs.2017.08.280.
20. A. Baldominos, A. Cervantes, Y. Saez, and P. Isasi, "**A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition using Mobile Devices**," Sensors, vol. 19, no. 3, p. 521, Jan. 2019, doi: 10.3390/s19030521.
21. D. Suleiman, M. Al-Zewairi, and G. Naymat, "**An Empirical Evaluation of Intelligent Machine Learning Algorithms under Big Data Processing Systems**," Procedia Computer Science, vol. 113, pp. 539–544, 2017, doi: 10.1016/j.procs.2017.08.270.
22. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. **Calibrating Probability with Undersampling for Unbalanced Classification**. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015.
23. Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) **Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS Computational Biology**, 5(9), e1000498
24. Cattral, F. Oppacher, D. Deugo. **Evolutionary Data Mining with Automatic Rule Generalization**. Recent Advances in Computers, Computing and Communications, pp.296-300, WSEAS Press, 2002