

Agriculture data analysis using parallel k-nearest neighbour classification algorithm

Vimala Muninarayanappa^{1,2}, Rajevee Ranjan²

¹Department of Agricultural Statistics, Applied Mathematics and Computer Science, University of Agricultural Sciences, Bangalore, India

²School of Computer Science and Applications, REVA University, Bangalore, India

Article Info

Article history:

Received Feb 14, 2023

Revised Sep 12, 2023

Accepted Oct 4, 2023

Keywords:

Artificial intelligence

Crop classification

Data mining

Feature extraction selection

Hyperspectral information

Machine learning technique

ABSTRACT

A cost-effective and effective agriculture management system is created by utilizing data analytics (DA), internet of things (IoT), and cloud computing (CC). Geographic information system (GIS) technology and remote sensing predictions give users and stakeholders access to a variety of sensory data, including rainfall patterns and weather-related information (such as pressure, humidity, and temperatures). They have unstructured format for sensory data. The current systems do a poor job of analysing such data since they cannot effectively balance speed and memory usage. An effective categorization model (ECM) on agriculture management system is proposed to address this research difficulty. First, a classification technique called priority-based k-nearest neighbour (KNN) is provided to categorize unstructured multi-dimensional data into a structured form. Additionally, the Hadoop MapReduce (HMR) framework is used to do classification utilizing a parallel approach. Data from real-time IoT sensors used in agriculture is the subject of experiments. The suggested approach significantly outperforms previous approaches that are computing time, memory efficiency, model accuracy, and speedup.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Vimala Muninarayanappa

School of Computer Science and Applications, REVA University

Rukmini Knowledge Park, Yelahanka, Kattigenahalli, Bengaluru, Karnataka 560064, India

Email: vimalam514@gmail.com

1. INTRODUCTION

To improve the agricultural productivity, it is essential to update the system with data such as yield, crop type, and crop growth conditions along with rainfall pattern data as well as weather related information (such as pressure, humidity, and temperature) time to time. The agro data captured by these sensors is usually in unstructured form and is moved to cloud environment through gateway or internet. For smart agro farming, an effective system is needed for storing, and analysing such unstructured type of data on cloud platforms.

This research sought to address these issues and propose effective categorization model (ECM) methodology. In order to categorise unstructured type of multi-dimension high-dimensional data to structural form, a priority-based k-nearest neighbour (KNN) algorithm is first developed. Additionally, a concurrent categorization approach using the Hadoop MapReduce (HMR) architecture is provided. Figure 1 illustrates the design of a quick and effective agro data classification algorithm for an agricultural management system.

The significance of proposed crop classification technique are as follows. First, a multi-dimension, high-dimensional, unstructured agro data classification system based on priority was developed. Next, a parallel classification approach using the HMR is described. The proposed classification model can perform

analysis considering real-time agro sensory data with good accuracy, reduced time, higher memory efficiency, and speedup.

Because it can analyse enormous volumes of data and extract crucial information, big data (machine learning and deep learning) is used in precision agriculture. For the purpose of monitoring environmental factors on a farm, this project uses internet of things (IoT) technology for intelligent agriculture. Three-dimensional cluster analysis (3D CA) was used to study the environmental factors impacting the farm. The hyperspectral series of images or videos accelerates the rate at which data is generated and the volume at which it is produced, which poses challenges for big data, especially in applications for agricultural remote sensing. We provide an overview of the IoT, big data, and artificial intelligence (AI), as well as how these technologies will impact the agri-food sector in the future [1]–[4]. We undertake an analysis of the most recent research on the application of intelligent data processing technologies in agriculture, particularly in the production of rice. We provide a unified vision for IoT technology, data processing, and practical analytics in digital agriculture. Thanks to coronavirus disease-2019 (COVID-19), more people are now concerned about food safety, which is advantageous for the market share of smart agriculture. Contrary to existing solutions, the framework for integrating and analysing agricultural data from various sources provided in this research uses cloud computing (CC), which improves the solution's scalability, flexibility, affordability, and maintainability [5]–[8].

We thoroughly assess agriculture mobile crowd sensing (AMCS) and offer recommendations for approaches to agricultural data collection. Using a small quantity of ground truth data, this work offered Gaussian kernel regression for estimating rice yield from optical and synthetic aperture radar (SAR) imaging. We provide a unique joint federated learning (FL) model based on partial least squares (PLS) regression and neural networks (NN) (FL-NNPLS). This paper suggested a high-resolution spatiotemporal image fusion approach (HISTIF) made up of multiplicative modulation of temporal change (MMTC) and filtering for cross-scale spatial matching (FCSM). First, we evaluate the state of industrial agriculture and the takeaways from industrialized agricultural production patterns in this essay [9]–[12]. We start by suggesting an image compression method for data gathering. Initially provide a picture compression method for data gathering. We analyse how close a drone using a long range (LoRa) radio essential fly toward sensors in order to gather the data within a certain level of data quality [13]–[16].

In this study, a brand-new mechanism for automatically defining zones for variable rate application is proposed. In this work, we demonstrate an embedded system enhanced with AI that enables continuous analysis and on-site prediction of plant leaf growth dynamics. Finding the significant technologies towards the advancement of intelligent agriculture that may successfully enhance the production efficiency to ensure the quality of the agricultural yields is done using data visualization analysis along with cluster analysis [17], [18].

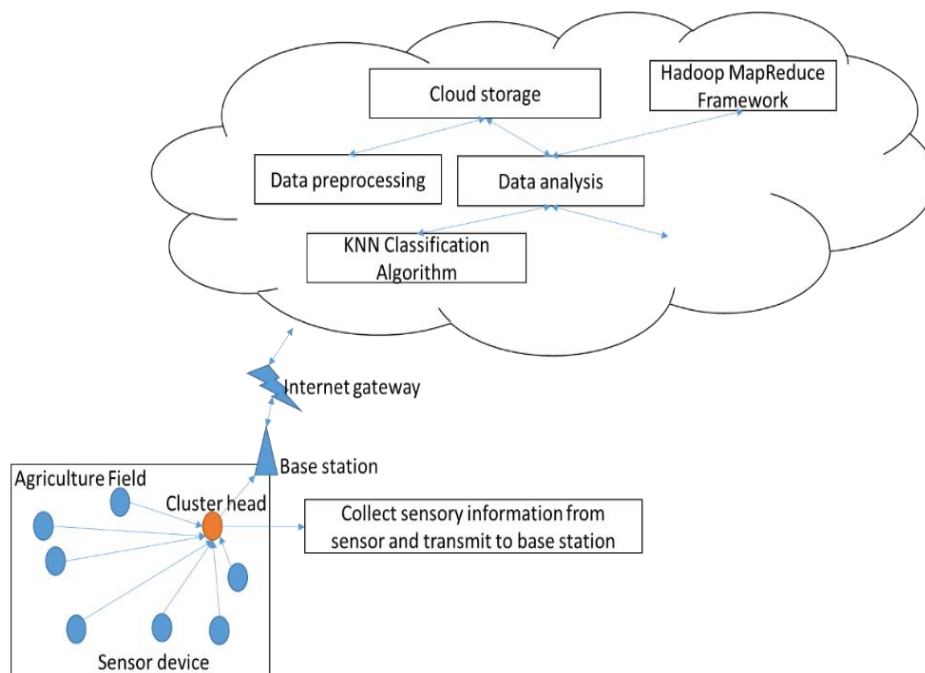


Figure 1. Accurate classification model's architectural design for a multi-level cloud storage concept

The paper is organized as following. In second section of paper provides the efficient classification methodology for analyzing raw unstructured data is presented. In penultimate section, experiment is conducted for evaluating accuracies of classification model is presented. The conclusion of research and future work is defined in last section.

2. PRIORITY-BASED K-NEAREST NEIGHBOR CLASSIFICATION MODEL TO ANALYZE UNSTRUCTURED AGRO DATA

This research provides a quick and effective classification algorithm for analysing unstructured agricultural data and storing it at various cloud storage levels (provider). Agriculture-related unstructured data is classified into structured data, for that a priority KNN algorithm is first introduced. To speed up the classification process for relatively large data, a parallel classification model utilising the HMR framework is then given. Figure 2 shows the block architecture of proposed classification model.

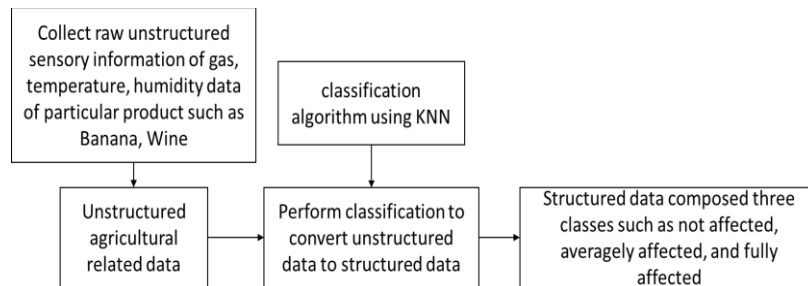


Figure 2. Block architecture of proposed classification model

For analysis or categorization in this work, crop-monitoring datasets gathered from [19] are used. Sensory data acquired from various temperature, humidity, and gas sensors makes up the information. The circumstances under which wine and banana fruits mature are determined using this data. The data comprises 11 attributes or dimensions, including id, time, R1, R2, R3, R4, R5, R6, R7, and R8, as well as temperature and humidity, and is made up of 919,438 data points that are dispersed throughout various locations and periods. The dataset used in this investigation is described in full in [19]. We categorised these data using priority clustering. Set to 3, the K (i.e. we take into consideration three groups, such as not affected, averagely affected, and totally impacted). The K can be modified to meet the criteria for user categorization. This is why we separate the data into three groups and store it in the cloud.

2.1. Clustering model for classifying unstructured raw data into structured data

The suggested priority-based KNN classification model is constructed by utilising k-mean clustering to divide the data points at each stage into L distinct areas. The data points in a location region are iteratively subjected to the same procedure following clustering. When there are less data points in an area than L , the iterative calculation is finished. Algorithm 1 presents the proposed priority-based KNN model.

Algorithm 1. Building priority-based KNN algorithm

Input: Agriculture Dataset E , diverging influence L , maximum iteration J_1 , center selection strategy to be applied D_{str} .

Output: Structured (Classified) data.

```

if  $|E| < L$  then
  build terminal node with feature points in  $E$ .
else
   $Q \leftarrow$  choose  $L$  data points from  $E$  using  $D_{str}$ .
  Converged  $\leftarrow$  false
  Iterations  $\leftarrow$  Zero
  while converged = false && iteration  $<$   $J_1$  do
     $D \leftarrow$  cluster the feature points in  $E$  around closest centers  $Q$ 
     $Q_N \leftarrow$  averages of clusters in  $D$ 
    if  $Q = Q_N$  then
      Converged  $\leftarrow$  true
    end if
  end while
   $Q \leftarrow Q_N$ 

```

```

iterations←iteration + 1
end while
for each cluster  $D_j \in D$  do
build non-terminal node with center  $Q_j$ 
Continuously apply clustering method to the feature points in  $D_j$ 
end for
end if

```

The algorithm's feature or attribute known as the diverging influence is the number of clusters L that should be taken into account while separating the data at each node, and choosing L is important for achieving a successful classification conclusion. J_{\max} , which represents the maximum clustering iterations, is another parameter of the priority-based KNN clustering method. Smaller iterations can speed up clustering at the expense of accuracy. Finally, yet importantly, the parameter D_{str} is utilised to govern the initial centres selection in the clustering algorithm. The suggested priority-based KNN clustering, however, achieves good convergence with minimal time. The raw input data used to perform classification is displayed in Figure 3.



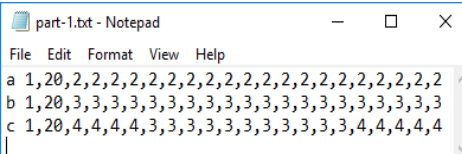
```

20D.txt - Notepad
File Edit Format View Help
1,1,20,0.622202995788401,0.714188919963403,1.26630484075579,1.59789882916394,5.60752121593222,5.25692852105337,3.64875449623855,2.44471695535105,4.39299242325266,4.320
85475998039,5.03900120307645,5.20457446245224,3.76644669767769,5.30509568699873,4.98645473500549,0.795237525713275,4.05968089421265,0.671586638885358,1.48078415180127,
3.85740185703496
2,1,20,2.49779381093001,4.57073263259639,4.32998606361448,4.22107234584683,3.23750648511458,2.3766404676934,1.18018015180723,0.618083922443019,3.78344912486777,2.34062
274321011,0.834369993025609,2.000397320823,0.13875370836293,0.729723845948337,1.55837021273019,4.57838619546424,4.70467559487777,5.64591279200554,2.42115184534395,1.21
440837701057
3,1,20,1.78483111565226,2.38275629171299,2.97213264203264,0.310095165493011,0.436561507646256,2.602778326549,4.86986984544893,0.49198671214375,1.42300435311767,2.86993
14070353,2.48144231969558,5.93961460818053,0.110870295674945,0.44693396267804,5.23726502690803,2.05147036319201,2.38420285634892,3.04523249490896,4.36039234404936,3.24
088074823882
4,1,20,4.08674836924614,3.04067057379553,4.66126783048327,2.06236968004255,3.49648476695478,3.04860388225345,5.07582075408931,3.04672285172936,3.75672547426854,0.41806
8539541247,2.09056193083551,4.41376810220711,4.25043479354141,0.276375774790708,2.08257074138269,0.125037519217021,1.60859999056375,4.02989089812147,3.79451083820058,0
.644728158565577
5,1,20,3.50677062224446,0.698193439663478,5.19507956932535,0.923517717832475,5.85918042351454,0.89475538484508,1.89952702514805,5.69181974568489,1.49245881904031,5.506
89984953818,2.20418333911532,1.66944309996881,1.29552044795617,5.18027109710512,3.12766000776443,4.74416054185674,3.72148846122971,0.91976296199009,2.81361039853823,3.
3879576253928
6,1,20,3.5269575245969,1.83314699371958,2.73526123787521,1.01493581357642,0.789369476199788,4.06767421748846,1.69917981463912,5.83449452088424,4.36436322129535,5.1222
27507691,0.452870372628267,4.35431195822745,5.25072165670838,3.72566418548378,5.73411012599436,3.39855492958266,3.84559100447017,3.7477501126508,2.20318925652801,2.611
7293821097
7,1,20,1.1590768571054,5.38400100840442,1.56426403268905,4.24982562473036,4.21388520042593,0.431118830866701,3.12824765425094,5.07016369643629,3.14871576702628,3.33841
019952596,2.40354905798731,0.698837028573657,1.22628049958324,0.416553212807771,3.11777058610589,1.68362362852489,3.96293220178826,4.19014375574894,0.144148241632687,1
.80150620602607
8,1,20,5.19557557020596,2.06503250426381,1.33745652438954,2.37023131134931,5.06402947690991,3.84987138088787,2.41872144326042,3.56985626247239,5.44616097111076,5.72910
507818177,0.908569537386563,5.97373834857891,4.71657370546207,0.211881016428527,2.23888076828741,4.14295656405527,0.278882961030529,4.48543561314486,2.80105061315049,0
.368250645333086
9,1,20,0.949426563237527,2.62901611820283,2.69577875385796,1.03595302176939,2.7286784497037,5.75403334405461,3.13803145864421,4.84416844390061,0.617893599377896,4.0604
8649746947,5.4356227277944,5.65544724872124,3.61360187101811,0.411683050501944,3.41615381190374,5.09404435284158,4.05654448401488,5.19403272133974,5.18579614782044,0.3
74013819538063
10,1,20,4.01239738760628,5.55839139177855,5.62400272532553,3.41961068884451,2.43984066213939,2.4298107094084,1.98226332311158,1.02439948315471,4.17767244452037,4.97481
402205062,1.41474066749436,5.47749658864341,1.39909314259845,5.7758975962996,0.862079642788544,2.57655945200313,3.36925375040586,0.311503502620153,5.64155286164561,5.3
099961328553
11,1,20,0.721839922243655,3.57455299935981,2.9619626630945,1.39567447364129,0.303482350387369,1.25312228866533,2.3701364775608,4.31489065496013,2.79572720452479,3.1448
3641544582,0.096412080040934,1.08485023969079,5.29963946532441,2.43703679361243,2.35585325772681,1.39103517145898,3.00501339287731,5.27616806501344,4.59611235963,0.30
88677402953
12,1,20,3.78177002094303,2.1236319754895,0.676495999380246,1.64307957732728,2.65574582997511,5.95812613915099,4.26635410612279,3.83770427471851,5.5496034077134,0.84959
4359080613,5.72479060243573,3.75504098139472,5.55247985969413,5.87580001496514,5.65217389729441,0.447844388246464,2.60642872868405,4.23832825168424,3.12612833845714,1.
1967183734707
13,1,20,0.0696742318475965,3.66737266815145,4.22867227862993,1.04283602907454,4.88840194201022,0.33989042780357,0.187857123318993,4.96205634898602,3.43004433857279,5.4
7104447132956,5.57154605912582,0.103085685392416,1.71539114299016,5.34268512135031,1.538226116191227,4.59820794675416,2.90805699997957,1.32888308576722,4.73992864366617

```

Figure 3. Raw input dataset used for performing classification operation

From Figure 3 it is visible the raw data is composed of 20-dimension point, which is generated similar to [19], [20]. The complexity of computation mainly dependent on dimension size rather than size of data (rows). Classification is carried out to identify least affected (i.e. class a), averagely affected (i.e. class b) and most affected (i.e. class c) under assumption described in Figure 4. The outcome of classification model is shown in Figure 5.



```

part-1.txt - Notepad
File Edit Format View Help
a 1,20,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
b 1,20,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3
c 1,20,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4

```

Figure 4. Classification input data for performing classification operation

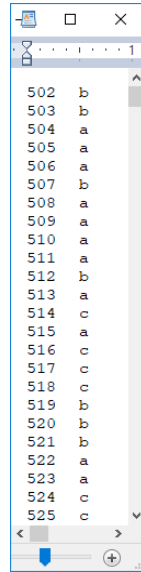


Figure 5. Classification outcome attained using priority-based KNN classification

2.2. Parallelizing classification using HMR framework

Additionally, this paper proposes a parallel classification system that makes use of the HMR framework [20]. Figure 6 depicts the HMR framework's fundamental design. Since HMR follows the execute-once paradigm, all state data for iterative execution should be put into distributed file system (DFS) and then read back in for each stage of algorithm calculation or evaluation. HMR is a widely used software model for MR computations that is accessible to the public (i.e. it is open source in nature).

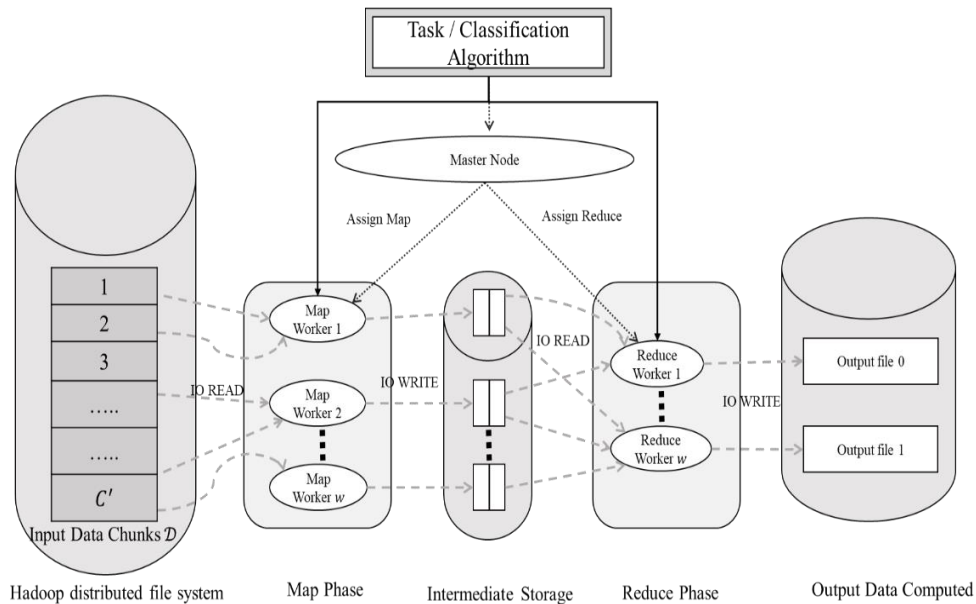


Figure 6. The architecture of HMR framework

2.3. Advantage of HMR

Hadoop [21] is a distributed computing framework designed using java programming language adopting cloud-computing environment, which supports the MR architecture as shown in Figure 7. HMR has execute-once paradigm, implying that with iterative execution strategy all state data should be written into DFS and after that read back in for each progression of the algorithm calculation or evaluation. HMR is a

publicly available software model (i.e. it is open source in nature), and broadly utilized for MR calculations. Owen *et al.* [22] has been worked to keep running over HMR and the Hadoop distributed file system [23], [24]. Hadoop distributed file system (HDFS) is an execution of the google file system (GFS) where an extensively large dataset is fragmented into equal length of small blocks and a duplicate copy of each blocks maintained (this process is known as data replication). While handling the information, the framework pushes calculations to the virtual computing nodes where these chunks are facilitated to expand information location awareness amid computing for quicker algorithm computation makespan. At the point when HMR is initiated with HDFS, HMR can exploit information location awareness and push calculations to the information they should work on, eliminating the systems or network administration overhead, which might be caused when collecting from HDFS. This may offer the HMR based usage an edge in computing overheads when contrasted with other distributed and parallel processing architecture.

2.4. Parallel classification algorithm for Hadoop MapReduce framework

HMR is a combination of two important functions known as map and reduce as shown in Figure 7. Map function takes input data of same domain and generate list of pair value of result in different domain $M(key_1, val_1) \rightarrow l(key_2, val_2)$. This created key key_2 a list of various values that was combined, and a reducer function. The reducer function uses the intermediate key key_2 and the values to create a new set of values called $l(val_3)$.

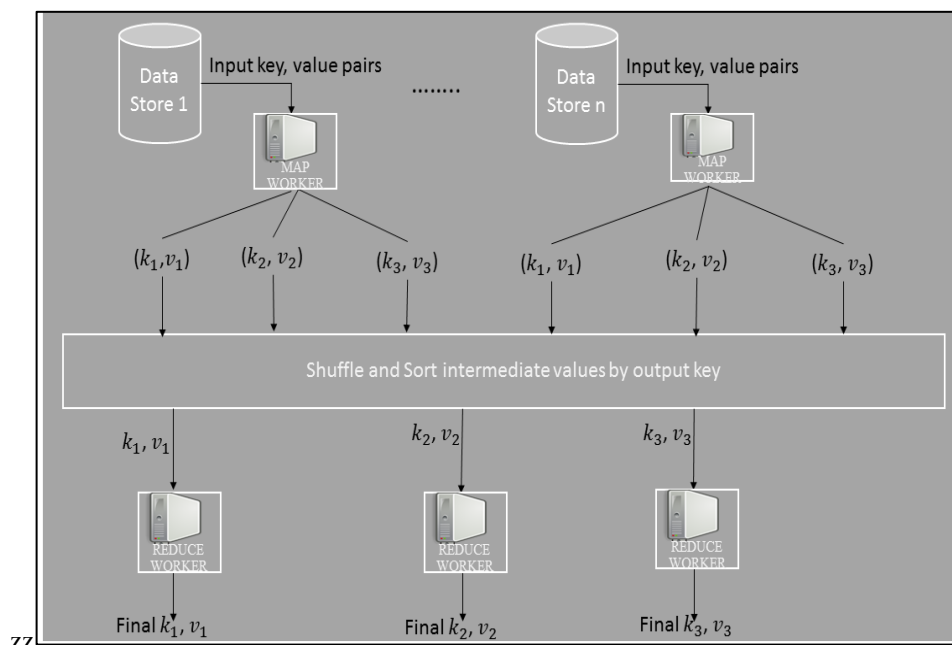


Figure 7. HMR computation model

In Hadoop, distributed system MapReduce job execution performed on multiple system or machine. Where one is master nodes and other is worker nodes or known as slave nodes. Master node distribute task among the worker nodes. Each slave nodes has fixed number of mapper and reducer function or can be called as map and reduce slots. Worker nodes periodically send their free or engaged map or reduced slots detail to the master node. Master nodes schedule the task based on availability of mapper reducer function in the cluster.

The MapReduce function combines the tasks of mapping and reducing. The input dataset is divided into uniformly sized blocks of data, which are then distributed among the nodes of the Hadoop cluster. Applying a user-defined mapper function to the input from the map task results in intermediate output that serves as data for the reduce task's input. Reduced stage combines reduction phase and two-phase shuffle. The output data to the map job is used as an input into the shuffle phase, where the already completed map task is shuffled and then sorted. The sorted data is now sent into the user-defined reduce function, and the output is written back into HDFS. A map stage involves several distinct map tasks, each of which is listed.

Reduce stage is combination of shuffle/sort and reduce phases. In reduce stage shuffle/sort phase start working only after the first map task completed. Working of shuffle phase completed after the all map

task work is completed. Once the shuffle/sort work over reduce task start working. Shuffle phase result obtained in first cycle may differ from result obtained in 2nd cycle. Result of shuffle phase varies due to dependency on Map cycle. Reduce shuffle phase measurement based on two reduce cycle one is called initial shuffle and other is called typical shuffle. Reduce phase begins once the shuffling phase is finished [20]. Provides information on HMR operation details. The Hadoop HDInsight cluster's distributed key building technique is displayed in Algorithm 2. This work uses distributed architecture to classify agricultural data, and our model achieves good accuracy, reduces computing time, and satisfies the real-time requirement, as empirically demonstrated in the next section.

Algorithm 2. Building distributed Key on Hadoop HDInsight cluster

Input: Data E , keyVal Q

Output: $ConstructKey(E, Q)$

$j \leftarrow MR_function()$

E_j read chunk of the data E with respect to function j using Hadoop distributed file system.

construct key in parallel on each worker with data E_j and keyVal Q

$MR_Cumulate()$ // Synchronize all workers.

3. RESULT AND DISCUSSION

This section compares the proposed effective categorization model (ECM) approach to the current approach [25] and evaluates how well it performs in terms of speedup, accuracy, central processing unit (CPU) time, as well as memory overhead. The information is used to determine how temperature and humidity affect the effects of gases on wine and bananas. In general, spreading sensor devices around the agricultural area improves yield. The sensors keep a look on conditions such as temperature along with humidity and make decisions depending on them, such as whether to release water or use pesticides, among other things. Additionally, by keeping an eye on the wind, which helps predict the onset of rain, cyclones, and other weather events in a specific location with less delay, agriculture production can be improved. So that the right decision can be made at the right moment with the least amount of harm to the crops. To assess the performance in terms of memory and time efficiency when taking into account real-time agrosensor dataset received from [19] such as Inspiral, this work compares with previous technique [26]. This research is carried on the Windows 10 operating system (OS) along with I-7 processor (64-bit). The memory use in this research is 16 GB RAM along with 4 GB GPU dedicated with compute unified device architecture (CUDA) support. One master worker node and four slave worker nodes are taken into account while designing the HDInsight cluster utilizing the database Azure HDInsight cluster and A3.

An experiment was carried out to evaluate the performance of ECM with the existing models [25], [26] in terms of total CPU time, memory overhead, and accuracy attained in generating classification trees for turning unstructured input into structured data. Table 1 shows the comparison along with several state of art approach for developing classification tree. Table 2 lists the results of this evaluation. The outcome demonstrates that artificial neural network (ANN) performs better than a random categorization model. Figure 8 shows the classification performance assessment considering different dimension size. We contrast the proposed outcome performance improvement to the ANN classification model therefore. While decreasing overall CPU time and memory overhead by 32.85% and 55.07%, respectively, the ECM-local classification model improves accuracy by 1.82%. Additionally, the ECM-Hadoop classification model obtains a speedup of 16, increases accuracy by 1.82%, decreases overall CPU time and memory overhead by 95.86% and 84.05%, respectively. Additionally, we assessed how dimension size affected classification ability, as shown in Figure 8. As shown in Table 2, we modified the dimension size to be 4, 6, 8, and 10 and assessed the classification result in terms of total CPU time, accuracy, and memory overhead. The results of the experiment demonstrate that when dimension size rises, computation time and memory overhead also increase. Similar to this, precision is achieved when dimension size is 5 and increases to 11 to get accuracy of 2.17. This makes it obvious that the size of the dimension affects categorization accuracy. The entire outcome demonstrates the ECM model's scalable performance in comparison to state-of-the-art models.

Table 1. Comparison along with several state of art approach for developing classification tree

	Random [7]	ANN [7]	ECM-Local	ECM-Hadoop
Total CPU time (s)	129.69	52.5	35.25	2.37
Average accuracy	0.977	0.971	0.989	0.989
Memory overhead (kilobytes)	0.71	0.69	0.31	0.11
Speedup	14	14	-	16

Table 2. Classification performance assessment by considering different dimension size

Dimension size	Total CPU time (s)	Average accuracy	Memory overhead (kilobytes)
4	1.79	0.983	0.09
6	1.86	0.986	0.099
8	1.93	0.987	0.106
10	2.17	0.989	0.11
Average	1.9375	0.986	0.101

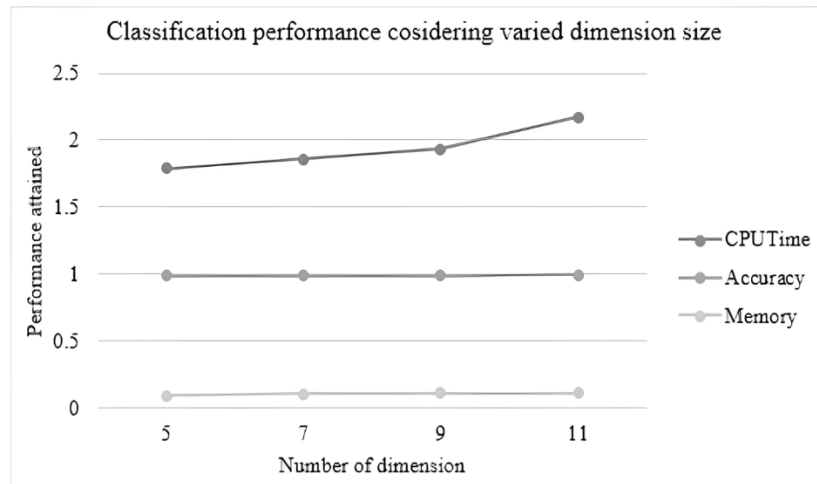


Figure 8. Classification performance assessment considering different dimension size

4. CONCLUSION

From the above research, we can establish an efficient classification technique regarding the performance analysis based on agro related data in unstructured form. Here a priority-based KNN classification model is presented, which performs the analysis on multi-dimensional data (high dimensional data). Here we have adopted a distributed computing framework for the analysis purpose. Parallel clustering algorithm approach by applying Hadoop framework is developed for establishing scalable performance during analysis of high dimensional data. All the research are carried out on real-time data scrapped from agro sensors. Further, the results display that the ECM-local reduces the total CPU time as well as memory overhead by 32.85% along with 55.07% respectively. Here the accuracy improves by 1.82%. Likewise, the ECM-Hadoop model for classification decreases the total CPU time by 95.86% along with memory overhead by 84.05% respectively. Here the accuracy is improved by 1.82% and the speedup enhances to 16. The overall performance result displays the scalable performance of developed ECM model when compared with several state-of-art paradigms on several parameters such as total CPU time as well as accuracy and memory efficiency along with speedup. Further, the future research would consider evaluating considering different dataset and minimize the storage and processing cost.




REFERENCE

- [1] S. A. Bhat and N.-F. Huang, "Big data and AI revolution in precision agriculture: survey and challenges," *IEEE Access*, vol. 9, pp. 110209–110222, 2021, doi: 10.1109/ACCESS.2021.3102227.
- [2] F.-H. Tseng, H.-H. Cho, and H.-T. Wu, "Applying big data for intelligent agriculture-based crop selection analysis," *IEEE Access*, vol. 7, pp. 116965–116974, 2019, doi: 10.1109/ACCESS.2019.2935564.
- [3] K. L.-M. Ang and J. K. P. Seng, "Big data and machine learning with hyperspectral information in agriculture," *IEEE Access*, vol. 9, pp. 36699–36718, 2021, doi: 10.1109/ACCESS.2021.3051196.
- [4] N. N. Misra, Y. Dixit, A. Al-Mallahi, M. S. Bhullar, R. Upadhyay, and A. Martynenko, "IoT, big data, and artificial intelligence in agriculture and food industry," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6305–6324, May 2022, doi: 10.1109/IIOT.2020.2998584.
- [5] R. Alfred, J. H. Obid, C. P.-Y. Chin, H. Haviluddin, and Y. Lim, "Towards paddy rice smart farming: a review on big data, machine learning, and rice production tasks," *IEEE Access*, vol. 9, pp. 50358–50380, 2021, doi: 10.1109/ACCESS.2021.3069449.
- [6] S. Chaterji *et al.*, "Lattice: a vision for machine learning, data engineering, and policy considerations for digital agriculture at scale," *IEEE Open Journal of the Computer Society*, vol. 2, pp. 227–240, 2021, doi: 10.1109/OJCS.2021.3085846.
- [7] J. Song, Q. Zhong, W. Wang, C. Su, Z. Tan, and Y. Liu, "FPDP: flexible privacy-preserving data publishing scheme for smart agriculture," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17430–17438, Aug. 2021, doi: 10.1109/JSEN.2020.3017695.
- [8] A. Goldstein, L. Fink, and G. Ravid, "A cloud-based framework for agricultural data integration: a top-down-bottom-up approach," *IEEE Access*, vol. 10, pp. 88527–88537, 2022, doi: 10.1109/ACCESS.2022.3198099.




- [9] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc fuzzy C means and auto-encoder CNN," in *Inventive Computation and Information Technologies*, vol. 563, S. Smys, K. A. Kamel, and R. Palanisamy, Eds., in Lecture Notes in Networks and Systems, vol. 563., Singapore: Springer Nature Singapore, 2023, pp. 353–368. doi: 10.1007/978-981-19-7402-1_25.
- [10] Y. Alebele *et al.*, "Estimation of crop yield from combined optical and SAR imagery using gaussian kernel regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10520–10534, 2021, doi: 10.1109/JSTARS.2021.3118707.
- [11] D. Vimalajeewa, C. Kulatunga, D. Berry, and S. Balasubramaniam, "A service-based joint model used for distributed learning: application for smart agriculture," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2022, doi: 10.1109/TETC.2020.3048671.
- [12] J. Jiang *et al.*, "HISTIF: a new spatiotemporal image fusion method for high-resolution monitoring of crops at the subfield level," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4607–4626, 2020, doi: 10.1109/JSTARS.2020.3016135.
- [13] Y. Liu, X. Ma, L. Shu, G. P. Hancke, and A. M. Abu-Mahfouz, "From Industry 4.0 to agriculture 4.0: current status, enabling technologies, and research challenges," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4322–4334, Jun. 2021, doi: 10.1109/TII.2020.3003910.
- [14] S. Nesteruk *et al.*, "Image compression and plants classification using machine learning in controlled-environment agriculture: antarctic station use case," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17564–17572, Aug. 2021, doi: 10.1109/JSEN.2021.3050084.
- [15] A. Caruso, S. Chessa, S. Escolar, J. Barba, and J. C. Lopez, "Collection of data with drones in precision agriculture: analytical model and LoRa case study," *IEEE Internet Things Journal*, vol. 8, no. 22, pp. 16692–16704, Nov. 2021, doi: 10.1109/JIOT.2021.3075561.
- [16] J. Xu, N. V. Bermeo, M. Zheng, D. Langton, M. O'Grady, and G. M. P. O'Hare, "Automated zone identification for variable-rate services in precision agriculture," *IEEE Access*, vol. 9, pp. 163242–163252, 2021, doi: 10.1109/ACCESS.2021.3134488.
- [17] D. Shadrin, A. Menshchikov, A. Somov, G. Bornemann, J. Hauslage, and M. Fedorov, "Enabling precision agriculture through embedded sensing with artificial intelligence," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4103–4113, Jul. 2020, doi: 10.1109/TIM.2019.2947125.
- [18] J. Chen and A. Yang, "Intelligent agriculture and its key technologies based on internet of things architecture," *IEEE Access*, vol. 7, pp. 77134–77141, 2019, doi: 10.1109/ACCESS.2019.2921391.
- [19] F. Huerta and R. Huerta, "Gas sensors for home activity monitoring data set," 2016, Accessed: Jul. 26, 2018. [Online]. Available: <http://archive.ics.uci.edu/ml/datasets/Gas+sensors+for+home+activity+monitoring>.
- [20] "Apache Hadoop," The Apache Software Foundation, 2006. Accessed: Oct. 21, 2017. [Online]. Available: <http://hadoop.apache.org>.
- [21] T. White, *Hadoop: The definitive guide*, 1st ed. O'Reilly Media, Inc., 2009.
- [22] S. Owen, B. E. Friedman, R. Anil, and T. Dunning, *Mahout in Action*, Manning Publications, 2011.
- [23] D. Borthakur, "The hadoop distributed file system: architecture and design," The Apache Software Foundation, 2007, Accessed: Jul. 26, 2018, [Online]. Available: https://svn.apache.org/repos/asf/hadoop/common/tags/release-0.16.3/docs/hdfs_design.pdf.
- [24] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008, doi: 10.1145/1327452.1327492.
- [25] L. Verdoliva, D. Cozzolino, and G. Poggi, "A reliable order-statistics-based approximate nearest neighbor search algorithm," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 237–250, Jan. 2017, doi: 10.1109/TIP.2016.2624141.
- [26] L. Wan, Q. Cao, F. Wang, and S. Oral, "Optimizing checkpoint data placement with guaranteed burst buffer endurance in large-scale hierarchical storage systems," *Journal of Parallel and Distributed Computing*, vol. 100, pp. 16–29, 2017, doi: 10.1016/j.jpdc.2016.10.002.

BIOGRAPHIES OF AUTHORS



Vimala Muninarayanappa    is a research scholar at School of Computer Science and Applications, REVA University, Bangalore. She has master's degree in computer applications from R V College of Engineering. She is currently an assistant professor at Department of Agricultural Statistics, Applied Mathematics and Computer Science, University of Agricultural Sciences, Bangalore. Her area of interest includes wireless sensor networks, IoT, and cloud computing. She can be contacted at email: vimalam514@gmail.com.



Dr. Rajeev Ranjan    after completing a Ph.D. in wireless sensor network at Indian Institute of Information Technology, Allahabad (IIIT-A), he is associate professor in the School of Computer Science and Applications at REVA University, Bangalore. His area of work includes wireless sensor networks-coverage and connectivity, sensor deployment and localization, IoT, and wireless sensor statistical routing. He can be contacted at email: rajeev.ranjan@reva.edu.in.