scientific reports

OPEN

MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network

Surendran Rajendran[1]✉, Osamah Ibrahim Khalaf[2], Youseef Alotaibi[3] & Saleh Alghamdi[4]

In recent times, big data classification has become a hot research topic in various domains, such as healthcare, e-commerce, finance, etc. The inclusion of the feature selection process helps to improve the big data classification process and can be done by the use of metaheuristic optimization algorithms. This study focuses on the design of a big data classification model using chaotic pigeon inspired optimization (CPIO)-based feature selection with an optimal deep belief network (DBN) model. The proposed model is executed in the Hadoop MapReduce environment to manage big data. Initially, the CPIO algorithm is applied to select a useful subset of features. In addition, the Harris hawks optimization (HHO)-based DBN model is derived as a classifier to allocate appropriate class labels. The design of the HHO algorithm to tune the hyperparameters of the DBN model assists in boosting the classification performance. To examine the superiority of the presented technique, a series of simulations were performed, and the results were inspected under various dimensions. The resultant values highlighted the supremacy of the presented technique over the recent techniques.

Big data is a group of information that cannot be processed, captured, and managed by means of traditional software implements in a particular interval. This can be a higher growth rate, huge and differentiated data resource[1]. It needs novel processing modes to have strong decision-making control, understanding, optimization capability, and discovery powers. Big data analysis is performed by many approaches, such as databases, web crawlers, analysis, data warehouses, big data mining, and visualization algorithms. Among others, web crawlers are a widespread analysis technology[2]. They could extract text data/numerical values from a webpage and set up it to analyze information. In addition to Python language, this technique could be executed with few software packages. The big data mining algorithm needs to manage a huge amount of information. Furthermore, the speediness of handling is deliberate. This algorithm includes optimization technologies (such as GDA and PSO algorithms), scientific modeling or other predictive technologies (such as NN, NB, roughset, DT), decision analysis technologies (such as multicriteria decision, gray decision, and so on), performance assessment technologies (for example, fuzzy comprehensive evaluation and data envelopment analyses). Since this mining information was hardly examined and big data in the historical, particularly used to engineer problems, it has is a very big problem while employing this mining information algorithm to big data[3].

Clustering and classification are the two key classes of algorithms in mining information[4]. Nevertheless, the performances of classification and clustering methods are considerably caused by the increasing dataset dimension because the algorithm in this category operates on the dataset dimension. Additionally, the drawback of higher dimension datasets includes redundant data, higher module construct time, and degraded quality, which makes information analyses highly complicated. To resolve these problems, the selection of features is employed as a major preprocessing phase for choosing subsets of features from a large dataset and increases the accuracy of clustering and classification models, which triggers foreign, ambiguous, and noisy data elimination[5]. The FS

[1]Center for Artificial Intelligence and Research (CAIR), Chennai Institute of Technology, Chennai, India. [2]Al-Nahrain Nanorenewable Energy Research Center, Al-Nahrain University, Baghdad, Iraq. [3]Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia. [4]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia. ✉email: surendran.phd.it@gmail.com

method depends upon search techniques and a performance assessment of subsets. Since the preprocessing stage, feature selection was essential for removing duplications, minimizing the amount of information, and irrelevant and unnecessary characteristics. It has various techniques to select the feature that assists in choosing the actual dataset as the effective feature. Filter, embedded and wrapper are the 3 approaches of the FS model[6]. The selection of features should achieve 2 aims: to eliminate/reduce the amount of FS and increase the output performance. As already mentioned, meta heuristics in previous decades simulate organisms' collective behaviors. In particular, this algorithm has generated an important development in several regions associated with optimization[7]. The optimal selection is made by a metaheuristic algorithm; in a rational interval, the cloud generates better solutions[8]. Sometimes, it is a better solution to mitigate the limitation of comprehensive time-consuming searches[9]. Various metaheuristic methods, alternatively, suffer from the optimum location, missing search multiplicity and imbalance among exploitative and explosive performances[10]. Recently, EA has shown itself to be efficient and attractive for solving challenges using optimizations. There are few approaches, such as PSO, CSA[11], GA[12], and ACO algorithms[13]. PSO was hybridized for constant search space issues in the works with another metaheuristics approach.

This study focuses on the design of a big data classification model using chaotic pigeon inspired optimization (CPIO)-based feature selection with an optimal deep belief network (DBN) model[14]. The proposed model is executed in the Hadoop MapReduce environment to manage big data. Initially, the CPIO algorithm is applied to select a useful subset of features. In addition, the Harris hawks optimization (HHO)-based deep belief network (DBN) model is derived as a classifier to allocate appropriate class labels. The design of the HHO algorithm to tune the hyperparameters of the DBN model assists in boosting the classification performance. To examine the superiority of the presented technique, a series of simulations were performed, and the results were inspected under various dimensions. This paper structure is defined as follows. In "Results" section, the limitation of the proposed research work was identified through a literature review. Hadoop map reduction models defined in "Discussion" section. In "Methods" section, the performance analysis of the proposed system is briefly elaborated. This is followed by conclusions and some probable future directions are recommended in "Conclusion" section. In Al-Thanoon et al., BCSA was stimulated by natural phenomena to perform the FS method. In BCSA, the flight length variable plays a significant part in the performance[15]. To enhance the classification performances by rationally elected features, a development of defining the flight length parameters through the concepts of the opposition-based learning method of BCSA is presented. BenSaid and Alimi proposed an OFS method that solves these problems[16]. The presented method named MOANOFS explores the new developments of the OML method and conflict resolution method (Automated Negotiation). MOANOFS employs a 2-decision level. Initially, deciding k(s) among the learner (OFS method) is trustful (trust value/higher confidence). This selected k learner will take part in the next phase in which this presented MANOFS technique is incorporated.

In Pooja et al., the TC-CMECLPBC method is projected. Initially, the features and data were collected from large climate databases[17]. The TCC model is employed to find the comparison among the features to select appropriate features through high FS precision. The clustering method consists of 2 stages, maximization (M) and expectation (E), for discovering the maximal likelihood of grouping information into clusters. Next, the clustering results are provided to linear program boosting classifiers to improve the predictive performance. Lavanya et al. examine FS techniques such as rough set and entropy on the sensor's information[18]. Additionally, a representative method of FW is presented that represents Twitter and sensor information effectively for additional analyses of information. Few common classifications, such as NB, KNN, SVM, and DT, are employed to validate the efficiency of the FS. An ensemble classification method is presented that is related to many advanced methods. In Sivakkolundu and Kavitha, a new BCTMP-WEABC method is presented to predict upcoming outcomes with high precision and less time consumption[19]. This method includes 2 models, FS and classifier, to handle a large amount of information. Baldomero-Naranjo et al. proposed a strong classification method depending upon the SVM method that concurrently handles FS and outlier discovery[20]. The classifiers are made to consider the ramp loss margin error and involve budget constraints for limiting the number of FSs. The search of classifiers is modeled by a mixed integer design through a large M parameter. Two distinct methods (heuristic and exact) are presented for solving this method. The heuristic method is authenticated by relating the quality of the solution given to this method using an accurate method.

Guo et al. proposed a WRPCSP method for executing the FS method. Next, the study incorporates BN and CBR systems for reasoning knowledge[21]. According to the possible reasoning and calculation, WRPCSP algorithms and BN permit the presented CBR scheme to work in big data. Furthermore, to solve these problems created with a large number of features, this study also proposed a GO method for assigning the computation process of big data for similar data processing. Wang et al. proposed a big data analytics approach to the FS process for obtaining each explanatory factor of CT, which will shed light on the fluctuations of CT[22]. Initially, the relative analyses are executed among all 2 candidate factors through mutational data metrics to construct the experiential system. Next, the system deconvolutions are explored to infer the direct dependencies among the candidate's factors and the CT by eliminating the effect of transitive relationships from the system. In Singh and Singh, the 4-phase hybrid ensemble FS method was proposed. Initially, the datasets are separated by the cross validation process[23]. Next, several filter approaches that depend on the weight score are ensembles for generating a rating of features, and then the consecutive FS method is used as a wrapper method for obtaining the best subsets of features. Finally, the resultant subsets are treated for the succeeding classifier's task. López et al. proposed a distributed feature weight method that accurately estimates feature significance in a large dataset with the popular method RELIEF in smaller problems[24]. The solution named BELIEF integrates new redundant removal measures that generate schemes related to this entropy but with low time costs. Furthermore, BELIEF provides a smoother scale-up, while additional cases are needed to increase the accuracy of the estimation.

Results

In this study, a new big data classification model is designed in the MapReduce environment. The proposed model derived a novel CPIO-based FS technique, which extracts a useful subset of features. In addition, the HHO-DBN model receives the chosen features as input and performs the classification process. The detailed working of these processes in the Hadoop MapReduce Environment is offered in the following sections. An essential component that develops the Hadoop architecture is as follows: Hadoop Distributed File System (HDFS): Initially, it can be a Google File System. This component was a distributed file system utilized as distributed storage to data; additionally, it gives access to information with maximum throughput. Hadoop YARN (MRv2): This component has been responsible for job scheduling and managing cluster resources. Hadoop MapReduce: Initially, Google's MapReduce, these modules are scheme dependent upon YARN to the parallel process of data.

There are several studies compared with Hadoop, namely, Mahout, Hive, Hbase, and Spark. The most essential feature that describes Hadoop is which the HDFS is a maximum fault tolerance to hardware failure. Certainly, it can be capable of repeatedly handling and resolving these cases. Additionally, HDFS can interface among the nodes going to cluster to manage the data, i.e., to rebalance them[25]. The model of data storing the HDFS was carried out using the MapReduce structure. While Hadoop has expressed mainly in Java and C languages, it can be near several other programming languages. The MapReduce structure allows separation of nodes going to cluster the task, which is also finished. An essential disadvantage of Hadoop is the absence of execution capable of real-time tasks. However, it could not be a vital restriction because of these particular features, another technology is utilized. MapReduce automatically parallelizes and applies the program to a large cluster of commodity technologies. It mechanism by break model as to 2 stages, the map as well as reduce stage. All the stages are key-value pairs as input as well as output, this kind of that can be elected as programmer. The map and reduces operations of MapReduce are combined and determined in terms of data structured form (key, value) pairs. The calculation obtains the group of input key-value pairs and makes the group of output key-value pairs. The map as well as reduce operations in Hadoop MapReduce is the subsequent common procedure:

$$map : (k_1, v_1) \rightarrow list(k_2, v_2)$$

$$reduce : (k_2, list(v_2)) \rightarrow list(v_2)$$

Design of CPIO-FS technique.

At this stage, the CPIO-FS technique is applied to derive a subset of features. PIO has been a newly developed bioinspired technique extremely utilized for solving the optimized issue. It can be dependent upon social organisms of swarms that utilize the base of learning. It tried to mathematically enhance the solution quality based on the natural performance of the swarm to adapt to the position and velocity of all individuals. The homing nature of pigeons derives from 2 important functions, mapping and compassing and landmark functions. Map and compass operator: At this point, the rules were explained as the place X_i and velocity V_i of pigeon i, and the position and velocity in the D-dimensional search space were maximized at each iteration. The new place and velocity of pigeons i at the t-th iteration have been defined utilizing the provided function in Eqs. (1)–(2):

$$V_i(t) = V_i(t-1) \cdot e^{-Rt} + rand \cdot \left(X_g - X_i(t-1)\right) \tag{1}$$

$$X_i(t) = X_i(t-1) + V_i(t) \tag{2}$$

where R stands for the mapping and compassing factors, rand demonstrates the arbitrary value in [0–1], and X_g illustrates the present global optimum place, which is obtained in the evaluation of every location[26].

Landmark operator: During this metric, the partial amount of pigeons is diminished from every generation. To accomplish the target immediately, residual pigeon flies to the destination place[27]. Let X_c be the middle place of pigeons, and the place upgrading rule of pigeon i at the t-th iteration has been written as Eqs. (3)–(5):

$$N_p(t) = \frac{N_p(t-1)}{2} \tag{3}$$

$$X_c(t) = \frac{\sum X_i(t) \cdot fitness(X_i(t))}{N_p \cdot \sum fitness(X_i(t))} \tag{4}$$

$$X_i(t) = X_i(t-1) + rand \cdot (X_c(t) - X_i(t-1)) \tag{5}$$

where N_p refers to the amount of pigeons, but the fitness is the cost function of pigeons. To reduce optimization, the target function has been elected from the rate of minimum.

Optimal feature selection process.

The FF objective is a terminology utilized to estimate the solution. The FF evaluates the solution, which is a subset of obvious features, by means of the true positive rate (TPR), false positive rate (FPR), and number of features. The number of features comprises FF, and there are features obtainable without affecting TPR or FPR. During these cases, it can be necessary to eliminate individual features. Equation (6) schemes the function executed to estimate the fitness of the pigeon or solution.

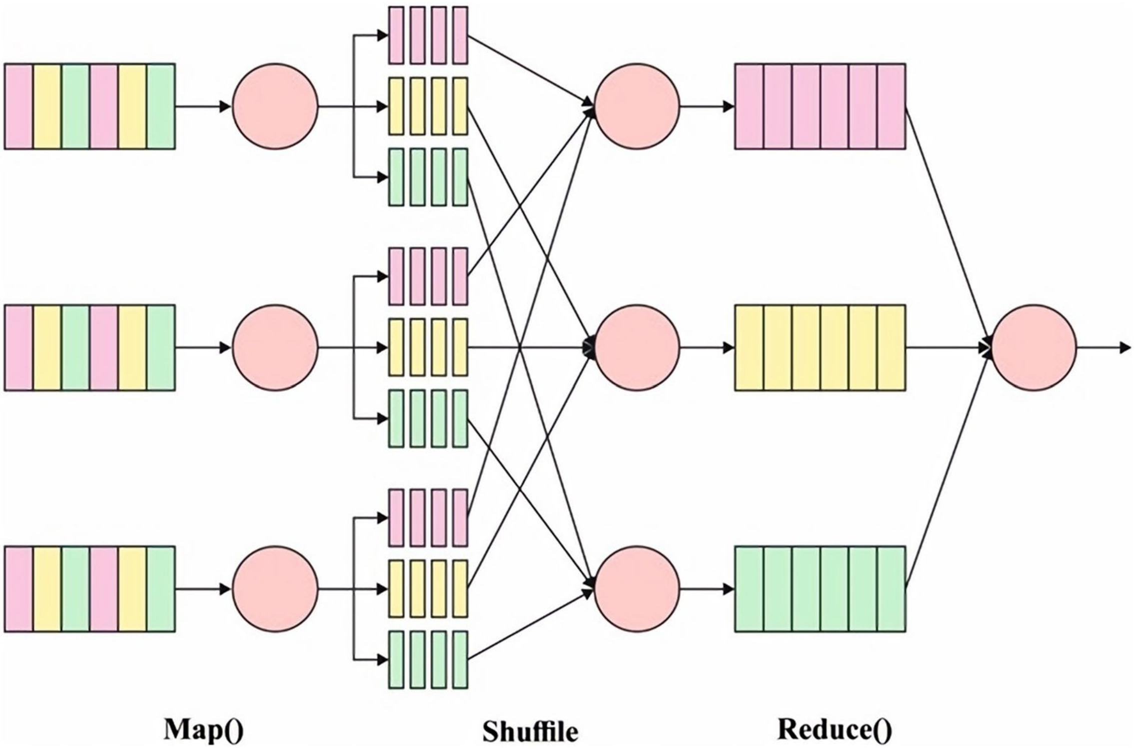

Figure 1. Framework of Mapreduce in Hadoop. It consists of map as well as reduce operations. The map stage obtains the various inputs and spilt and shuffled it. During this reduction stage, all reduce tasks procedure the input in-between data with a decrease function and create output data. Figure illustrates the block diagram of MapReduce.

$$FF = w_1 * \frac{SF}{NF} + w_2 * FPR + w_3 * \frac{1}{TPR} \tag{6}$$

where SF determines the number of elected features, NF refers to the entire feature and $w_1 + w_2 + w_3 = 1$. The measures of weights are set as $w_1 = 0.1$ and $w_2 = w_3 = 0.45$ because TPR and FPR are important. An initial method of projected PIO-FS defines the outcome or pigeon by vector with length, which is similar to feature count. As the PIO techniques fundamental approach acts on the place of pigeon regularly, the proposed PIO for FS explains the solution inference as vector of a definite length, but the measures of place and velocity vectors are arbitrary values in 0 and 1.

Usually, the velocity of pigeons is monitored by a sigmoidal function that has been utilized to transfer velocity as a binary version by utilizing Eq. (7). To resolve the binarized SI technique, the pigeon place is upgraded depending upon the sigmoid function value and the possibilities of arbitrarily uniform values in 0 and 1 by Eq. (8). The residual manner was operated similar to convention PIO except for the upgrading place of landmark operators. In addition, the sigmoid function was implemented transmission, and the velocity and place were upgraded as:

$$S(V_i(t)) = \frac{1}{1 + e^{\frac{-v_i}{2}}} \tag{7}$$

$$X(t)_{(i,p)}[i] = \begin{cases} 1, if\,(S(V_i(t)) > r) \\ 0, otherwise \end{cases} \tag{8}$$

where $V_i(t)$ stands for the pigeon velocities from iteration t and r refer to the uniformly arbitrary values.

Discussion

The Design of HHO-DBN Model is discussed here. The features are passed into the DBN model to perform the classification process. The DBN was a probability generation approach that is opposite the classic discriminative approach. This network is a DL technique that is stacked by RBM and trained in a greedy approach. The resultant prior layer was utilized as the input of the succeeding layer. Eventually, the DBN network has been generated. In the DBN, hierarchical learning has been simulated as the framework of the human brain. All the layers of the deep network are regarded as a logistic regression (LR) approach. The joint distribution function of x and h^k in Layer l is in Eq. (9).

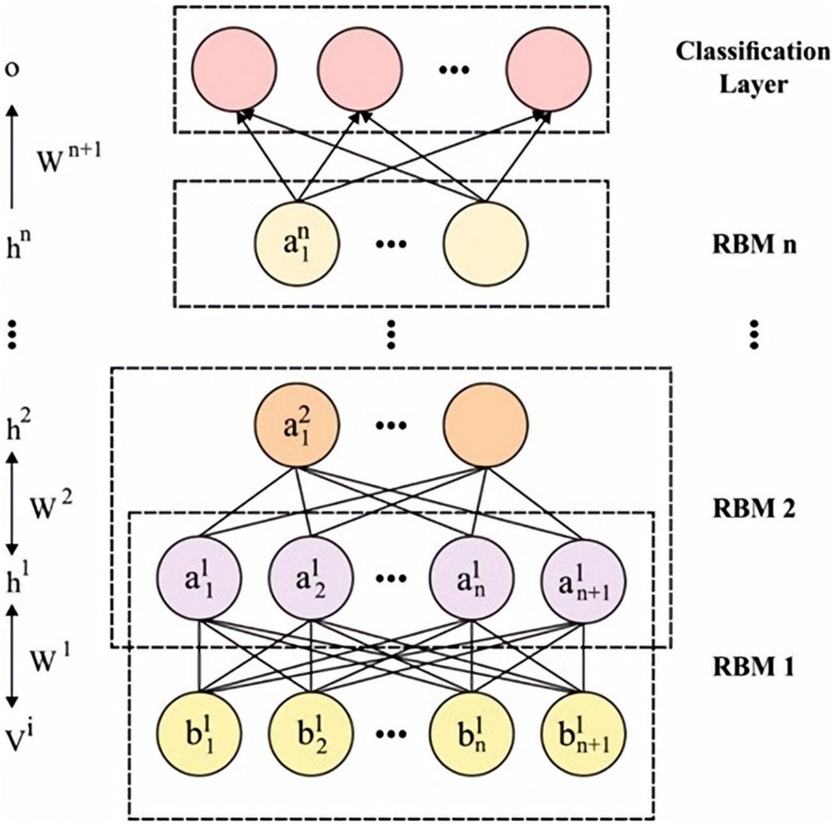

Figure 2. DBN structure. The hyperparameters of the DBN model take place using the HHO algorithm to boost the classifier results. During the HHO, the candidate solution is the Harris hawks, and an optimum/global solution has been planned prey. Therefore, HHO illustrates the exploratory and exploitative phases.

$$p\left(x, h^1, h^2, \ldots, h^l\right) = \left(\prod_{k=0}^{l-2} P\left(h^k | h^{k+1}\right)\right) P\left(h^{l-1}, h^l\right). \tag{9}$$

Input data of the DBN method compose the 2D vector reached in preprocessing. The RBM layers were trained one-by-one in pretrained. The following visible variable is a duplicate hidden variable from the preceding layer[28]. The parameter is transmitted in a layerwise approach, and the features have been learned in the preceding layer. The LR is maximum layers trained by fine-tuning, where the cost function has been revised using BP for optimizing the weight w. The 2 steps are contained in the procedure of trained a DBN technique. All the RBM layers are unsupervised trained, input has been mapped as to distinct feature space, and data has been saved about feasible. Afterward, the LR layer was added on top of the DBN as supervised classification. Figure 2 demonstrates the framework of the DBN model.

The HHO technique is simulated as the hunting performance of Harris hawk to rabbit. If the rabbits have considerable energy, the Harris hawk explores the definitional domain [LB, UB] with the subsequent formula in Eq. (10):

$$x_i(t+1) = \begin{cases} x_r - r_1|x_r - 2r_2x_i(t)| & q \geq 0.5 \\ (x_b - x_m) - r_3(LB + r_4(UB - LB)) & q < 0.5 \end{cases} \tag{10}$$

where $x_i(t+1)$ implies the place of the ith individual from the next iteration of t, x_r refers to the place of an arbitrarily elected candidate at the present iteration, and x_b and x_m are optimum and averaged places in/of the swarms. r_1, r_2, and r_3 are 3 arbitrary numbers from the Gauss distribution. q implies the chance of an individual following that most 2 manners, which represents which it can also be an arbitrary number[29]. The energy of rabbits, signified as symbol E, declines linearly in the maximal value to 0[30–32] in Eq. (11):

$$E = 2E_0\left(1 - \frac{\tau}{\max Iter}\right) \tag{11}$$

where E_0 refers to the primary phase of energy that also fluctuates in the interval of [0, 1]. $maxIter$ stands for the maximal permitted iteration number that is set up at the start. When $|E| < 1$, the Harris hawk is a manner for rabbits with approaches that are explained. τ refers the current iteration.

Dataset	Training instances	Test instances	Features
Epsilon	400,000	100,000	2000
ECBDL14-ROS	65,003,913	2,897,917	631

Table 1. Dataset description.

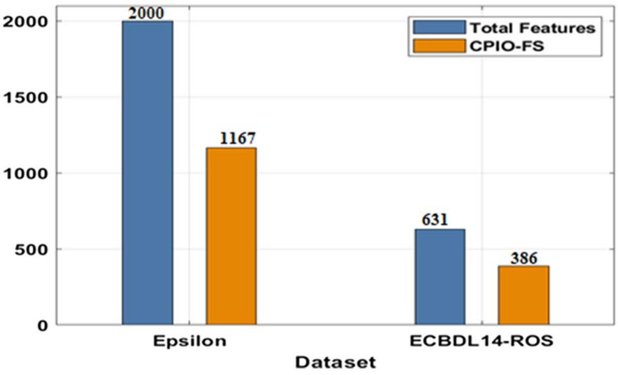

Figure 3. FS analysis of CPIO-FS model. The results depicted that the CPIO-FS technique has chosen an optimal set of features from the applied dataset. For instance, on the test Epsilon dataset, the CIPO-FS technique selected a set of 1167 features out of 2000 features. In addition, on the applied ECBDL14-ROS technique, the CIPO-FS technique has chosen a collection of 631 features out of 386 features.

Soft besiege. *If* $|E| \geq 0.5$ and $r \geq 0.5$, the Harris hawk encloses rabbits softly and scares rabbits to run to make the rabbits tired. During this approach, the Harris hawk is upgrading its places with the subsequent formula in Eq. (12):

$$x_i(t+1) = x_b - x_i(T) - E|\mathrm{J} \cdot x_b - x_i(t)| \tag{12}$$

where $J = 2(1 - r_s)$ signified the arbitrary jumps near the rabbit.

Hard besiege. *if* $|E| < 0.5$, and $r \geq 0.5$, the rabbits were previously exhausted to minimal energy; afterward, the Harris hawk was carrying out hard besiege and made the surprised pounce in Eq. (13).

$$x_i(T+1) = x_b - E|x_b - x_i(t)| \tag{13}$$

Soft besiege with progressive rapid dives. If$|E| \geq 0.5$ and $r < 0.5$, the rabbits are sufficient energy; therefore, the Harris hawks until soft besiege is performed, but in further intelligence, one in Eqs. (14)–(16):

$$x_i(T+1) = \begin{cases} Yf(Y) < F(x_i(t)) \\ Zf(Z) < F(x_i(\mathrm{t})) \end{cases} \tag{14}$$

$$Y = x_b - E|\mathrm{J} \cdot x_b - x_i(t)| \tag{15}$$

$$Z = Y + S \times LF(D) \tag{16}$$

where D implies the dimensionality of the issues, and S refers to the arbitrary number. $LF(x)$ shows the Levy flight from D dimensions in Eq. (17):

$$LF(x) = 0.01 \times \frac{\mu \times \sigma}{|v|^{\frac{1}{\beta}}}, \sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2\left(\frac{\beta-1}{2}\right)} \right)^{\frac{1}{\beta}} \tag{17}$$

Hard besiege with progressive rapid dives. If$|E| < 0.5$ and $r < 0.5$, the rabbits were tired, and Harris hawks were performed hard besiege with intelligence. The influence of the upgrading formulas is similar to Eq. (15), but the middle parameter of Y is altered, which exists to the averages x_m in Eq. (18):

$$Y = x_b - E|J \cdot x_b - x_m| \tag{18}$$

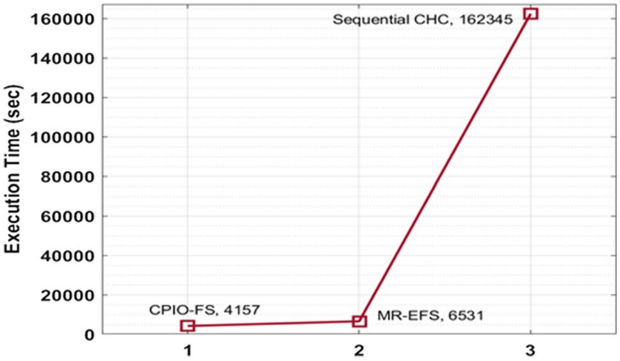

Figure 4. Execution time analysis of CPIO-FS model. The results showed that the CPIO-FS technique obtained an effective outcome with a minimal execution time of 4157 s, whereas the MR-EFS and sequential CHC techniques achieved higher execution times of 6531 s and 162,345 s, respectively.

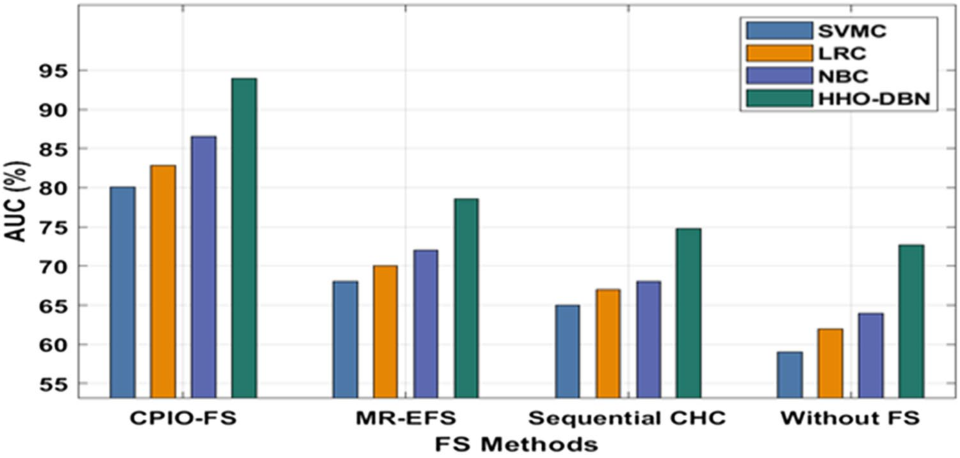

Figure 5. AUC analysis of HHO-DBN model under epsilon dataset. The results ensured that the CPIO-HS with the HHO-DBN model offered the maximum outcome over the other techniques. For instance, with no FS technique, the HHO-DBN model attained a higher AUC of 72.70%, whereas the SVMC, LRC, and NBC techniques attained a lower AUC of 59%, 62%, and 64%, respectively. Following the sequential CHC approach, the HHO-DBN manner reached an increased AUC of 74.80%, whereas the SVMC, LRC, and NBC methods obtained minimal AUCs of 65%, 67%, and 68%, respectively. In line with MR-EFS, the HHO-DBN method reached a superior AUC of 78.60%, whereas the SVMC, LRC, and NBC methodologies achieved a minimal AUC of 68%, 7%, and 72%, respectively. Last, with the CPIO-FS technique, the HHO-DBN methodologies reached a superior AUC of 93.90%, whereas the SVMC, LRC, and NBC approaches attained minimal AUCs of 80.10%, 82.80%, and 86.50%, respectively.

Methods

The evaluation of the experimental results of the presented technique takes place on two benchmark datasets, namely, Epsilon and ECBDL14-ROS. The former dataset has 400,000 samples, whereas the latter dataset has 65,003,913 samples[33], as shown in Table 1. Figure 3 offers the FS outcome of the CIPO-FS technique. Figure 4 demonstrates the execution time analysis of the CPIO-FS technique with existing techniques with 400,000 training instances. Figure 5 investigates the AUC analysis of different classification models under different FS approaches on the epsilon dataset[34,35]. A comprehensive training runtime analysis of different classification models under different FS approaches on the epsilon dataset is provided in Fig. 6. A brief AUC analysis of different classification models under distinct FS methods on the ECBDL14-ROS dataset is provided in Fig. 7. A detailed training runtime analysis of different classification approaches under different FS approaches on the ECBDL14-ROS dataset is provided in Fig. 8.

Conclusion

In this study, a new big data classification model is designed in the MapReduce environment. The proposed model derived a novel CPIO-based FS technique, which extracts a useful subset of features. In addition, the HHO-DBN model receives the chosen features as input and performs the classification process. The design of the HHO-based hyperparameter tuning process assists in enhancing the classification results to a maximum extent. To

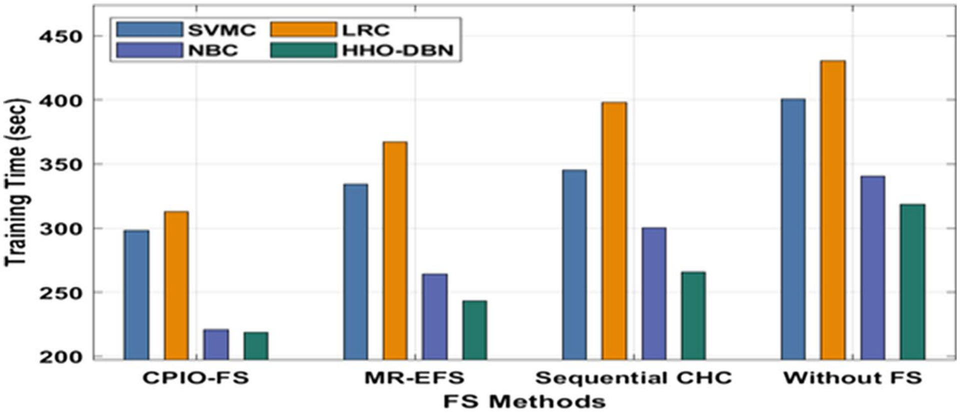

Figure 6. Training time analysis of the HHO-DBN model under the epsilon dataset. The results portrayed that CPIO-FS with the HHO-DBN technique resulted in the least training runtime compared to the other methods. For instance, with no FS, the HHO-DBN technique has gained a lower runtime of 318.54 s, whereas the SVMC, LRC, and NBC techniques have accomplished a higher runtime of 400.38 s, 430.48 s, and 340.42 s, respectively. Moreover, sequential CHC, the HHO-DBN approach has attained a minimal runtime of 265.98 s, whereas the SVMC, LRC, and NBC approaches have accomplished superior runtimes of 345.27 s, 398.07 s, and 300.21 s, respectively. Furthermore, MR-EFS, the HHO-DBN method has reached an increased runtime of 243.09 s, whereas the SVMC, LRC, and NBC methodologies have accomplished a higher runtime of 334.18 s, 367.29 s, and 264.26 s, respectively. Finally, CPIO-FS, the HHO-DBN method has attained a lower runtime of 218.96 s, whereas the SVMC, LRC, and NBC algorithms have accomplished superior runtimes of 298.36 s, 312.78 s, and 220.90 s, respectively.

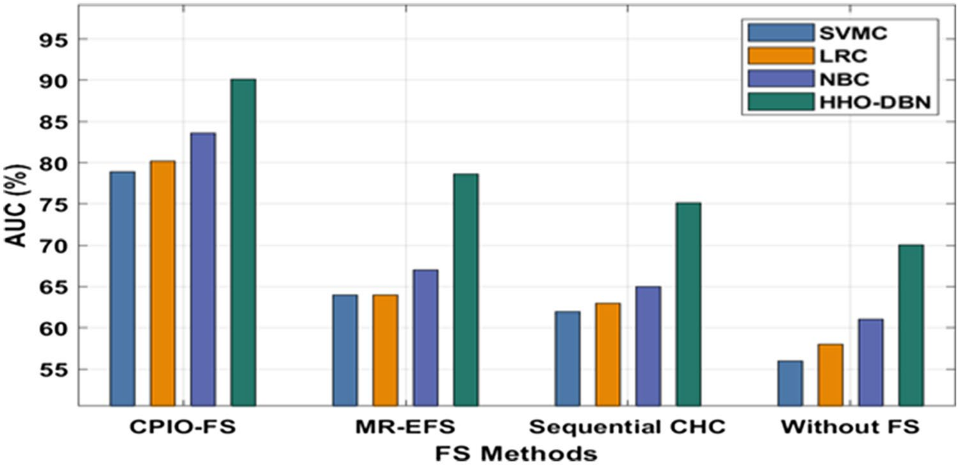

Figure 7. AUC analysis of HHO-DBN model under ECBDL14-ROS dataset. The outcomes demonstrated that CPIO-FS with the HHO-DBN technique resulted in the lowest training AUC compared to the other methods. For instance, with no FS, the HHO-DBN manner has reached a maximum AUC of 70.10%, whereas the SVMC, LRC, and NBC techniques have accomplished a lower AUC of 56%, 58%, and 61%, respectively. Additionally, sequential CHC and the HHO-DBN technique achieved a higher AUC of 75.14%, whereas the SVMC, LRC, and NBC techniques achieved a minimum AUC of 62%, 63%, and 65%, respectively. Likewise, MR-EFS, the HHO-DBN technique, gained an increased AUC of 78.60%, whereas the SVMC, LRC, and NBC techniques accomplished a decreased AUC of 64%, 64%, and 67%, respectively. Finally, CPIO-FS and the HHO-DBN technique achieved a higher AUC of 90.10%, whereas the SVMC, LRC, and NBC techniques achieved minimal AUCs of 78.90%, 80.15%, and 83.60%, respectively.

examine the superiority of the presented technique, a series of simulations were performed, and the results were inspected under various dimensions. The resultant values highlighted the supremacy of the presented technique over the recent techniques. The proposed work utilized a limited number of quality service parameters for the implementation. In the future, hybrid metaheuristic algorithms and advanced DL architectures can be designed to further improve the classification outcomes of big data with an extended number of quality of service parameters.

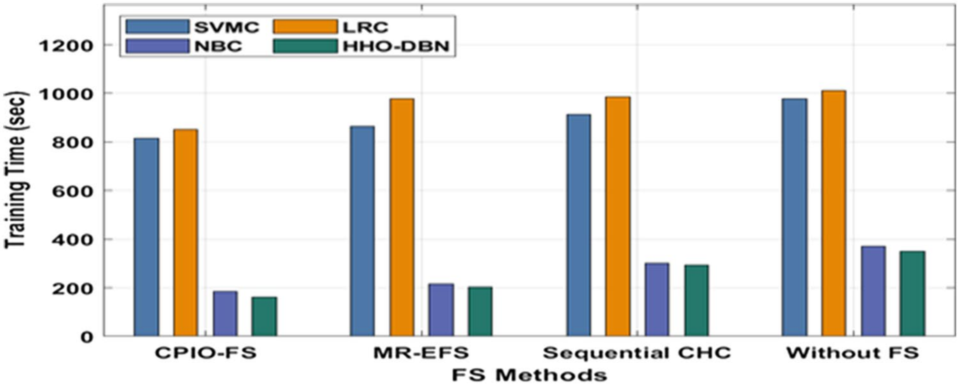

Figure 8. Training time analysis of the HHO-DBN model under the ECBDL14-ROS dataset. The results portrayed that CPIO-FS with the HHO-DBN technique resulted in the least training runtime compared to the other techniques. For instance, with no FS, the HHO-DBN technique has gained a minimal runtime of 350.76 s, whereas the SVMC, LRC, and NBC methods have accomplished superior runtimes of 978.37 s, 1012.47 s, and 369.98 s, respectively. In addition, sequential CHC, the HHO-DBN approach has reached a reduced runtime of 293.87 s, whereas the SVMC, LRC, and NBC algorithms have accomplished superior runtimes of 912.40 s, 986.45 s, and 300.26 s, respectively. In addition, MR-EFS, the HHO-DBN method has reached a lesser runtime of 203.76 s, whereas the SVMC, LRC, and NBC methodologies have accomplished maximal runtimes of 864.28 s, 978.38 s, and 215.09 s, respectively. Last, CPIO-FS, the HHO-DBN manner, obtained a minimal runtime of 160.90 s, whereas the SVMC, LRC, and NBC methodologies accomplished superior runtimes of 815.89 s, 850.53 s, and 184.34 s, respectively.

Data availability

Data are available upon reasonable request for researchers who meet the criteria for access to confidential data.

Received: 21 September 2021; Accepted: 26 November 2021
Published online: 17 December 2021

References

1. Awan, M. J., Rahim, M. S. M., Nobanee, H., Khalaf, O. I. & Ishfaq, U. A big data approach to black Friday sales. *Intell. Autom. Soft Comput.* **27**, 785–797 (2021).
2. El-Hasnony, I. M., Barakat, S. I., Elhoseny, M. & Mostafa, R. R. Improved feature selection model for big data analytics. *IEEE Access* **8**, 66989–67004 (2020).
3. Qiu, M., Kung, S. Y. & Yang, Q. Editorial: IEEE transactions on sustainable computing special issue on smart data and deep learning in sustainable computing. *IEEE Trans. Sustain. Comput.* **4**, 1–3 (2019).
4. Sudhakar Sengan, P., Sagar, V., Khalaf, O. I. & Dhanapal, R. The optimization of reconfigured real-time datasets for improving classification performance of machine learning algorithms. *Math. Eng. Sci. Aerospace* **12**, 1–10 (2021).
5. Zhao, W., Han, S., Meng, W., Sun, D. & Hu, R. Q. BSDP: Big sensor data preprocessing in multisource fusion positioning system using compressive sensing. *IEEE Trans. Veh. Technol.* **68**, 8866–8880 (2019).
6. Emary, E. & Zawbaa, H. M. Feature selection via lèvy antlion optimization. *Pattern Anal. Appl.* **22**, 857–876 (2019).
7. Dhrif, H., Giraldo, L. G. S., Kubat, M. & Wuchty, S. A stable hybrid method for feature subset selection using particle swarm optimization with local search. *Proc. Genet. Evol. Comput. Conf.* **1**, 13–21 (2019).
8. Abdulsahib, G. M. & Khalaf, O. I. Comparison and evaluation of cloud processing models in cloud-based networks. *Int. J. Simul. Syst. Sci. Technol.* **19**, 1–6 (2018).
9. Guo, Y., Chung, F. L., Li, G. & Zhang, L. Multilabel bioinformatics data classification with ensemble embedded feature selection. *IEEE Access.* **7**, 103863–103875 (2019).
10. Al-Khanak, E. N., Lee, S. P., Ur Rehman Khan, S., Verbraeck, A. & van Lint, H. A heuristics-based cost model for scientific workflow scheduling in cloud. *Comput. Mater. Continua.* **67**(3), 3265–3282 (2021).
11. De Souza R. C. T., Coelho L. D. S., De Macedo C. A., & Pierezan J. A V-Shaped binary crow search algorithm for feature selection. in *Proc. IEEE Congr. Evol. Comput. (CEC).* 1–8 (2018).
12. Khan, N. A., Khalaf, O. I., Romero, C. A. T., Sulaiman, M. & Bakar, M. A. Application of Euler neural networks with soft computing paradigm to solve nonlinear problems arising in heat transfer. *Entropy* **23**, 1053. https://doi.org/10.3390/e23081053 (2021).
13. Alsufyani, A., Alotaibi, Y., Almagrabi, A. O., Alghamdi, S. A. & Alsufyani, N. Optimized intelligent data management framework for a cyber-physical system for computational applications. *Compl. Intell. Syst.* **1**, 1–13 (2021).
14. Khan, H. H., Malik, M. N., Alotaibi, Y., Alsufyani, A. & Algamedi, S. Crowdsourced requirements engineering challenges and solutions: A software industry perspective. *Comput. Syst. Sci. Eng.* **39**, 221–236 (2021).
15. Al-Thanoon, N. A., Algamal, Z. Y. & Qasim, O. S. Feature selection based on a crow search algorithm for big data classification. *Chem. Intell. Lab. Syst.* **212**, 104288 (2021).
16. BenSaid, F. & Alimi, A. M. Online feature selection system for big data classification based on multiobjective automated negotiation. *Pattern Recogn.* **110**, 107629 (2021).
17. Pooja, S. B., Balan, R. S., Anisha, M., Muthukumaran, M. S. & Jothikumar, R. Techniques Tanimoto correlated feature selection system and hybridization of clustering and boosting ensemble classification of remote sensed big data for weather forecasting. *Comput. Commun.* **151**, 266–274 (2020).
18. Lavanya, P. G., Kouser, K. & Suresha, M. Effective feature representation using symbolic approach for classification and clustering of big data. *Expert Syst. Appl.* **173**, 114658 (2021).

19. Sivakkolundu, R. & Kavitha, V. Bhattacharyya coefficient target feature matching based weighted emphasis adaptive boosting classification for predictive analytics with big data. *Mater. Today* **5**, 63 (2021).
20. Baldomero-Naranjo, M., Martínez-Merino, L. I. & Rodríguez-Chía, A. M. A robust SVM-based approach with feature selection and outliers detection for classification problems. *Expert Syst. Appl.* **178**, 115017 (2021).
21. Guo, Y., Zhang, B., Sun, Y., Jiang, K. & Wu, K. Machine learning based feature selection and knowledge reasoning for CBR system under big data. *Pattern Recogn.* **112**, 107805 (2021).
22. Wang, J., Zheng, P. & Zhang, J. Big data analytics for cycle time related feature selection in the semiconductor wafer fabrication system. *Comput. Ind. Eng.* **143**, 106362 (2020).
23. Singh, N. & Singh, P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chem. Intell. Lab. Syst.* **1**, 104396 (2021).
24. López, D., Ramírez-Gallego, S., García, S., Xiong, N. & Herrera, F. BELIEF: A distance-based redundancy-proof feature selection method for Big Data. *Inf. Sci.* **558**, 124–139 (2021).
25. Alotaibi, Y. Automated business process modeling for analyzing sustainable system requirements engineering. in *2020 6th International Conference on Information Management (ICIM), IEEE*, 157–161 (2020).
26. Alotaibi, Y. *et al.* Suggestion mining from opinionated text of big social media data. *Comput. Mater. Continua.* **68**, 3323–3338 (2021).
27. Metawa, N., Nguyen, P. T., Nguyen, Q. L. H. T. T., Elhoseny, M. & Shankar, K. Internet of things enabled financial crisis prediction in enterprises using optimal feature subset selection-based classification model. *Big Data.* **9**, 331–342 (2021).
28. Almanaseer, W., Alshraideh, M. & Alkadi, O. A deep belief network classification approach for automatic diacritization of arabic text. *Appl. Sci.* **11**, 5228 (2021).
29. Suryanarayana, G. *et al.* Accurate magnetic resonance image super-resolution using deep networks and Gaussian filtering in the stationary wavelet domain. *IEEE Access* **9**, 71406–71417 (2021).
30. Li, G. *et al.* Research on the natural language recognition method based on cluster analysis using neural network. *Math. Probl. Eng.* **2021**, 1–13 (2021).
31. Alotaibi, Y. A new database intrusion detection approach based on hybrid meta-heuristics. *Comput. Mater. Continua* **66**, 1879–1895 (2021).
32. Rout, R., Parida, P., Alotaibi, Y., Alghamdi, S. & Khalaf, O. I. Skin lesion extraction using multiscale morphological local variance reconstruction based watershed transform and fast fuzzy C-means clustering. *Symmetry* **13**, 2085 (2021).
33. Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A. & Yu, X. Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustain. Cities Soc.* **60**, 102177 (2020).
34. Tian, Z. *et al.* User and entity behavior analysis under urban big data. *ACM/IMS Trans. Data Sci.* **1**, 1–19 (2020).
35. Luo, C. *et al.* A novel web attack detection system for internet of things via ensemble classification. *IEEE Trans. Ind. Inf.* **17**, 5810–5818 (2021).

Acknowledgements

We deeply acknowledge Taif University for supporting this study through Taif University Researchers Supporting Project Number (TURSP-2020/313), Taif University, Taif, Saudi Arabia.

Author contributions

S.R. and S.A. conceived the experiment. Y.A. and O.K. conducted the experiment and performed statistical analysis and figure generation. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.