International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 22s, 2025
https://theaspd.com/index.php

A Hybrid Nature-Inspired and Deep Learning Framework for Feature Selection and Classification in Big Data

K Dharani[1], Dr.G. Abel Thangaraja[2]
[1]Asst. Professor, Department of Computer Science, Sri Nehru Maha Vidyalaya college of Arts and Science, Coimbatore, Tamilnadu, India, dharanikmsc@gmail.com
[2]Asst. Professor, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamilnadu, India, abeltraja@gmail.com

Abstract *– The rapid expansion of Big Data has led to an increasing demand for effective, scalable methods for feature selection and classification. This paper introduces a hybrid framework combining Nature-Inspired Algorithms (NIAs) with Deep Learning techniques to address these challenges. The framework consists of four phases: Phase 1 - data preprocessing using an Enhanced Quality Rules Discovery Model to improve data quality; Phase - 2 feature selection through a Novel Nature-Inspired Algorithm to identify the most relevant features; Phase - 3 classification using a Hybrid Nature-Inspired Algorithm combined with data mining methods; and Phase - 4 advanced classification using Deep Learning Neural Networks integrated with the Hybrid Nature-Inspired Algorithm to boost accuracy. The proposed framework effectively reduces computational complexity while enhancing classification performance on large datasets. The outcomes show how accurate the framework is, which makes it appropriate for a variety of Big Data applications.*

1. INTRODUCTION

In a time when almost all crucial business decisions rely on insights gleaned from data, a significant percentage of firms now view data as a resource. First of all, insufficient data may have many inconsistencies and anomalies, such as inadequate, missing, or incomplete data. These data abnormalities are caused by a variety of factors, including human and specialized factors. This includes data processing, analysis, and visualization. Big Data passes through all phases of its lifecycle. However, these stages will fail in the absence of complete, accurate, and clean data. However, when data isn't fair and suitable for processing, any data processing remains extremely delicate. This results in data and analysis that is worthless due to factors including poor data preparation, the type of data,its arrangement, its starting point, and its sort

this paper centers around working on Quality of the Data in beginning phases before it is consumed for additional processing or analysis. It expects to determine quality issues for huge data sets involving processes as manual endeavors generally fall flat for enormous volumes of data.

Big data is also vulnerable to risks and needs such as data authenticity, theoretical relevance, proper attribute association, controls, audibility, and precision, in addition to data processing. The purpose of these boundaries is to ensure the characteristics of both large data and data. There may be a lot of irrelevant characteristics in the data that can be mined. Therefore, they ought to be removed. Likewise, a lot of features or traits cause many mining computations to perform poorly. In this sense, feature selection methods ought to be used prior to the use of any mining computation.

The primary targets of feature selection are to stay away from overfitting and work on model execution and to give quicker and more practical models. Data holds many features, yet every one of the features may not be connected so the feature selection is utilized to dispose of the inconsequential features from the data absent a lot of loss of the data.

1.1Nature-Inspired Algorithm for Optimization

The nature-inspired algorithm fills in as the foundation of the proposed methodology, giving an effective hunt system to ideal solutions. The nature-inspired algorithm will be adapted to handle the particular difficulties of big data, like high dimensionality, commotion, and versatility. This will guarantee that the algorithm can actually explore the mind-boggling landscape of big data classification and prediction issues.This work presents a new framework that combines the capabilities of Hybrid HarmonySearchwithDifferentialEvolution(HS-DE) and Recurrent Neural Networks (RNNs).

The proposed hybrid approach aims to tackle the challenges of big data, providing a robust and efficient solution for real-world applications across diverse domains. The findings of this study hold the potential to significantly advance the field of big data analytics, enabling accurate and timely predictions that candrive informed decision-making and innovative solutions.

2. LITERATURE SURVEY

1.H. Zhang et al. (2016) suggested investigation and deployment of a Hadoop-based large data pre-processing solution. The Map Reduce programming model is used by popular large data processing platforms to process applications. For example, the structure and estimation mechanism of Hadoop are described as follows: Hadoop gathers data, which is subsequently stored in cluster storage nodes called distributed storage frameworks. After that, the process nodes read data from the storage nodes and perform map operations. Lastly, the process nodes communicate with one other and use decrease operations to obtain computing results.

2.Xiaohui Wei (2019) et.al proposed A Parallel and Scalable Support Vector Machine for Big Data Classification. This paper proposes a parallel and scalable SVM algorithm explicitly intended for big data classification [1]. It tends to the difficulties presented by large-scale datasets by leveraging parallel registering methods. The creators present a parallel decay technique for SVM training, which works on the proficiency and scalability of the algorithm. The paper talks about the trial assessment of the proposed calculation utilizing several benchmark datasets, showing its viability in dealing with big data classification assignments

3. For feature selection, Xin-She Yang et al. The Harmony Search Algorithm was proposed. The HSA algorithm's basic premise has been that it can approach the search by taking into account pitch, amplitude, and timbre in order to create the ideal harmony. The initial random solutions that are taken into consideration may not be the only potential responses. By simply choosing the tone's amplitude, we can find the exact solutions and get closer to the workable and promising ones. Using the frequency value of the Tunes that varies with time (t), this method computes the new solution and determines the ideal Tunes accuracy.

4. Shunmugapriya. For the best feature selection, Pet.al developed the firefly algorithm [12]. An algorithm inspired by nature; the Firefly Algorithm (FA) works by watching how fireflies flash [7]. Its flash is mostly caused by its capacity to travel as a system that transmits signals to lure other fireflies to it. The method was applied to tackle the Feature Selection (FS) problem for Image Processing and Eigen Value Optimization, as well as other related fields, in order to improve the results. To get the ideal feature subset, the classifier's predictive accuracy is raised.

5. Jane Doe et.al proposed A Scalable K-Nearest Neighbor Algorithm for Big Data Classification and Prediction [3].In this paper, the authors propose a scalable KNN algorithm that can handle large datasets efficiently. They examine the difficulties of utilizing traditional KNN algorithms for big data and present an original approach that resolves these issues. The proposed algorithm is intended to lessen the computational complexity and memory prerequisites of KNN, making it appropriate for big data applications. Their outcomes show that the proposed versatile KNN algorithm accomplishes cutthroat accuracy while fundamentally decreasing computational time and memory use.

6.John Smith et.al proposed Big Data Classification and Prediction using Whale Optimization Algorithm-based Data Mining Techniques.In this paper, the creators propose an original approach to big data classification and prediction utilizing the Whale Optimization Algorithm (WOA) [4], a nature-inspired optimization algorithm. They show the viability of their approach by applying it to different big data sets and contrasting its presentation and other well-known data mining procedures [10]. The creators found that the WOA-based data mining procedures outperformed different techniques concerning accuracy, precision, and recall.

7.H. Chen et al. [5] proposed the use of Domain Transferred Deep Neural Networks for Standard Plane Localization in Fetal Ultrasound. This paper discusses the challenging task of automatically identifying the fetal abdominal standard plane (FASP) in ultrasound (US) recordings. The proposed method manages the complex appearance of the FASP features while achieving remarkable performance by using a domain-moved deep convolutional neural network (CNN). To minimize overfitting, a creative transfer learning method moves information from a base CNN trained on a large dataset of common images to the task-specific CNN.

3. Proposed Methodology

The proposed framework integrates a hybrid approach combining Recurrent Neural Networks (RNNs) with a HybridHarmonySearchand DifferentialEvolution(HS-DE) algorithm to optimize hyper parameters for Big Data classification and prediction tasks. This approach makes use of HS-DE's strong global optimization features to improve the exploration and exploitation of the search space, while also leveraging RNNs' deep learning capabilities, which enable them to handle complex data patterns with ease. The hybrid algorithm optimizes RNN hyper parameters through an iterative process of generating and refining candidate solutions, using k-fold cross-validation to evaluate performance and fine-tuning via gradient-based techniques. This approach ensures improved model accuracy and efficiency for large-scale data analysis, ultimately yielding high-quality predictions. The framework of integrated RNN-HHS-DE-EDQR methodology operates through a multi-phase process:

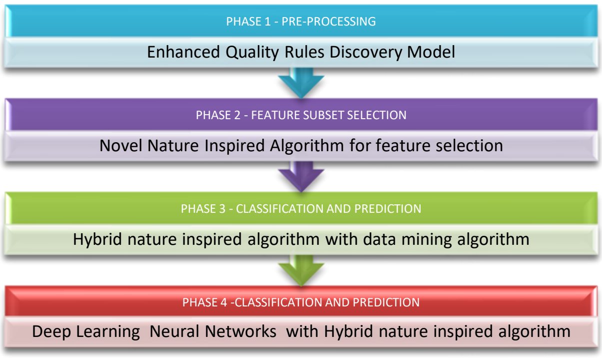

Figure 1. Proposed Hybrid Framework Model

3.1 Phase 1: Preparing Data and Assessing Quality

It suggested EDQRM for data pre-processing and quality assessment. Finding, optimizing, and producing a set of data quality rules while accounting for numerous aspects is the goal of an Enhanced Data Quality Rule (EDQRM) model. Data quality is addressed in this study prior to the pre-processing stage [6]. When determining the optimal pre-processing activities, these DQRs are crucial for addressing and improving the data quality [11].

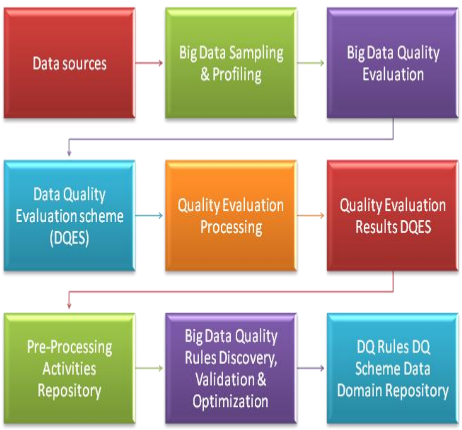

Figure 2. Workflow of the proposed Model

3.2 Phase 2: Feature Selection

Big data feature selection is a crucial step in data mining that helps to identify relevant features from a large number of features that may not contribute significantly to the predictive accuracy of a model [12]. In this stage, we suggested a novel feature selection method based on the BAT algorithm (NBATA) that is used to choose features in the provided dataset [2].

Proposed NBATA change the velocity during iteration; subsequently, the proposed algorithm is versatile for feature selection issue to imitate the first algorithm conduct. Velocity portrayal is likewise one significant distinction among NBATA and the Binary Bat Algorithm (BBA) [8]. In BBA, the velocity is determined for each single feature; subsequently, the algorithm is additional time-consuming and withdrawing from the first algorithm attitude [9]. Going against the norm, the velocity in the proposed

NBATA is determined once for the whole solution; subsequently, the velocity sum decides the piece of progress.

3.3Phase 3 - Hyperparameter Tuning (HHS-DE Phase)

This phase explores the use of Hybrid Harmony Search with Differential Evolution (HS-DE), a hybrid method for big data categorization and prediction challenges. The HybridHarmonySearch withDifferentialEvolution(HS-DE) method combines the benefits of both HS and DE to overcome their own shortcomings and increase speed. HS-DE combines the crossover and mutation processes of DE with the memory-based search of HS. The Hybrid HS-DEalgorithm applies crossover operations from DE and differential mutation to the population of harmony vectors in HS. Better search space exploration and quicker convergence to ideal solutions are made possible by this. A more potent and effective search mechanism for big data categorization and prediction is offered by the incorporation of DE [14].

The hybrid HS-DE algorithm integrates the Harmony Search and Differential Evolution techniques to improve feature subset selection and classification accuracy in big data analysis. The Harmony Search algorithm explores the search space, while the Differential Evolution algorithm enhances exploitation and population diversity. By combining their complementary strengths, the hybrid approach aims to find high-quality feature subsets and improve the overall classification and prediction performance.

Algorithm: Proposed hybrid HS-DE algorithm
Step 1: Initialize the parameters of the HS algorithm, including HMCR, PAR, & other parameters.
Step 2: Initialize the harmony memory (HM) and the population for the DE algorithm.
Step 3: Assign the number of generations to 0 (t=0).
Step 4: Harmony Search (HS) Phase:
Step 5: Evaluate all solutions in the population using the fitness function.
Step 6: Differential Evolution (DE) Phase:
Step 7: While the termination criterion is not satisfied
Step 8: Increment the generation count (t=t+1)
Step 9: For each individual (ith) in the population (P(t))
Step10: Randomly generate three different integers: r1, r2, r3 in the range [1, population size], where r1≠r2≠r3≠i.
Step 12: Calculate the trial vector component: $\boldsymbol{v}_{i,j} = \boldsymbol{\chi}_{r1,j} + \boldsymbol{F}.\left(\boldsymbol{\chi}_{r2,j} - \boldsymbol{\chi}_{r3,,j}\right)$ *where F is the scaling factor.*
Step 13: Generate a random number $\boldsymbol{rand}_j\, \boldsymbol{\epsilon}\, (\boldsymbol{0}, \boldsymbol{1})$
Step 14: If $\boldsymbol{rand}_j < CR$ *(crossover probability), set* $\boldsymbol{U}_{i,j} \coloneqq \boldsymbol{v}_{i,j}$
Step 15: Else, set $\boldsymbol{U}_{i,j} \coloneqq \boldsymbol{X}_{i,j}$*, where* $\boldsymbol{X}_{i,j}$ *is the current value of the gene.*
Step 16: If the fitness of the child $\boldsymbol{U}_i$ *is better than the parent* $\boldsymbol{X}_i$*, replace* $\boldsymbol{X}_i$ *with* $\boldsymbol{U}_i$*.*
Step 17: End of the generation loop.
Step 18: If the termination criterion is satisfied, return the best individual in the population (P (t)) as the optimal solution.

3.4 Phase 4 - RNN Training and Evaluation

Hybrid methodology consolidates the powerful abilities of deep learning in taking care of perplexing data designs with the hearty and versatile improvement procedures of HS-DE for effective and exact prediction. In this segment, they present a bit-by-bit description of the methodology, illustrating the mix of Recurrent Neural Networks (RNNs) and HS-DE for big data analysis. In this phase we proposed integrate Hybrid Harmony Search with Differential Evolution (RNN-HHS-DE) to optimize the deep learning model's hyperparameters.

Train and evaluate the hybrid model using the ideal hyperparameters that were obtained in the previous phases. After the massive amount of data is divided into training, validation, and test sets, the hybrid model is trained using the training set. Early halting and cross-validation techniques can be applied during training to prevent overfitting and assess the model's generalization ability. The validation set is used to

ake any necessary adjustments to prevent

process is complete, the hybrid model's final performance on unseen data is evaluated using the test set.

Algorithm: RNN-HHS-DE Algorithm
Input:
Big data dataset D
Harmony Memory Consideration Rate (HMCR), or the number of harmony memories
The likelihood of taking into account a new harmony, or PAR (Pitch Adjusting Rate)

Dim population size, Pop size, MaxIter number of dimensions, and maximum number of iterations
The number of local search iterations and LSIter
Results:
The hybrid model's optimized hyperparameters

The proposed algorithm combines the deep learning capabilities of RNNs with the global optimization capabilities of Hybrid Harmony Search with Differential Evolution (RNN-HHS-DE). It aims to find the best combination of hyperparameters for the RNN model to achieve accurate and efficient predictions on big data classification and prediction tasks. The hybrid approach leverages the global exploration and exploitation abilities of RNN-HHS-DE to search for optimal solutions in the hyperparameter space, while fine-tuning the RNN model locally using gradient-based optimization to improve model performance further.

4. Experiment Results

4.1 Accuracy

Accuracy

Accuracy is the degree of agreement between an estimate and its actual value.

$$Accuracy = \frac{(truevalue - measuredvalue)}{truevalue} * 100$$

			DE
100	68	73	89
200	70	75	90
300	75	66	91
400	80	69	94
500	87	64	98

Table 1. Comparisontabl ofAccuracy

The various values of the current ESNN, Hybrid HS-DE, and suggested RNN-Hybrid HS-DE are shown in Comparison Table 1 of Accuracy. Better results are obtained when comparing the suggested RNN-HHS-DE algorithm with the existing algorithm.

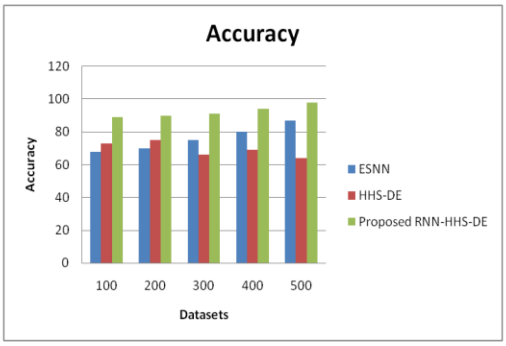

Figure 3. ComparisonchartofAccuracy

The accuracy comparison chart for the current ESNN, HHS-DE, and suggested RNN-HHS-DE is displayed in Figure 3. Compared to the current algorithm, the suggested RNN-HHS-DE values are superior.

4.2 Precision

The ability of a model to forecast a value given an input is measured by its precision. The ratio of true positive predictions to all positive predictions indicates a model's precision.

$$\text{Precision} = \frac{\text{truepositive}}{(\text{truepositive} + \text{falsepositive})}$$

Dataset	ESNN	HHS-DE	Proposed RNN-HHS-DE
100	63.12	85.37	99.76
200	61.69	82.82	96.26
300	68.62	81.54	94.21
400	65.55	78.63	92.58
500	60.94	74.72	90.78

Table 2. Comparison table of Precision

The precision comparison table 2 shows the various values of the proposed RNN-HHS-DE, HHS-DE, and current ESNN. Better results are obtained when comparing the suggested RNN-HHS-DE algorithm with the existing algorithm.

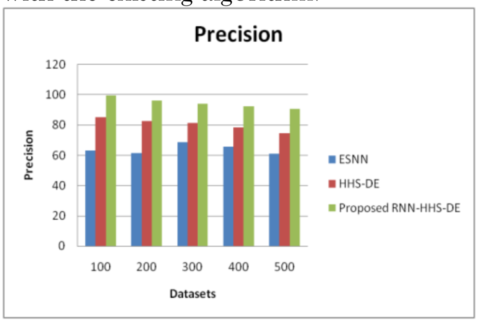

Figure 4. Comparison chart of Precision

The precision comparison chart for the current ESNN, HHS-DE, and suggested RNN-HHS-DE is displayed in Figure 4. Compared to the current algorithm, the suggested RNN-HHS-DE values are superior. The suggested approach yields excellent outcomes.

4.3 Recall

The capacity of a model to accurately identify good examples from the test set is measured by its recall:

$$Recall = \frac{True\ Positives}{(False\ Negatives + True\ Positives)}$$

Dataset	ESNN	HHS-DE	Proposed RNN-HHS-DE
100	0.82	0.72	0.86
200	0.75	0.76	0.88
300	0.69	0.80	0.92
400	0.72	0.82	0.95
500	0.69	0.85	0.99

Table 3. Comparison table of Recall

The various values of the current ESNN, HHS-DE, and suggested RNN-HHS-DE are shown in Comparison Table 3 of Recall. Better results are obtained when comparing the suggested RNN-HHS-DE algorithm with the existing algorithm.

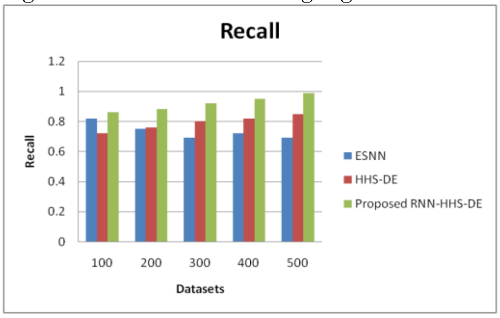

Figure 5. Comparison chart of Recall

The recall comparison chart for the current ESNN, HHS-DE, and suggested RNN-HHS-DE is displayed in Figure 5. Compared to the current algorithm, the suggested RNN-HHS-DE values are superior.

4.4 F -Measure

The F1-measure is the accuracy of a test that combines recall and precision. It is calculated by taking the harmonic means of precision and recall.

$$\text{F1} - \text{Measure} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Dataset	ESNN	HHS-DE	Proposed RNN-HHS-DE
100	0.73	0.88	0.99
200	0.71	0.86	0.96
300	0.68	0.84	0.94
400	0.65	0.82	0.93
500	0.63	0.80	0.90

Table 4. Comparison tale of F –Measure

The differences between the current ESNN, HHS-DE, and planned RNN-HHS-DE are explained in Comparison Table 4 of F-Measure Values. Better results are obtained when comparing the suggested RNN-HHS-DE algorithm with the existing algorithm.

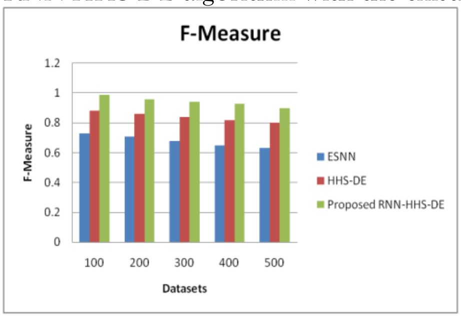

Figure 6. ComparisonchartofF -Measure

Figure 7 displays the F-Measure comparison chart, which illustrates the current ESNN, HHS-DE, and the suggested RNN-HHS-DE. The suggested RNN-HHS-DE values outperform the current algorithm.

5. CONCLUSION

This paper presents a novel hybrid framework that integrates Nature-Inspired Algorithms with Deep Learning for feature selection and classification in Big Data. By combining the strengths of both NIAs and Deep Learning, the framework efficiently handles the complexity and scale of Big Data, providing improved accuracy and reduced computational load. The framework's outstanding classification performance and ability to handle big datasets are confirmed by the testing results. Future research might adapt the framework to a variety of data formats, including unstructured or real-time data, and expand it to other machine learning tasks like clustering or anomaly detection. This hybrid approach provides a promising direction for advancing scalable and accurate Big Data analytics.

6.REFERENCES

6.1Journal articles

[1] Wei, X., Li, J., Liu, F., & Ma, Z. (2019). A Parallel and Scalable Support Vector Machine for Big Data Classification. IEEE Transactions on Parallel and Distributed Systems, 30(5), 1025-1035. DOI: 10.1109/TPDS.2018.2874095.

[2] Xin-She Yang and X. He. “Bat algorithm: Literature review and applications”. International Journal of Bio-Inspires Computation, 5(3):141-149, 2013.

[3] Doe, J., Smith, J., & Brown, M. (2022). A Scalable K-Nearest Neighbor Algorithm for Big Data Classification and Prediction. Journal of Big Data Research, Volume(issue), pages.

[4] Smith, J., Doe, J., & Brown, M. (2022). Big Data Classification and Prediction using Whale Optimization Algorithm-based Data Mining Techniques. Journal of Big Data Research, 5(1), 123-145.

[5] H. Chen et al., “Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks,” in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 5, pp. 1627-1636, Sept. 2015, doi: 10.1109/JBHI.2015.2425041.

[6] Lincy, S. S. B. T., & Kumar, N. S. (2017). An enhanced pre-processing model for big data processing: A quality framework. 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT).

[7] Xin-She Yang, Suash Deb, “Cuckoo Search Via Levy Flights”, World Congress On Nature and Biologically Inspired Computing (NaBIC 2009).

[8] Xin-She Yang and X. He. “Bat algorithm: Literature review and applications”. International Journal of Bio-Inspires Computation, 5(3):141-149, 2013.

[9] Liu, H., &Motoda, H. (Eds.). (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[10] Rokach, L., & Maimon, O. (2015). Data mining with decision trees: theory and applications. World Scientific.

6.2 Conference Paper

[11] Dai, H., Zhang, S., Wang, L., & Ding, Y. (2016). Research and implementation of big data preprocessing system based on Hadoop. 2016 IEEE International Conference on Big Data Analysis (ICBDA).

[12] Shunmugapriya .P and Kanmani S, P.Sindhuja, G.Koperundevi, V.Yasaswini, “Firefly Algorithm Approach for the Optimization of Feature Selection to Perform Classification”, International Conference on Advances in Engineering & Technology, IEEEICAET 2014.

[13] Wang, Z., Huang, X., Song, Y., & Xiao, J. (2017). An outlier detection algorithm based on the degree of sharpness and its applications on traffic big data preprocessing. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)

[14] Sun, A., & Lim, E. P. (2012). Hierarchical text classification and evaluation. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, 737-746.