Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 8, pp. 2203–2212
DOI : 10.47974/JIOS-1783

Improving imbalanced data classification in healthcare systems : A novel optimization approach with statistical insights

Eshwar Dara *
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota
Florida 33620
U.S.A.

and

Department of Computer Science and Engineering
Kommuri Pratap Reddy Institute of Technology (KPRIT)
Hyderabad 500088
Telangana
India

Bhuvan Unhelkar †
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota
FL 34243
USA

S. Siva Shankar §
Department of Computer Science and Engineering
KG Reddy College of Engineering and Technology
Hyderabad 501504
Telangana
India

* *E-mail:* **dreshwardara@gmail.com (Corrsponding Author)**
† *E-mail:* **bunhelkar@usf.edu**
§ *E-mail:* **drsivashankars@gmail.com**

Tulika Chakrabarti [‡]
Department of Chemistry
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

Prasun Chakrabarti [@]
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

B. Sivaneasan [#]
Singapore Institute of Technology
Specialist Adult Educator Engineering
Singapore Institute of Technology
1 Punggol Coast Road 828608
Singapore

Martin Margala [$]
School of Computing and Informatics
University of Louisiana at Lafayette
Louisiana 70503 (337) 482-6768
U.S.A.

Abstract

Imbalanced data classification poses a challenge in medical intelligent diagnosis due to limited samples and high-dimensional features in real-world biomedical datasets. This issue impacts the classification accuracy and can lead to erroneous disease diagnoses. Addressing this challenge requires effective methods tailored for imbalanced and limited biomedical datasets. In this study, we introduce a novel approach, the Artificial Jellyfish Optimization Based Advanced Dynamic Neural Fuzzy (AJO-ADNF) classification model, coupled with a Hybrid Shuffle Shepherd based Sampling Method (HSS-SM). Initially, irrelevant features are removed through preprocessing, while significant pathogenic traits are identified. The HSS-SM algorithm balances the dataset by providing reasonable minority and majority class data, mitigating computational burden. Subsequently, the AJO-ADNF classifier is applied to

[‡] *E-mail:* **tulika.chakrabarti@spsu.ac.in**
[@] *E-mail:* **drprasun.cse@gmail.com**
[#] *E-mail:* **sivaneasan@singaporetech.edu.sg**
[$] *E-mail:* **Martin.margala@louisiana.edu**

classify the balanced dataset. This approach generates genuine minority and majority class data and autonomously selects optimal learning model parameters, enhancing robustness and efficiency. Evaluation on COVID-19 healthcare datasets using MATLAB demonstrates superior classification performance compared to conventional methods in terms of Accuracy, Sensitivity, Specificity, AUC, F-measure, and other metrics.

Subject Classification: *68T07, 92C55.*

Keywords: *Imbalanced data, Machine learning, Statistical insights, Optimization, Neural fuzzy, Sampling, Performance measures, Healthcare and big data.*

I. Introduction

Big data presents challenges in classification due to its volume, variety, and velocity, particularly in medical contexts. Imbalanced datasets, where one class significantly outnumbers the other, further complicate analysis, leading to biased results and inaccurate conclusions [1-3]. Traditional classifiers struggle with this imbalance, affecting classification accuracy. To address this, we propose the AJO-ADNF

Method with HSS-SM, which preprocesses data to remove irrelevant features, balances datasets, and optimizes classification. This approach enhances accuracy and precision, particularly in healthcare applications like COVID-19 diagnosis [4-5].

Our contributions include effective preprocessing, dataset balancing, and improved classification through optimized parameters. Experimental results on healthcare datasets demonstrate the model's efficacy, offering insights for biomedical applications. The paper proceeds with a review of related work, a discussion of the problem statement, presentation of the proposed framework, experimental results, and analysis, concluding with implications and future directions.

II. Problem statement

Big data is integral to modern technology, spanning social networks and smart infrastructure. However, the challenge lies in handling imbalanced datasets, which can skew classification results and lead to erroneous conclusions [6-8]. Existing deep learning techniques struggle with class imbalances, requiring adjustments to network structures and optimization processes. In medical datasets, imbalances hinder illness prediction, necessitating sampling approaches for data balancing [9-12]. However, these methods have drawbacks, such as information loss and

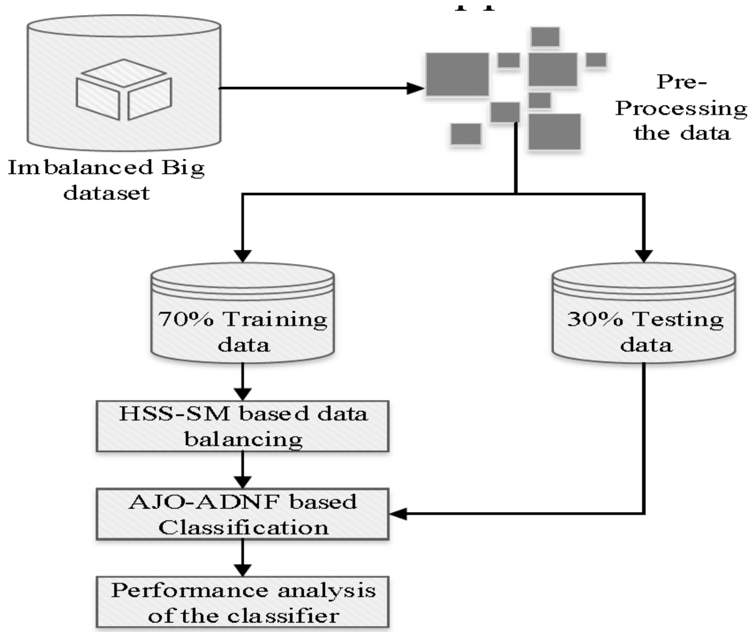

Figure 1

The block representation of the proposed system

prolonged training time [13-17]. Conventional classification techniques are also affected by probability bias, making it difficult to distinguish between minority and majority classes, especially in multiclass scenarios. This poses significant challenges in accurately assessing performance, particularly in healthcare contexts. As a result, alternative metrics and approaches are needed to address the complexities of imbalanced datasets [18-23].

III. Proposed methodology

The proposed system for imbalanced big data classification includes three stages, pre-processing, sampling and classification. In this research, the hybrid sampling approach is utilized named as HSS-SM based on the fitness of this approach.

Also, the balanced testing and training test of classification is executed by the proposed AJO-ADNF method. The parameters of the ADNF method is tuned by the fitness function of AJO technique. The block representation of the proposed system is shown in Figure 1.

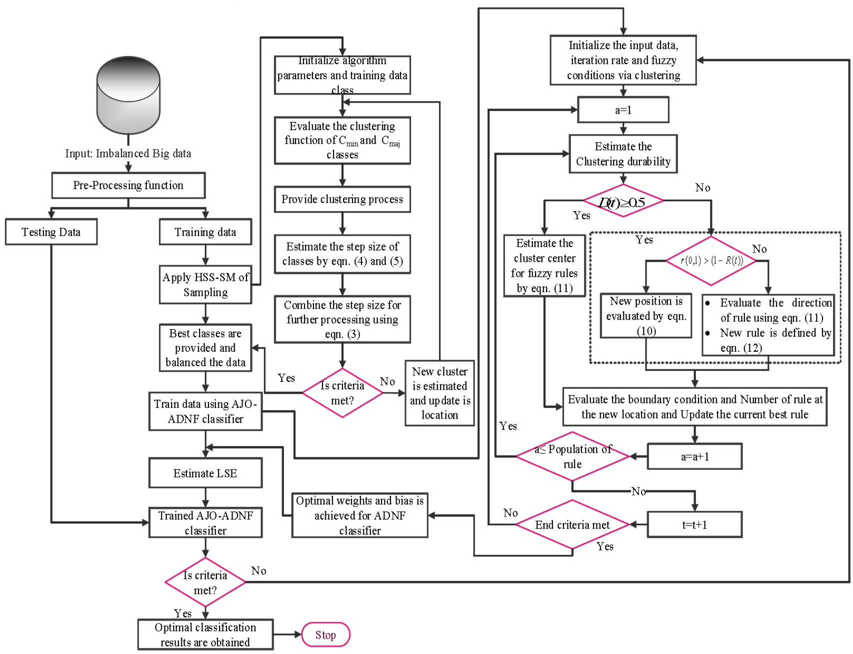

Figure 2

The flowchart of proposed AJO-ADNF with HSS-SM method

- Pre-Processing Stage
- HSS-SM based data balancing
- AJO-ADNF based Classification

1. AJO based parameter tuning

   Step 1: Initialization

   Step 2: Clustering Durability

   Step 3: Cluster center

   Step 4: Cluster of Fuzzy rule

2. ADNF classifier

   The flowchart of proposed AJO-ADNF with HSS-SM method is illustrated in Figure 2.

IV. Experiment setup and analysis

Research had been carried out in this part to validate the performance of the suggested strategy. The following is a list of experimental system setups: Windows 7 Operating System Intel(R) dual-core, 3.20 GHz, MATLAB R2019a tool, 8 GB RAM.

- Experimental analysis

The experimental dataset contribution is provided in Table.1.

Table 1

The experimental dataset contribution

Datasets	Attributes	Ratio of Imbalance	Majority	Minority
A- Fleury Institute	25	6.44	20,097	1,29,597
B- Sri-Libanês Hospital	23	1.4	1,946	2,732
C- Israelita Albert Einstein Hospital	21	4.66	9,632	44,879

- Performance Comparison.

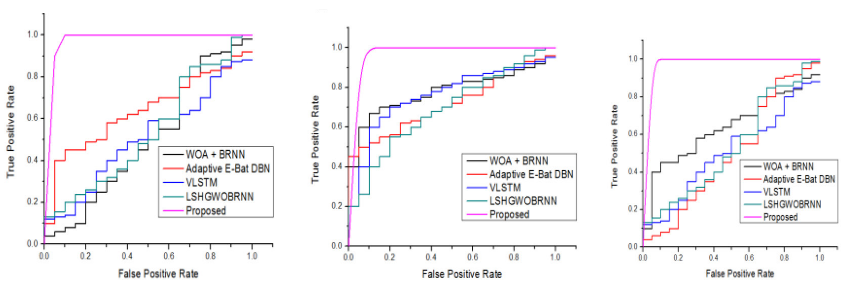

Figure 3

ROC curve of COVD-19 imbalanced data a) Fleury Institute b) Sri-Libanês Hospital c) Israelita Albert Einstein Hospital

This study (Figure 3) created Receiver Operating Characteristic (ROC) curves for each of the COVID-19 datasets.

The classification Accuracy, Precision, Sensitivity, Specificity and AUC, F-Measure, FPR, FNR, MCC and Computational time of imbalanced COVID-19 data by the proposed method has achieved finest performance over the conventional methods, which is illustrated in Fig.4 (a-j).

V. Discussion

The proposed AJO-ADNF with HSS-SM approach demonstrates superior performance in classifying imbalanced big data across various

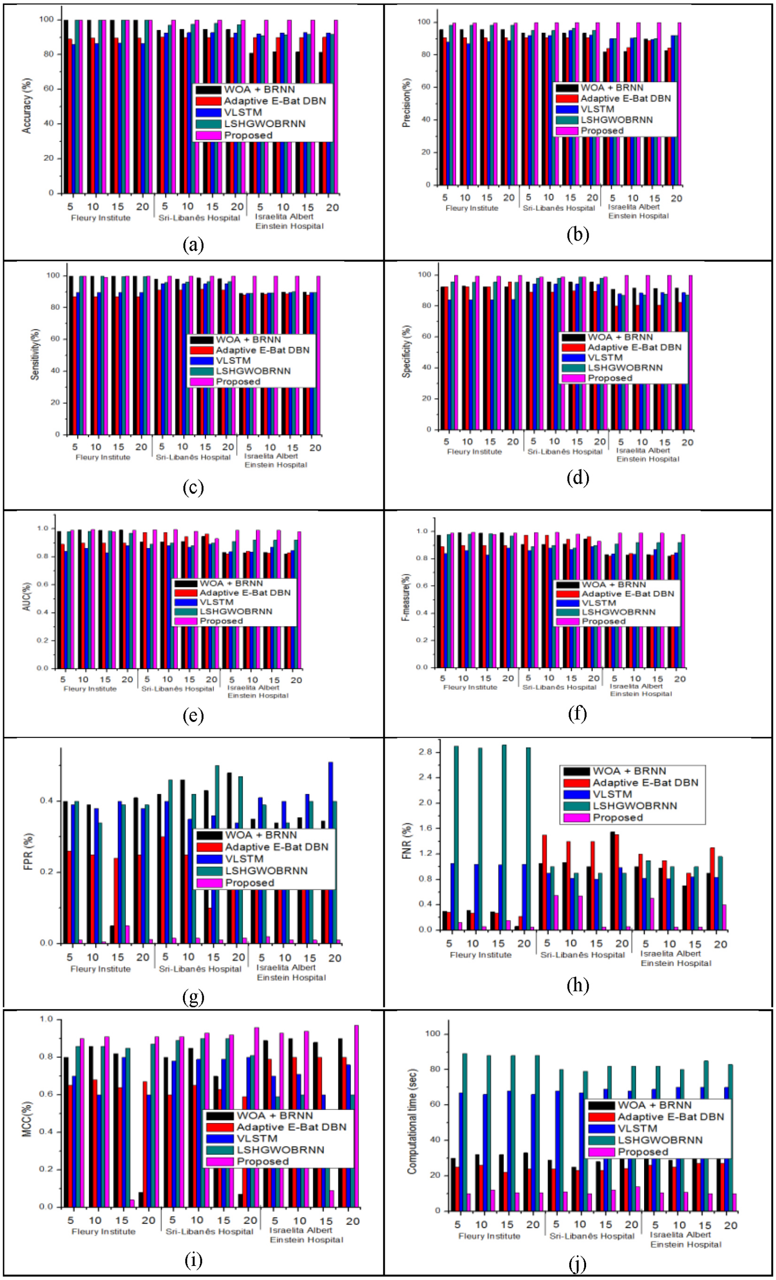

Figure 4

Performance of conventional methods

training stages (5%, 10%, 15%, and 20%). Compared to conventional techniques, it consistently achieves higher positive rates for all performance metrics. Specifically, in the proposed approach achieves remarkable accuracy: 100% for Fleury Institute, 99.95% for Sri-Libanês Hospital, and 100% for Israelita Albert Einstein Hospital. In contrast, existing methods like WOA + BRNN, Adaptive E-Bat DBN, VLSTM, and LSHGWOBRNN show lower accuracy rates for these hospitals. Moreover, the proposed model outperforms existing works across all parameters, demonstrating significant improvement in imbalanced big data classification using intelligent optimized machine learning techniques.

VI. Conclusion

Imbalanced real-world data challenges effective classification. A novel optimization-tuned machine learning approach enhances accuracy for large datasets. Using the COVID-19 dataset as an example, the study introduces the Hybrid Sampling approach, HSS-SM, to create balanced data and achieve remarkable accuracy with the AJO-ADNF classifier. This method holds promise for biomedical data classification, with potential applications in clinical settings. Future research may explore additional deep learning approaches to further improve model performance on complex imbalanced big data.

References

[1] Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," *Cities*, vol. 89, pp. 80–91 (2019).

[2] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, "Internet-of-things-based smart cities: Recent advances and challenges," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 16–24, Sep. (2017).

[3] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94-101, Feb. (2018).

[4] M. Gohar, M. Muzammal, and A. U. Rahman, "SMART TSS: Defining transportation system behavior using big data analytics in smart cities," *Sustainable Cities and Society*, vol. 41, pp. 114–119 (2018).

[5] L. D. Ribeiro-Reis, "Truncated exponentiated-exponential distribution: A distribution for unit interval," J. Stat. Manag. Syst., vol. 25, no. 8, pp. 2061-2072, 2022.

[6] N. Mahankale, S. Gore, D. Jadhav, G. S. P. S. Dhindsa, P. Kulkarni, and K. G. Kulkarni, "AI-based spatial analysis of crop yield and its relationship with weather variables using satellite agrometeorology," in 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 1–7, Oct. (2023).

[7] R. Vinayakumar, M. Alazab, S. Srinivasan, Q. V. Pham, S. K. Padannayil, and K. Simran, "A visualized botnet detection system based deep learning for the Internet of Things networks of smart cities," *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 4436–4456, Jul.–Aug. (2020).

[8] M. A. Ahad, S. Paiva, G. Tripathi, and N. Feroz, "Enabling technologies and sustainable smart cities," *Sustainable Cities and Society*, vol. 61, p. 102301 (2020).

[9] G. Chen, Y. Liu, and Z. Ge, "K-means Bayes algorithm for imbalanced fault classification and big data application," *Journal of Process Control*, vol. 81, pp. 54–64 (2019).

[10] M. I. Khan, "Dual generalized order statistics with moments properties using powered inverse Rayleigh distribution," *Journal of Statistics and Management Systems*, vol. 25, no. 8, pp. 2087-2099 (2022).

[11] S. Gore, Y. Bhapkar, J. Ghadge, S. Gore, and S. K. Singha, "Evolutionary Programming for Dynamic Resource Management and Energy Optimization in Cloud Computing," in 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 1–5, Oct. (2023).

[12] R. A. Bauder and T. M. Khoshgoftaar, "The effects of varying class distribution on learner behavior for Medicare fraud detection with imbalanced big data," *Health Information Science and Systems*, vol. 6, pp. 1–14 (2018).

[13] W. C. Sleeman IV and B. Krawczyk, "Multi-class imbalanced big data classification on spark," *Knowledge-Based Systems*, vol. 212, p. 106598 (2021).

[14] J. M. Johnson and T. M. Khoshgoftaar, "The effects of data sampling with deep learning and highly imbalanced big data," *Information Systems Frontiers*, vol. 22, no. 5, pp. 1113-1131 (2020).

[15] Y. Soni, G. C. Gandhi, and D. Goyal, "A secure e-health framework for rural Rajasthan," *Journal of Statistics and Management Systems,* vol. 25, no. 8, pp. 2113-2122 (2022).

[16] S. Gore, D. Jadhav, M. E. Ingale, S. Gore, and U. Nanavare, "Leveraging BERT for Next-Generation Spoken Language Understanding with Joint Intent Classification and Slot Filling," in 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 1–5, Oct. (2023).

[17] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data,* vol. 7, pp. 1–47 (2020).

[18] W. D. Xie and X. Cheng, "Imbalanced big data classification based on virtual reality in cloud computing," *Multimedia Tools and Applications*, vol. 79, no. 23, pp. 16403–16420 (2020).

[19] C. Banchhor and N. Srinivasu, "Integrating Cuckoo Search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification," *Data Knowledge Eng.,* vol. 127, p. 101788 (2020).

[20] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics,* vol. 107, p. 103465 (2020).

[21] C. M. Paredes, D. Martínez-Castro, V. Ibarra-Junquera, and A. González-Potes, "Detection and isolation of DoS and integrity cyber attacks in cyber-physical systems with a neural network-based architecture," *Electronics,* vol. 10, no. 18, p. 2238 (2021).

[22] L. Jerlin Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm," *Int. J. Imaging Syst. Technol.*, vol. 30, no. 3, pp. 660-673 (2020).

[23] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," *Computers & Industrial Engineering*, vol. 140, p. 106266 (2020).