A Privacy Based Deep Learning Algorithm for Big Data Analytics

D. Franklin Vinod*, Neha Ahlawat
Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, NCR Campus, Delhi-NCR Campus, Delhi-Meerut Road, Modinagar, Ghaziabad, UP, India.
E-mail: datafranklin@gmail.com, nehablow@gmail.com
*Corresponding author

Thesis summary

Keywords: Big Data, Deep Belief Network, Heterogeneous data, pattern recognition, privacy preserving

Received: March 31, 2025

This thesis addresses critical challenges in privacy-preserving feature selection and classification for big data analytics. Specifically, four novel methodologies are proposed: Hierarchical Classification Feature Selection (HCFS), Privacy-Preserving Classification Selection with p-stability (PPCS), Local N-ternary Pattern combined with Modified Deep Belief Network (LNTP-MDBN), and Privacy-Preserving Cosine Similarity integrated with Multi-Manifold Deep Metric Learning (PPCS-MMDML). These approaches collectively enhance classification accuracy, optimize feature extraction from heterogeneous image sets, and robustly preserve privacy, demonstrating significant improvements in data-driven analytical applications.

Povztek: V disertaciji razbiti algoritem globokega učenja omogoča učinkovito klasifikacijo in izbiro značilk pri analizi velikih podatkov.

1 Introduction

The exponential growth of big data has significantly increased the risks of privacy breaches. Traditional processing systems face challenges managing the high volume, velocity, and variety of data generated by modern technologies such as sensors and the Internet of Things (IoT). Distributed architectures like Hadoop facilitate efficient big data management; however, handling sensitive data, particularly in sectors such as healthcare, necessitates stringent privacy and security measures. This research focuses on innovative deep learning techniques that balance effective data utilization with essential privacy protection. Four advanced methodologies—HCFS, PPCS with p-stability, LNTP-MDBN, and PPCS-MMDML—are introduced, targeting privacy-preserving feature selection and efficient classification, especially in complex image datasets [1-3].

The thesis summery highlights the significance of big data analytics in the modern digital era., emphasizing the need for classification in big data environments, the role of deep learning in big data classification, and the privacy challenges in big data analytics.

2 Methodology and designs

The thesis introduces four distinct privacy-preserving methods:

HCFS (Hierarchical Classification Feature Selection) enhances classification by identifying optimal feature subsets, significantly improving accuracy [4].

PPCS with p-stability is a bi-level approach safeguarding individual and network-level privacy, effectively preventing intrusions and preserving data integrity [5].

LNTP-MDBN (Local N-ternary Pattern with Modified Deep Belief Network) specializes in feature extraction from heterogeneous images, minimizing reconstruction errors and maximizing classification accuracy through strategic modifications to deep belief networks [6].

PPCS-MMDML (Privacy-Preserving Cosine Similarity with Multi-Manifold Deep Metric Learning) addresses privacy-preserving classification challenges specifically in cancer image analysis, maintaining high classification accuracy while enforcing stringent privacy constraints [7].

3 Results and discussion

The OSIRIX viewer [8] and the Mammographic Image Analysis Society (MIAS) [9] are used to build datasets for the brain, breast, and bone in order to support the suggested performance. Every image used has a 1024 x 1024-pixel dimension. Figure 1 illustrates the metrics used to validate the performance of the proposed LNTP-MDBN, including classification accuracy, macro-averaged F1 score, and running time.

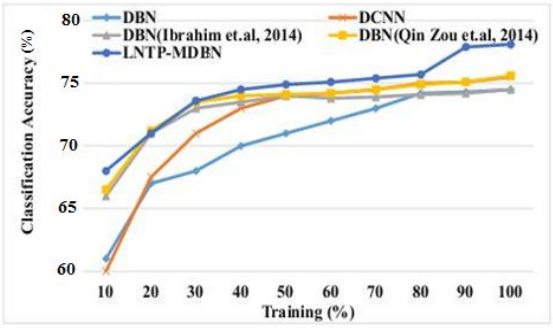

Figure 1: Classification accuracy analysis

With 100% training, the DBN and LNTP-MDBN have respective accuracy percentages of 75.6 and 78.1. The comparison analysis demonstrates that the suggested LNTP-MDBN work is superior to the current DBN.

Because of enhanced pattern extraction and the redesigned DBN that carries out heterogeneous image set classification, the suggested LNTP-MDBN uses less computing time for each level. LNTP spends 0.095, 1.25, and 0.023 seconds on feature extraction, training, and testing levels, respectively.

Table 1: Classification accuracy analysis of PPCS-MMDML

S.No	Data sets	Classification Accuracy (%)	
		MMDML	PPCS-MMDML
1	Bone Cancer	66.5	74.2
2	Brain Cancer	68.5	72.6
3	Breast Cancer	71.6	77.8

The performance assessments of the proposed PPCS-MMDML and the existing MMDML in three datasets related to cancer diseases are shown in Table 1. The PPCS-MMDML accuracy rates for breast, brain, and bone cancer diseases are 77.8%, 74.2, and 72.6, respectively. According to the comparison analysis, the suggested PPCS-MMDML improves the categorization of the bone, brain, and breast datasets by 7.7%, 4.1%, and 6.2%, respectively, over the current MMDML.

4 Conclusion and future work

The proposed methods, particularly LNTP-MDBN and PPCS with p-stability, achieve superior results in privacy preservation and classification accuracy within big data environments. Future research directions include integrating PPCS and LNTP-MDBN for enhanced privacy-preserving deep learning frameworks, exploring online learning techniques for real-time classification, and expanding application. The comparison analysis proves that the proposed work LNTP-MDBN provides improvement over the existent DBN which is suitable for heterogeneous cancer disease detection.

References

[1] W. Dou, X. Zhang, J. Liu and J. Chen, Hiresome-II: Towards privacy aware cross-cloud service composition for big data applications, IEEE Trans Parallel Distrib Syst., 6(2), (2014), 455–466.

[2] A.T. Azar and A.E. Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier, Soft Computing, 19, (2015), 1115-1127.

[3] S. Gao, Z. Zeng, K. Jia, T.-H. Chan and J. Tang, Patch-Set-Based Representation for Alignment-Free Image Set Classification, IEEE Trans.Cir. and Sys. for Video Tech., 26, (2016), 1646-1658.

[4] D. F. Vinod and V. Vasudevan, "A filter-based feature set selection approach for big data classification of patient records," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 2016, pp. 3684-3687, doi: 10.1109/ICEEOT.2016.7755397.

[5] Franklin Vinod, D., Vasudevan, V. (2017). A Bi-level Security Mechanism for Efficient Protection on Graphs in Online Social Network. In: Arumugam, S., Bagga, J., Beineke, L., Panda, B. (eds) Theoretical Computer Science and Discrete Mathematics. ICTCSDM 2016. Lecture Notes in Computer Science, vol 10398. Springer, Cham.

[6] Vinod DF, Vasudevan V. LNTP-MDBN: Big Data Integrated Learning Framework for Heterogeneous Image Set Classification. Curr Med Imaging Rev. 2019;15(2):227-236. doi: 10.2174/1573405613666170721103949. PMID: 31975670

[7] Franklin Vinod, D., Vasudevan, V. (2019). PPCS-MMDML: Integrated Privacy-Based Approach for Big Data Heterogeneous Image Set Classification. In: Satapathy, S., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, vol 106. Springer, Singapore.

[8] OSIRIX. Available: http://www.osirix-viewer.com/resources/ dicom image-library

[9] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis and I. Ricketts, Mammographic Image Analysis Society(MIAS) database v1.21 [Dataset], (2015).

[10] R. Ibrahim, N. A. Yousri, M. A. Ismail and N. M. El-Makky, Multi-level gene/MiRNA feature selection using deep belief nets and active learning, In Proc Int. Conf. IEEE Engg. Med. Bio Society (EMBC), (2014), 3957-3960.

[11] Q. Zou, Y. Cao, Q. Li, C. Huang and S. Wang, Chronological classification of ancient paintings using appearance and shape features, Pattern Recognition Letters, 49, (2014), 146-154.