**Tuijin Jishu/Journal of Propulsion Technology**
**ISSN: 1001-4055**
**Vol. 44 No. 5 (2023)**

# Preventing Measures In Bipolar Disorder Via Machine Learning

**[1]Kavita Agarwal, [2]Dr. Tapsi Nagpal**

[1] [2] Department of Computer Science and Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India,
E-mail: [1]goel.kavita15@gmail.com,[2]dr.tapsi@lingayasvidyapeeth.edu.in

**Abstract:** The technical advancement in recent days has made a drastic increase in the amount of data generated from the physical, cyber, and human world. The collection of data at a huge scale makes sense only if the data is actionable and can also be used for making decisions. Data mining provides needed assistance at this stage by inspecting the relationship in the data and offers needed insights to the data owners. Moreover, the significant insights obtained are shared with third parties for further analysis. In this situation, numerous varieties of information is processed and the integration of machine learning with big data technology has made prominent aspect in the identification of bipolar disorder. It is a complex genetic disorder characterized by episodes of mania and depression. It affects 1% of the population worldwide. It is a major under-addressed public health problem which causes a significant burden on caregivers. High heritability and familial relative risk indicate the role of genetics in the etiology of the disorder. In this research, data is handled using an improved auto encoder and deep neural network. The feature selection is accomplished using ISAE and classification is accomplished using DNN. The proposed approach outperforms the existing state of art techniques.
**Keywords**: Bipolar disorder, big data, EMR, Registries, Claims, Patient Monitoring

## 1. Introduction

Bipolar disorder is a severe mood disorder with a lifetime prevalence of 1% in both males and females [1, 2]. It is characterised by recurrent episodes of dysregulated moods and is associated with high morbidity and suicide risk. Life time risk of an unrelated member of general population is 0.5 to 1.5%, that in first degree relative of bipolar disorder (BD) is 5-10% and that in monozygotic co- twin is 40-70% [3, 4]. It is a leading cause of global disability and its treatment is unsatisfactory. Family, twin, and adoption studies have provided strong evidence for the significant involvement of genes in predisposition to BD [5].

Estimates of heritability is as high as 89% and 93% in twin studies from hospital register in the UK and population register in Finland respectively [6]. However, the genetic basis for the disorder remains obscure [7-11]. It has been suggested that environmental stressors may trigger mood episodes [12] though they are more likely to be involved in the precipitation of the first episodes, but less with subsequent episodes [13].

Antioxidants suppress the process of oxidation that eventually led to the formation of free radicals. These free radicals are generated through lipid peroxidation. At higher concentration, free radicals from oxygen (reactive oxygen species - ROS) can damage the integrity of various biomolecules including DNA [14] and promote the activation of autophagy, apoptosis, and necrosis [15]. Oxidative stress has been implicated in the pathogenesis of neuropsychiatric diseases [16] as the brain has greater vulnerability to oxidative damage.

Big Data offers a huge value and has recognized as being impelling force behind the innovation of technology and economic progression. Emerging approaches likely smart grid, e-health system, social network etc, provides exceptional utilities by examining the data for obtaining better understanding and services. An issue related to privacy is faced during the communication, transmission, mining and aggregation process. The sophisticated mining approaches may increasingly efficient and can expose the underlying data. The privacy related issues had established viable elucidations to fulfil the necessities of privacy

The intent is to construct trade-off among utility and privacy which means effective use of these services while averting exposure or inference of private information during data analytic process. In the past few years,

researchers have retorted to the issues and projected various approaches [17]. While the privacy related research is still quickly establishing, it is significant to discuss the privacy related aspects to preserve the privacy of the data in the era of big data. This informative chapter offers a various facet of privacy in big data analytical process [18].

Big Data is a ubiquitous technology with the emergence of social network, outsourced cloud computing, Internet of Things (IoT) and data transmission generated huge data that witnessed the enormous growth in the data. The data generation is observed in name of huge volume, elevated velocity and diversified data variety [19]. The exceptional usage of networking among smart objects and the intelligent computational platform accords the Big Data but pretences towards privacy where the location of privacy, transaction and behaviour are recorded digitally [20, 21]. Nowadays, Hospitals maintaining patient medical records digitally and it contains personal medical data that increases the privacy concerns [22]. Most of the developed companies are utilizing Big Data to observe the workforce by tracking the performance and the productivity of the employees [23]. The issues faced in hospital and companies revealed the gap among the regulatory policies (convention), Big Data and the necessities of new policies to address the complete concerns of privacy [24, 25-33].

Evaluating the term of big data conceptually from privacy perspective is inadequate to understand privacy comprehensively in big data domain. Hence, if 7 V's of big data can be evaluated from privacy perspective, a better understanding of big data can be presented. As presented in Figure 1, Volume, Velocity, Variety, Veracity, Value, Variability and Vulnerability are considered and evaluated from privacy perspective in Table 1.

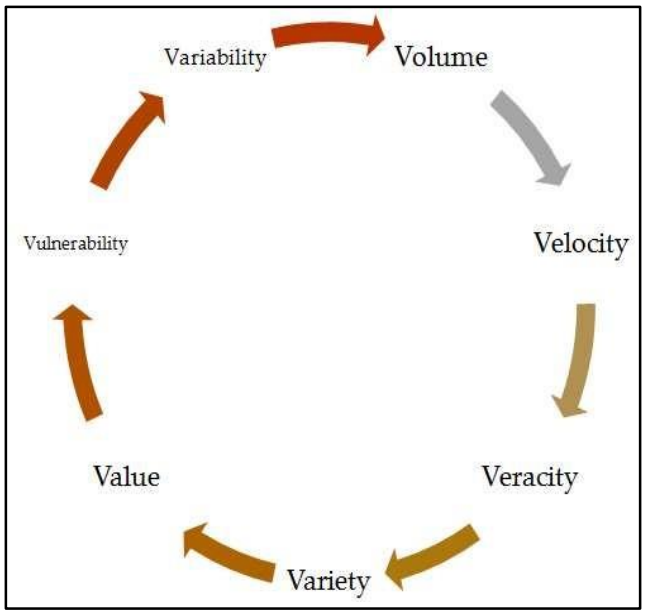


**Figure 4.** 7V's of Big Data

**Table 1:** Privacy Perspective of 7V's of Big Data

| V's of Big Data | View of Privacy Insights |
|---|---|
| Volume | The raise in the count of the data records is directly proportional to the raise of the volume of data. The increase of data volume results in the decreasing of privacy threats and increases the utility data. |
| Velocity | The lifetime of data utility is more effective and the process of stream, batch as well as real time is accompanied in terms of high utility. |
| Variety | The combined form of diversified data is utilized to enhance the processing of big data analytics and processing. Thereby, the privacy issues are decreased with the data utility rate. |
| Veracity | The more reliable and accurate data available with high level of utility rate. The integrity of the data is devastated by the entrusted data source. Accurate data offers meaningful and high rate of utility. |
| Value | High utility value achieves best potential rate whereas data value plays significant role in preserving the privacy. |
| Variability | Most of the inconsistencies of the data namely extreme values, outlier and negative accurate analysis were rectified. Variability is one of the important concerns in the privacy domain. |
| Vulnerability | Big data is holds several personal information and it needs privacy preserving approaches in order to prevent the data from exposure to the unknown or any third parties. |

## 2. Classification of Bipolar Disorder

This section discusses about the classification of bipolar disorder using improved stacked auto encoder and deep neural network. The overall proposed methodology is illustrated in Figure 1.

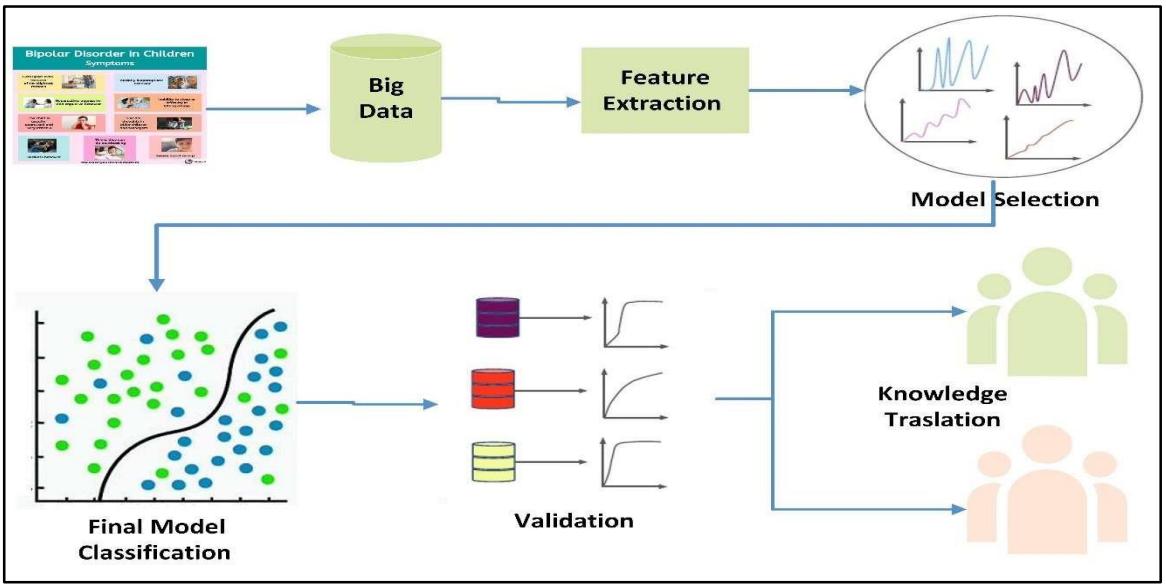


**Figure 1.** Overall Proposed Methodology

Auto encoders are used to build improved stacked autoencoders (SAE). In a neural network, the hidden layer of every auto-encoder is linked to the hidden layer of the following auto-encoder. The hidden layer of the first auto-encoder should be the input to the second one during the training phase. The improved stacked encoder is given in Figure 2.

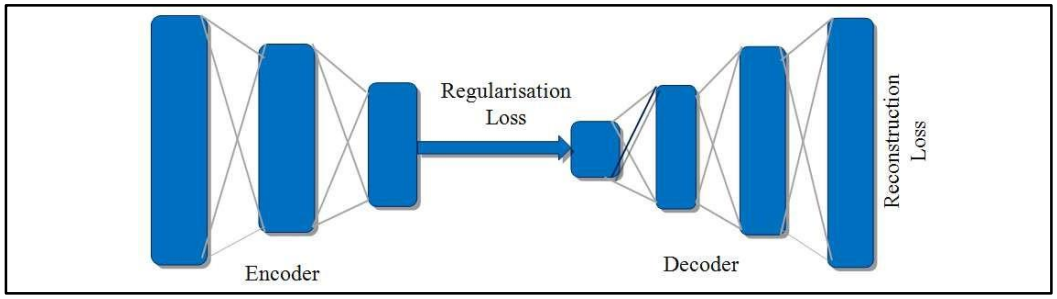


**Figure 2.** Improved Stacked Auto Encoder

New representations utilizing the improved SAE is constructed by piling them on upper edge of older ones. The high-level properties of the datadata are included in the eventual hidden layer output after rebuilding. The qualities of an item in the object field determine how conductivity is distributed. To calculate the conductance distribution, the layer of logistic regression is employed. Figure 2 indicates the structure of improved stacked auto encoder.

The DNN receives the normaliseddata choices as input. The symbol for selected data is denoted by S = {SG(1), SG(2),..., SG(A)}, where A stands for the amount of training sets and SG(l) ∈ [0,1]a is the count of normaliseddatas. In a set of randomly selected data sequences, the letter a denotes the presence of an undetermined count of data values. Distribution sample for internal connectivity is (S

= {σ(1), σ(2), …, σ(a)}), where σ(l) ∈ [0,1]n and n indicate the possibility of every class. The bias matrices, vectors, and weight are seeded via unsupervised layer-by-layer training. The DNN is assigned with the responsibility of processing the data with the rate S = {SG(1), SG(2),…, SG(A)}.

The whole method is elucidated below: It's crucial to learn the initial hidden layer by utilising the outcomes of the preceding one. Until all hidden levels are revealed, the similar process is repeated. After the process of feature selection using ISAE, classification is done using Deep Neural Network. During the supervised fine-tuning step, the pre-trained model parameters from a DNN's last hidden layer are input into a model of logistic regression to activate the whole DNN. An authentic representative group of the conductivity distribution serves as the inspiration for the network's name. The back-propagation algorithm used in the top-down method of optimising network parameters is based on the principle of gradient improvement. Dropout may be used to improve a model by lowering overfitting and making it more generalizable. Through every training program, 0.5% of the network's hidden units are arbitrarily removed from the network.

Neuronal coadaptation can be made easier, leading to the development of a more robust network. When trained on big datasets, the dropout layer works superbly. Equations (2) and (3) are affected by dropout, as can be observed in a standard auto-encoder.

$$y_i = (\sum_{j=1}^{a} w_{ij}\, Bernoulli(p) \times x_i + b_i) \text{----------(2)}$$

$$z_i = (\sum_{j=1}^{a} w^T\, Bernoulli(p) \times x_i + b_i) \text{----------(3)}$$

A randomized vector either of zero or one with a frequency of p equivalent to 0.5 is produced by the Bernoulli() operator.

During the process of training weights are updated in every iteration that improve the features of the encoder

output. The proposed framework is trained for limiting both the reconstruction and regularisation loss that is given in equation 4.

$$\min_{\theta,\emptyset} \beta F_{rec} = \min \frac{1}{2} \sum_{i=1}^{n} \| x_i - g_\emptyset (f_\theta (x_i)) \|^2 \text{----------} (4)$$

where the encoder and decoder parameters are indicated by $\theta$ and $\emptyset$, respectively. The hyper parameter $\beta$ stabilises significance of losses to the reconstruction of data. Equation (4) established the term Frec, which is then minimised to maximise the resemblance among the input and output data and enhance the latent space description.
SVM employs kernel trick to identify new features for conversion from low dimensional feature space to higher dimensional feature space. In the application of SVM classifier, the parameters should be tuned to reduce the tuning time and to overcome the over-fitting problem. SVM builds a model using a set of labelled training samples and classifies the new samples on the basis of distance to the hyper plane. Training of the samples is performed by minimizing the error function EF as presented in Equation 1

$$EF = \frac{1}{2} w^T w + c \sum_{i=1}^{N} s_i \text{----------} (5)$$

with the constraints $y_i(w^T \emptyset(x_i) + b) \geq 1 - s_i)$ $and$ $s_i \geq 0, i = 1,2,3,4, \ldots \ldots \ldots N$

C and b are constants; w is the vector of coefficients and $s_i$ denotes the factor for managing the non-separable input data. Index i is used to label the N training cases. Proper choice of C will avoid over- fitting problems. The categorization between malignant and non-cancerous samples will become poorer if β is high and the data reconstruction term dominates. However, if this term is set too low, the hidden space's features won't be properly optimised and the reconstructive losses will be minimal. As a result, the latent characteristics will be drastically varied from the input data, which lowering the generalization ability as well as accuracy. Consequently, the hyperparameter β have to be adjusted properly.

## 3. Result and Discussion

This section illustrates the numerical outcomes of the existing and proposed methodology. The existing approaches namely SVM, and CNN are compared with the proposed ISAE. These approaches are investigated using the classification accuracy.

Accuracy: Accuracy refers to how close the determined value from the classified occurrences is to the true value. The representation of quantitative bias and persistent flaws is known as accuracy. It is also the recognition (both TP and TN values) amongst number of the assessed classes, as well as the estimation's similarity to the true value. Variation between the outcome and genuine resulting values occurs when the lowest accuracy occurs. It's the proportion of successful fall detection to the number of information examined. The rate of accuracy is given n Figure 3 and Table 2. It's calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{----------} (6)$$

**Table 1.** Comparison of Accuracy

| eration Count | SVM | CNN | ISAE |
|---|---|---|---|
| 50 | 83 | 87 | 93 |
| 100 | 86 | 89 | 95 |
| 150 | 88 | 90 | 96 |
| 200 | 89 | 91 | 97 |

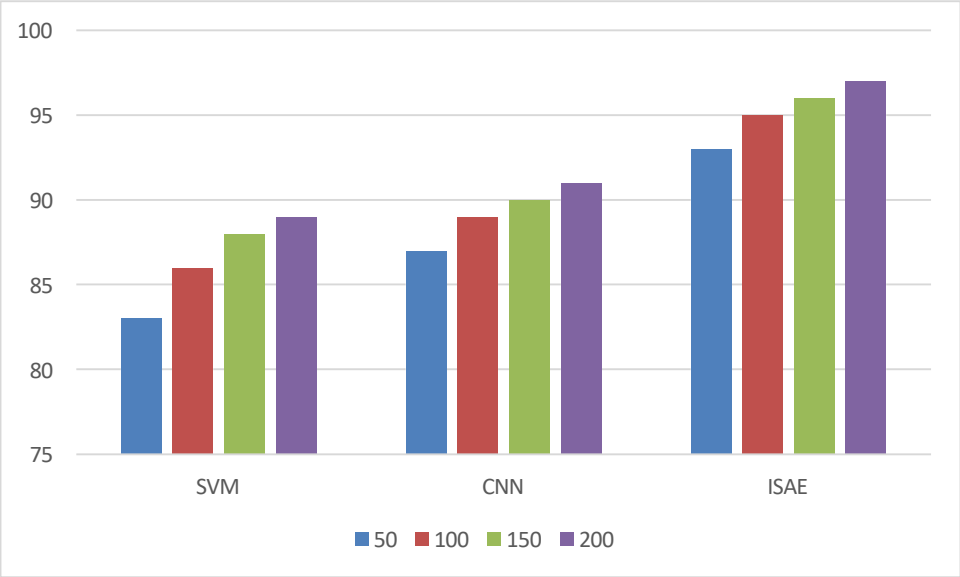


**Figure 3.** Comparison of Accuracy

## 4. Conclusions

By examining the relationships in the data and giving the data owners the necessary insights, data mining delivers the necessary support at this point. Also shared with outside parties for additional investigation are the major discoveries that were discovered. Numerous different types of information are processed in this condition, and the integration of big data and machine learning technology has become a crucial factor in the recognition of bipolar illness. It is a complicated hereditary condition that is characterised by manic and depressive periods. 1% of the world's population is impacted by it. It is a serious under-resolved public health issue that places a heavy strain on caretakers. High heritability and familial relative risk show that genetics play a part in the disorder's genesis. In this study, a deep neural network and an enhanced auto encoder are used to manage data. DNN and ISAE are used for feature selection and classification, respectively. The suggested method performs better than current state-of-the-art methods. The proposed approach attains 97% accuracy and outperforms SVM and CNN. In future, the approach can be extended with artificial intelligence based algorithms for attaining effective accuracy.

**References**

[1] Librenza-Garcia, D., Kotzian, B. J., Yang, J., Mwangi, B., Cao, B., Lima, L. N. P., ... &Passos, I. C. (2017). The impact of machine learning techniques in the study of bipolar disorder: a systematic review. *Neuroscience & Biobehavioral Reviews*, *80*, 538-554.

[2] Jan, Z., Noor, A. A., Mousa, O., Abd-Alrazaq, A., Ahmed, A., Alam, T., &Househ, M. (2021). The Role of Machine Learning in Diagnosing Bipolar Disorder: Scoping Review. *Journal of medical Internet research*, *23*(11), e29749.

[3] Claude, L. A., Houenou, J., Duchesnay, E., & Favre, P. (2020). Will machine learning applied to neuroimaging in bipolar disorder help the clinician? A critical review and methodological suggestions. *Bipolar disorders*, *22*(4), 334-355.

[4] Perez Arribas, I., Goodwin, G. M., Geddes, J. R., Lyons, T., & Saunders, K. E. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, *8*(1), 1-7.

[5] Passos, I. C., Ballester, P. L., Barros, R. C., Librenza-Garcia, D., Mwangi, B., Birmaher, B.,

[6] ... &Kapczinski, F. (2019). Machine learning and big data analytics in bipolar disorder: a position paper from the International Society for Bipolar Disorders Big Data Task Force. *Bipolar Disorders*, *21*(7), 582-594.

[7] Antosik-Wójcińska, A. Z., Dominiak, M., Chojnacka, M., Kaczmarek-Majer, K., Opara, K. R., Radziszewska, W., ... &Święcicki, Ł. (2020). Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. *International journal of medical informatics*, *138*, 104131.

[8] Nunes, A., Schnack, H. G., Ching, C. R., Agartz, I., Akudjedu, T. N., Alda, M., ... & Hajek,

[9] T. (2020). Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular psychiatry*, *25*(9), 2130-2143.

[10] Sonkurt, H. O., Altınöz, A. E., Çimen, E., Köşger, F., &Öztürk, G. (2021). The role of cognitive functions in the diagnosis of bipolar disorder: a machine learning model. *International Journal of Medical Informatics*, *145*, 104311.

[11] Tomasik, J., Han, S. Y. S., Barton-Owen, G., Mirea, D. M., Martin-Key, N. A., Rustogi, N.,

[12] ... & Bahn, S. (2021). A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data. *Translational psychiatry*, *11*(1), 1-12.

[13] de Siqueira Rotenberg, L., Borges-Júnior, R. G., Lafer, B., Salvini, R., & da Silva Dias, R. (2021). Exploring machine learning to predict depressive relapses of bipolar disorder patients. *Journal of Affective Disorders*, *295*, 681-687.

[14] Ceccarelli, F., & Mahmoud, M. (2022). Multimodal temporal machine learning for Bipolar Disorder and Depression Recognition. *Pattern Analysis and Applications*, *25*(3), 493-504.

[15] Wu, M. J., Mwangi, B., Bauer, I. E., Passos, I. C., Sanches, M., Zunta-Soares, G. B., ... & Soares, J. C. (2017). Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. *Neuroimage*, *145*, 254-264.

[16] Fernandes, B. S., Karmakar, C., Tamouza, R., Tran, T., Yearwood, J., Hamdani, N., ... & Leboyer, M. (2020). Precision psychiatry with immunological and cognitive biomarkers: a multi-domain prediction for the diagnosis of bipolar disorder or schizophrenia using machine learning. *Translational psychiatry*, *10*(1), 1-13.

[17] Li, H., Cui, L., Cao, L., Zhang, Y., Liu, Y., Deng, W., & Zhou, W. (2020). Identification of bipolar disorder using a combination of multimodality magnetic resonance imaging and machine learning techniques. *BMC psychiatry*, *20*(1), 1-12.

[18] Jadhav, R., Chellwani, V., Deshmukh, S., & Sachdev, H. (2019, January). Mental disorder detection: Bipolar disorder scrutinization using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 304-308). IEEE.

[19] Liu, Y. S., Chokka, S., Cao, B., &Chokka, P. R. (2021). Screening for bipolar disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. *Journal of Affective Disorders Reports*, *6*, 100215.

[20] Abawajy, J. H., Ninggal, M. I. H., &Herawan, T. (2016). Privacy preserving social network data publication. *IEEE communications surveys & tutorials*, *18*(3), 1974-1997.

[21] Cukier, K. (2010). Data, data everywhere: A special report on managing information. Economist Newspaper.

[22] Tene, O. (2011). Privacy: The new generations. *International data privacy law*, *1*(1), 15-27.

[23] Terry, N. P. (2012). Protecting patient privacy in the age of big data. *UMKC L. Rev.*, *81*, 385.

[24] Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges [guest editors' introduction]. *Computer*, *46*(6), 22-24.

[25] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., ... &Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, *2014*.

[26] Eisenstein, M. (2015). The power of petabytes. *Nature*, *527*(7576), S2.

[27] Zhou, M., Zhang, R., Xie, W., Qian, W., & Zhou, A. (2010, November). Security and privacy in cloud computing: A survey. In *2010 Sixth International Conference on Semantics, Knowledge and Grids* (pp. 105-112). IEEE.

[28] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

[29] Yang, Z., Zhong, S., & Wright, R. N. (2005, April). Privacy-preserving classification of customer data without loss of accuracy. In *Proceedings of the 2005 SIAM International Conference on Data Mining* (pp. 92-102). Society for Industrial and Applied Mathematics.

[30] Machanavajjhala, A., Kifer, D., Gehrke, J., &Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3-es.

[31] Li, N., Li, T., &Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k- anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106-115). IEEE.

[32] Tekiner, F., & Keane, J. A. (2013, October). Big data framework. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1494-1499). IEEE.

[33] Matwin, S. (2013). Privacy-preserving data mining techniques: survey and challenges. In *Discrimination and Privacy in the Information Society* (pp. 209-221). Springer, Berlin, Heidelberg.

[34] Li, X., Yan, Z., & Zhang, P. (2014, September). A review on privacy-preserving data mining. In *2014 IEEE International Conference on Computer and Information Technology* (pp. 769- 774). IEEE.

[35] Senosi, A., & Sibiya, G. (2017, September). Classification and evaluation of privacy preserving data mining: a review. In *2017 IEEE AFRICON* (pp. 849-855). IEEE.

[36] Wang, A., Wang, C., Bi, M., & Xu, J. (2018, June). A Review of Privacy-Preserving Machine Learning Classification. In *International Conference on Cloud Computing and Security* (pp. 671-682). Springer, Cham.

[37] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. *Ieee Access*, *2*, 1149-1176.

[38] Desai, S., Alhadad, R., Chilamkurti, N., & Mahmood, A. (2019). A survey of privacy preserving schemes in IoE enabled Smart Grid Advanced Metering Infrastructure. *Cluster Computing*, *22*(1), 43-69.

[39] Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE access*, *4*, 2751-2763.

[40] Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving.

*IETE Technical Review*, *34*(5), 544-560.

[41] Wang, T., Zheng, Z., Rehmani, M. H., Yao, S., &Huo, Z. (2018). Privacy preservation in big data from the communication perspective—A survey. *IEEE Communications Surveys & Tutorials*, *21*(1), 753-778.

[42] Lv, D., Zhu, S., Xu, H., & Liu, R. (2018, October). A Review of Big Data Security and Privacy Protection Technology. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)* (pp. 1082-1091). IEEE.

[43] Terzi, D. S., Terzi, R., &Sagiroglu, S. (2015, December). A survey on security and privacy issues in big data. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)* (pp. 202-207). IEEE.

[44] Ye, H., Cheng, X., Yuan, M., Xu, L., Gao, J., & Cheng, C. (2016, September). A survey of security and privacy in big data. In *2016 16th international symposium on communications and information technologies (iscit)* (pp. 268-272). IEEE.

[45] Pham, V. V. H., Yu, S., Sood, K., & Cui, L. (2017). Privacy issues in social networks and analysis: a comprehensive survey. *IET networks*, *7*(2), 74-84.

[46] Hu, J., &Vasilakos, A. V. (2016). Energy big data analytics and security: challenges and opportunities. *IEEE Transactions on Smart Grid*, *7*(5), 2423-2436.

[47] Tran, H. Y., & Hu, J. (2019). Privacy-preserving big data analytics a comprehensive survey. *Journal of Parallel and Distributed Computing*, *134*, 207-218.

[48] Tran, H. Y., & Hu, J. (2019). Privacy-preserving big data analytics a comprehensive survey. *Journal of Parallel and Distributed Computing*, *134*, 207-218.

[49] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR)*, *51*(4), 1-35.

[50] Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*(1), 8.

[51] Aggarwal, C. C., & Philip, S. Y. (2008). Privacy-preserving data mining: a survey. In *Handbook of database security* (pp. 431-460). Springer, Boston, MA.

[52] Wang, T., Xu, Z., Wang, D., & Wang, H. (2019). Influence of data errors on differential privacy. *Cluster Computing*, *22*(2), 2739-2746.

[53] Machanavajjhala, A., Kifer, D., Gehrke, J., &Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3-es.

[54] Li, N., Li, T., &Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k- anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106-115). IEEE.

[55] Dwork, C. (2006). Differential privacy, in automata, languages and programming. *ser. Lecture Notes in Computer Scienc*, *4052*, 112.