International Journal of Multidisciplinary Global Research
Vol. 01 (04), Oct-Dec 2024, pp. 45-55.
Online ISSN: 3107-4197

Supervised Learning used for Clinical Decision Support System

R. KARTHIKEYAN*

Dept. of CSE, University College of Engineering, Pattukkottai
Tamilnadu-614701, India
karthikeyanau2018@gmail.com

T. SATHIS KUMAR

Dept. of CSE, School of Engineering and Technology, Dhanalakshmi Srinivasan University,
Trichy, Tamilnadu-621112, India
sathiskumart.set@dsuniversity.ac.in

J. BRITTO DENNIS

Dept. of CSE, School of Engineering and Technology, Dhanalakshmi Srinivasan University,
Trichy, Tamilnadu-621112, India
brittodennis@dsuniversity.ac.in

Abstract

The privacy-preserving data publishing method resolves the obstacle of disclosing sensitive information when searching for valuable information. Anonymization techniques offer one of the most significant privacy assurances among the current methods. The issue of private data publication, in which many parties have disparate attributes for the same group of people, is discussed in this study. The request includes a method to securely combine confidential details about the person through multiple sources of information. A highly secure provider confidentiality protocol is needed while preserving the information to facilitate data mining jobs. This protocol is a sub-protocol associated with the exponential mechanism in a distributed environment.

Keywords: Elgamal encryption, Homomorphic encryption, SVM classification, and Anonymization methods.

1. Introduction

A decision-making system was developed to utilize patient data, pertinent clinical decision support systems, and computerized medical diagnosis processes that well-organized healthcare information to improve health and healthcare delivery. To create a clinical decision support system, biomedical engineering and artificial intelligence in machine learning transform one of the machine learning tools.

* Corresponding Author.

Enough reliable clinical data sets must be available for the system to function accurately, but these are only sometimes available. For instance, there are typically insufficient general practitioner (GP) surgery samples to cover all the disorders. Therefore, it is doubtful that a correct diagnosis can be made with minimal samples. The healthcare industry can benefit from the latest developments in remote outsourcing methods, such as cloud computing, offering precise and efficient decision support.

This disadvantage may impact the use of outsourcing methods in the healthcare industry. Additionally, look at The accessible medical information that will be incorporated towards treatment understanding with support vector machines (SVMs), among the tools for machine learning that have been utilized extensively in biomedical engineering to forecast various diseases. Training and testing are the two distinct stages that typically use an SVM. Features and training information will be utilized to train a model. Assigning labels to each unlabeled data sample that matches is another option during the evaluation process using the trained classifier.

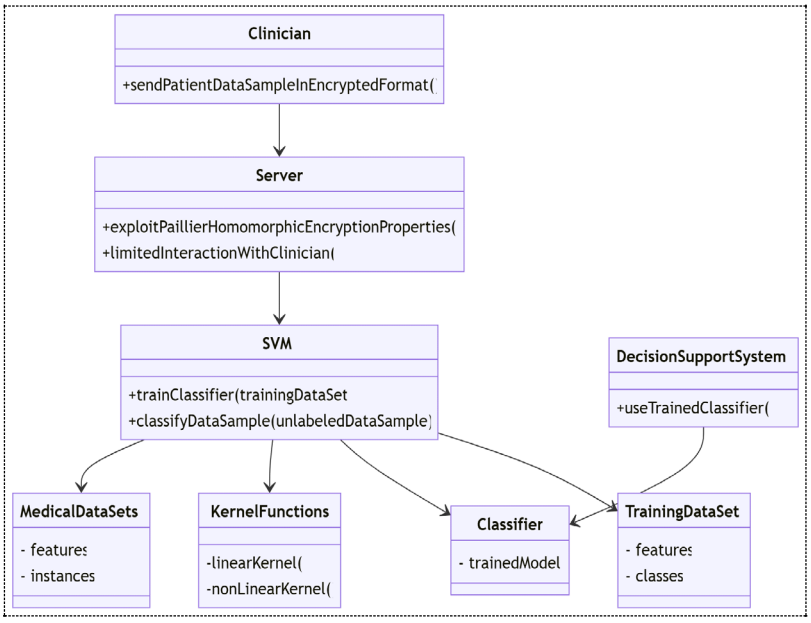

Fig. 1. An outline of a privacy-preserving clinical decision support system.

Figure 1 describes an outline of a privacy-preserving clinical decision support system. The accessible medical data set may be used to train a classifier. During the testing phase, the learned classification system can also serve as a decision-making system to make specific patient choices. The SVM uses several kernel techniques, such as straight and irregular kernels, based on the data's separability during training. If there are more features than instances, there's no need to convert the data to a

higher-dimensional space. The collecting of medical data is the precise reason why non-linear mapping does not improve performance.

A physician uses the technique displayed in Fig. 1 to transmit an encrypted small amount of patient data to the server over the global web. The clinician can then engage in limited interaction with the server through two-party secure estimation protocols. Alternatively, the server will use pallier homomorphic cryptography characteristics to execute tasks straight on the encrypted information when homomorphic characteristics do not support the actions.

2. Related Work

There are two stages: testing and training. Substantial data must be gathered for the first stage, which involves training a classifier. Numerous firms release their client data for both financial and research reasons. Publicating an individual's information (such as information about cancer hospital patients) could expose a person's identity and violate patient privacy.

2.1. *Support vector machine*

When it comes to data classification, SVMs have been utilized. The classification parameters w and b are obtained during training, the person's information, or the platform's evaluation in automated Learning. It uses a categorizing mechanism generated by learning the informational items. An SVM seeks to distinguish between classes. In the following subsection, we go over an SVM's classification function. This classification function is essential to create a decision support system that protects privacy.

2.1.1. *The linear approach classification of the problem*

Obtaining two Similar hyperplanes, $w1x + b = -1$ and $w1x + b = +1$, is the aim of linear classification; when it comes to data classification, SVMs have been utilized. While training, we collect patient data or the server's judgment with the classification parameters w and b. We may increase the separation between the hyperplanes by using distinct training data sets for each class. An unidentified sample for testing could be recognized during the training period.

3. Secure privacy decision-assisting system

Healthcare information is accessible remotely through Utilizing the Internet when defending security. As an outcome, we are looking at the client-server scenario in which the distant server is utilized in making choices through the Internet and gets server assistance to reach a decision. However, because of privacy concerns, the doctor is hesitant to share the individual's information, or the server's decision does not want to reveal any of the classification function's parameter values because doing

so would violate the security of the learning clinical information sampling who related to other people.

3.1. *Homomorphic the use of encryption*

Our analyses are founded on the Paillier cryptography system and use homomorphic encryption for concreteness without sacrificing generality. It is possible to employ any other homomorphic encryption technique. A decisional compositional residual challenge underlying the pallier cryptosystem's verified semantic protection is homomorphic encryption with the public keys method.

3.1.1. *Performance analysis*

Examine the suggested encrypted domain algorithm's performance. It compares the conventional plain-domain method and the proposed encrypted-domain method. The experiment uses two data sets from the UCI machine learning archive: the Wisconsin Breast Cancer (WBC) dataset and the Puma Indian Diabetic (PID) dataset. Both datasets are everyday (healthy) collections of information. Of the 681 samples in the WBC data set, 444 and 237 were all abnormal (harmful), whereas of the 768 samples in the PID data set, 500 were malignant, and 268 were benign. Except for the class label property, The WBC and PID data sets each involve nine features. A few instances of samples utilized for training obtained from WBC and PID information collection after normalization are included in Table II.

4. System Model

There are five system models:

- Input Selection,
- Data Normalization,
- Homomorphic Encryption,
- Decision Support Value Estimation and Client-side Decryption,
- Performance Evaluation.

4.1. *Input selection*

Our process's input module is the dataset selection module. The Pima Diabetes Dataset is the name of the dataset. This dataset pertains to medicine. It includes information about the patient's diabetes, including age, insulin level, and pedigree level. There are also attributes available in this class. The dataset was entered into the database following the selection. The dataset is preprocessed following loading.

This entails removing any undesirable characters or values from the collection. We can categorize the dataset values according to the class attribute present. Malignant and benign are the classification criteria. Unreliable data collection techniques frequently prepare information that may outcome in incomplete values, out-of-range numbers (including Earnings: -100), and even unattainable information

combinations (for example, Sex: Male, during pregnancy: Yes). Cleaning, normalization, transformation, feature extraction, and selection of the ultimate training set result from the information's purification. Kotsiantisetal (2006) presents every data preprocessing stage with a well-known method.

4.1.1. *Dataset normalization process*

The method of arranging elements and records in a database with relational elements so that the learning specimen values are similar. Decreased repetition is known as database normalization. Typically, normalization, which involves splitting records into smaller and bigger ones, reduces repetitive records and creates connections among them. Normalization stops sampling from having a big originating size out of skewing the answer by maintaining the numerical. The normalized information for training sets is represented by the notation Yi ∈ R n, and i = 1; P,

$$Yijnorm = Yij - \mu j / \sigma j \quad (1)$$

Depending on how separable the training data is, this problem can be further classified as linear and non-linear problem-solving.

4.1.2. *Homomorphism encryption technique*

The Paillier method uses a homomorphic encryption technique with an identification key length of 2048. The key size is very long. Because of this, hackers are unable to access the encrypted file. However, other homomorphic encryption algorithms, including the Paillier cryptosystem, might be applied. The decisional composite residuosity problem, which asks The Paillier system of cryptography, an in combination homomorphic encryption using public key methods, is conceptually safe if the value of z has an n-residue mod P2 during every component n (i.e., z = y n mod P2). Let p and q be two large integers prime, and let n = pq.

Using the Paillier cryptosystem, A piece of information m ∈ Z n could be encoded utilizing the Paillier encryption system as follows: JmK = g m r n mod n 2, where r ∈ Z ∗ n and g ∈ Z ∗ P 2. When g ∈ Z ∗ P2 and r ∈ Z ∗ n, JmK = g m r n mod p2. For sure, with encryptions, J m1 K and J m2 K, it encrypting J m1 + m 2 K the sum of m1 + m 2 Both the plain domain as well as encrypting J m1: K the good it is of m 1 Considering the constant value within the plain-domain can be efficiently measured in the encrypting area to be J m 1 + m 2 K = J m 1 KJ m. It makes Paillier a cryptosystem because it is, in addition, homomorphic.

4.1.3. *Encryption*

Compute cipher text as:
Let m be an encrypted message with m ∈ Zn., Choose r at random if r is in Zn*.

$$c = gm \cdot xn \, mod \, P2 \quad (2)$$

4.1.4. *Decision Support Value Estimation*

The user sends the encrypted file to the server. Some procedures are related to the encrypted file on the server. The server uses the Paillier-generated key to decode the file. The server uses the SVM classification algorithm to estimate the decision function value after decryption. As supervised learning models with corresponding learning algorithms, support vector machines (SVMs) sift through data, searching for patterns to use in regression and classification research.

Patient data is safeguarded from exposure, even if a server gets involved in the whole process. The encryption process is done using the clinician's public key, which everyone, including that server's information, can use to determine the values of the elements. A server must use homomorphic and two-party safe computation features in the domain's encryption, as it only has encrypted patient information.

4.1.5. *Decryption*

Decrypt the encrypted decision values that the server estimates during client-side decryption:

To decrypt, let c be the ciphertext, where c ∈ Z*c2. The plaintext message can be calculated as follows: m = L (c^λmodP2) · μ mod n

Decryption is "essentially one exponentiation modulo P2," as stated in the original publication.

4.1.6. *Performance Evaluation*

The performance is based on the complexity of computation and communication. To evaluate the better performance result by using the Paillier cryptography. The precision offered by the indicated encrypted-domain technique has been compared to that of a traditional plain-domain method to evaluate the accuracy. Choose the following data sets from the UCI machine learning repository: Puma Indian Diabetic (PID) and the Wisconsin state breast cancer data (WBC). While the PID information collection has 768 specimens, 500 are aggressive, and 268 are healthy.

The suggested technique's interaction cost depends significantly on Paillier encoding's length in our execution; an encrypted sample is 2048 bits in length. Sending a secure specimen for testing with N number of properties needs 2: 048 N K bits of throughput to exchange information channels. The system provides | S a certain amount of encrypted information (i.e., similar to accomplishing a profit in support vector), yet during the initial and last interaction, the intermediary mere a single data.

5. Proposed System

SVM—Support Vector Machine
Input: Both x and y have been provided for labeled training information; either α = 0 or α = in part trained it.

(1) C ⇐ is an amount (10, for instance)
(2) replicate
(3) for all {xi, yi}, {xj, yj} do
(4) Enhance αi & α
(5) end for
(6) While there are no modifications
Output: Retain only supporting variables (αi > 0)

5.1. *Characteristics with a homomorphic*

The Paillier encryption system's homomorphic quality is only one of its noteworthy characteristics. Because of the extra homomorphism of the encrypting activity, the following properties may be extracted.

5.1.1. *In addition to homomorphic plain text*

While multiple cipher texts are added up, the outcome will reflect the entirety number of the balancing plain texts.

$$D\ (E\ (P1, S1).\ E\ (P2, r2)\ mod\ D2)\ =\ P1 + P2\ mod\ n \tag{3}$$

While a cipher text and a plaintext have been multiplied by g, the outcome will reflect the total number of the corresponding plaintexts,

$$D\ (E\ (P1, S1).\ gP2\ mod\ D2)\ =\ P1 + P2\ mod\ n \tag{4}$$

5.1.2. *Multiplying by plaintexts homomorphic*

If the strength of one encrypted plaintext is increased to that of the other, it will be decoded for the final product of both plaintexts.

$$D\ (E\ (P1, S1)\ P2\ mod\ D2)\ =\ P1P2\ mod\ n \tag{5}$$

$$D\ (E\ (P2, r2)\ P1\ mod\ D2)\ =\ P1P2\ mod\ n \tag{6}$$

Furthermore, the item of plaintext with the value of k will have a decryption outcome for an encrypting plaintext raised more by a single value,

$$D\ (E\ (P1, S1)\ k\ mod\ P2)\ =\ kP1\ mod\ n \tag{7}$$

Still, no method currently exists for computing encryption for an item involving two communications using Paillier encryption methods, and a private key is unclear.

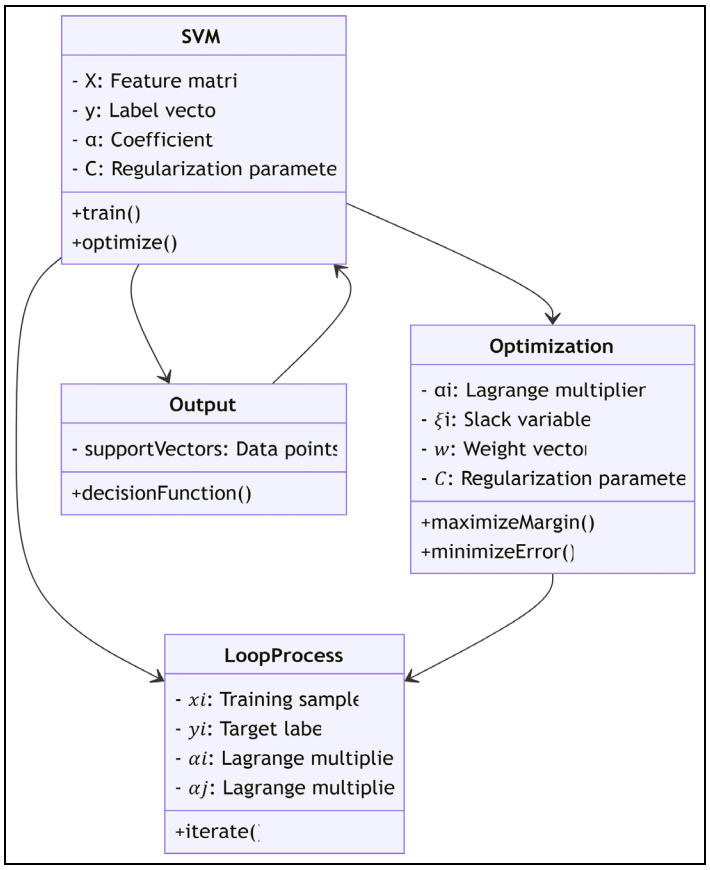

Fig. 2. Supervised Learning used for a decision support system in SVM.

5.2. *Elagamal encryption*

Take two huge prime numbers, such as p and q, randomly and distinct of each other to achieve gcd (pq, (p-1) (q-1)) =If both primes have the same length, this property is ensured. i.e., p, q ∈ 1 || {0, 1} s-1 about safety parameters.

- Find out n=pq and λ= lcm (p-1, q-1).
- Chose a number g from randomly with g ∈ Z*n2

Confirm that n splits into the correct sequence of g by evaluating for the existence of the next module multiplication reverse: μ = L (g λ mod n 2) a function is specified as follows L = (u-1)/n Note that the term (a/b) indicates the proportion of a and b, the most significant an integer value, instead of the modulo multiplier of periods its module multiplying reverse for v > 0and a > vb. The widely available secret key for encryption is (n,g), while the personal key is supplied by (λ, μ). When making use of p,q of equivalent size, an easier substitute for the mentioned before key generation stages would be: g= n+1, λ = Φ(n) and μ = Φ(n)-1 where Φ(n) = (p-1)(q-1). Let m represent the message that needs to be secret in C ε z n 2. Choose a random r to m =L

(c λ mod n 2). As pointed out in the primary research on decoding, it is "essentially one exponentiation modulo n 2.

6. Conclusion

Secure decision-support systems that protect user privacy using support vector machines with Gaussian kernels. By facilitating secure online access to large datasets of clinical information (or healthcare knowledge) held in off-site locations, the proposed algorithm may enhance healthcare providers' capacity to make informed decisions. New forms of outsourcing, including cloud computing, have made this a reality. Because the homomorphic properties of the Paillier cryptosystem are used, our approach takes advantage of this feature. You can only encrypt integer data using this approach. A novel approach is proposed to increase the process's continuous variables efficiently and privately. Results showed an accuracy of up to 97:21% when our method was evaluated on two sets of medical data, confirming its effectiveness. Crucially, our encrypted-domain technology prevents the disclosure of patient data to the remote server by keeping it encrypted throughout the whole process, including diagnosis.

7. References

Ajemba, P. O., Ramirez, L., Durdle, N. G., Hill, D. L., & Raso, V. J. (2005). A support vectors classifier approach to predicting adolescent idiopathic scoliosis's progression risk. IEEE Transactions on Information Technology in Biomedicine, 9(2), 276–282.

Amraee, S., Chinipardaz, M., & Charoosaei, M. (2022). Analytical study of two feature extraction methods compared with deep learning methods for classification of small metal objects. Visual Computing for Industry, Biomedicine, and Art, 5(13), 34. https://doi.org/10.1186/s42492-022-00111-6

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Transactions on Information Technology in Biomedicine, 14(4), 1114–1120.

Baskaran, V., Guergachi, A., Bali, R. K., & Naguib, R. N. G. (2011). Predicting breast screening attendance using machine learning techniques. IEEE Transactions on Information Technology in Biomedicine, 15(2), 251–259.

Chen, K., & Liu, L. (2005). Privacy-preserving data classification with rotation perturbation. In Proceedings of the 5th IEEE International Conference on Data Mining (pp. 589–592). Washington, DC, USA.

Damgård, I., Geisler, M., & Krigaard, M. (2007). Efficient and secure comparison for online auctions. In Proceedings of the 12th Australasian Conference on Information Security and Privacy (pp. 416–430). Townsville, Australia.

Fung, B. C. M., Wang, K., & Yu, P. S. (2007). Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering, 19(5), 711–725.

Gebre, A. M., Belete, M. D., & Belayneh, M. (2023). Object-based image analysis (OBIA)-based gully erosion dynamics, sediment loading rate, and sediment yield study in Lake Hawassa Sub-basin, Ethiopia. Natural Resource Modeling, 36(3), e12368.

Garg, A. X., Adhikari, N. J., & McDonald, H., et al. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. The Journal of the American Medical Association (JAMA, 293)(10), 1223–1238.

Indumathi, J., Shankar, A., Ghalib, M. R., Gitanjali, J., Hua, Q., Wen, Z., & Qi, X. (2020). Blockchain-based Internet of medical things for uninterrupted, ubiquitous, user-friendly, unflappable, unblemished, unlimited healthcare services (BCIoMT U6 HCS). IEEE Access, 8, 1–12.

Lisboa, P. J., & Taktak, A. F. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks, 19, 408–415.

Lin, K.-P., & Chen, M.-S. (2011). On the design and analysis of the privacy-preserving SVM classifier. IEEE Transactions on Knowledge and Data Engineering, 23(11), 1704–1717.

Mathew, G., & Obradovic, Z. (2011). A privacy-preserving framework for distributed clinical decision support. In Proceedings of the IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (pp. 129–134).

Musleh, A. A., & Jaber, H. S. (2021). Comparative analysis of feature extraction and pixel-based classification of high-resolution satellite images using geospatial techniques. E3S Web of Conferences, 318(1), 04007. https://doi.org/10.1051/e3sconf/202131804007

Nik Effendi, N. A. F., Mohd Zaki, N. A., Abd Latif, Z., Suratman, M. N., Bohari, S. N., & Zainal, M. Z., et al. (2021). Unlocking the potential of hyperspectral and LiDAR for above-ground biomass (AGB) and tree species classification in tropical forests. Geocarto International, 42, 1–26.

Pearson, S., & Charlesworth, A. (2009). Accountability as a way forward for privacy protection in the cloud. In Proceedings of the 1st International Conference on Cloud Computing (CloudCom) (pp. 131–144). Beijing, China.

Pearson, S., Shen, Y., & Mowbray, M. (2009). A privacy manager for cloud computing. In Proceedings of the International Conference on Cloud Computing (CloudCom) (pp. 90–106). Beijing, China.

Petrovska, B., Zdravevski, E., Lameski, P., Corizzo, R., Stajduhar, I., & Lerga, J. (2020). Deep Learning for feature extraction in remote sensing: A case-study of aerial scene classification. Sensors, 20(14), 3906. https://doi.org/10.3390/s20143906

Rahmadika, S., Astillo, P. V., Choudhary, G., Duguma, D. G., Sharma, V., & You, I. (2022). Blockchain-based privacy preservation scheme for misbehavior detection in lightweight IoMT devices. IEEE Journal of Biomedical and Health Informatics, 27(2), 1–12.

Rahulamathavan, Y., Phan, R. C.-W., Chambers, J. A., & Parish, D. J. (2013). Facial expression recognition in the encrypted domain based on local Fisher discriminant analysis. IEEE Transactions on Affective Computing, 4(1), 83–92.

Raphael, C.-W., & Phan, J. (2014). Privacy-preservation clinical decision support system using Gaussian kernel-based classification. IEEE Transactions, vol. no. 2, 227–248.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., et al. (2023). RingMo: A remote sensing foundation model with masked image modeling. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–22.

Sundareswaran, S., Squicciarini, A. C., & Lin, D. (2012). Ensuring distributed accountability for data sharing in the cloud. IEEE Transactions on Dependable and Secure Computing, 9(4), 555–567.

Xiao-Bai, L., & Sarkar, S. (2006). A tree-based data perturbation approach for privacy-preserving data mining. IEEE Transactions on Knowledge and Data Engineering, 18(9), 1278–1283.

Yuan, M., Chen, L., Yu, P. S., & Yu, T. (2013). Protecting sensitive labels in social network data anonymization. IEEE Transactions on Knowledge and Data Engineering, 25(3), 633–647.

R. Karthikeyan is working as a Guest Lecturer in the Department of Computer Science and Engineering at the University College of Engineering, Anna University, Pattukkottai Campus. He received a B.E. Degree in Computer Science and Engineering. Anna University, Chennai, in 2012 and received an M.E. Degree in Computer Science and Engineering. Anna University, Chennai, in 2015. Currently, he is pursuing PhD scholar (part-time) in Dhanalakshmi Srinivasan University Trichy. His areas of interest are IoT, Blockchain, Wireless Sensor Networks, Machine Learning, and Deep Learning.

T. Sathis Kumar is working as an Associate Professor in the Department of Computer Science and Engineering in (the School of Engineering and Technology) Dhanalakshmi Srinivasan University, Trichy. He holds a PhD from Anna University and has been actively involved in teaching and research. His areas of interest are machine learning, artificial intelligence, service-oriented architecture, and web technologies.

J. Britto Dennis is working as an Associate Professor in the Department of Computer Science and Engineering in (the School of Engineering and Technology) Dhanalakshmi Srinivasan University, Trichy. He holds a PhD from Anna University and has been actively involved in teaching and research. His areas of interest are IoT, Blockchain, Wireless Sensor Networks, Machine Learning, and Deep Learning.