BLOCKCHAIN BASED MULTI DISEASE PREDICTION USING SUPPORT VECTOR MACHINE ALGORITHM

Mrs. J. Dhanalakshmi, Assistant Professor, Dept.Of Computer Science, SSM Institute of Engineering and Technology, Dindigul-624002.
R. Periyasamy, S.Mohamed Fazil, K.S. Jayasuriya, UG Students, Dept. Of Computer Science, SSM Institute of Engineering and Technology, Dindigul-624002.

Abstract

The advent of Blockchain (BC) technology has become a remarkable, most revolutionary, and growing development in recent years. BT's open platform stresses data protection and anonymity. It also guarantees data is protected and valid through the consensus process. BC is mainly used in money-related exchanges; now it will be used in many domains, including healthcare; This paper proposes efficient Blockchain-based secure healthcare services for disease prediction. Diabetes and cardio diseases are considered for prediction. Initially, the patient health information is collected from Fog Nodes and stored on a Blockchain. The Machine learning algorithm is initially applied to the patient health records. Finally, diabetic and cardio diseases are predicted using classification based Support Vector Machine (SVM) algorithm. To evaluate the performance of the proposed work, an extensive experiment and analysis were conducted on data from the real world healthcare. The accuracy is achieved in better number in the prediction performance than the existing. The experimental results show that the proposed work efficiently predicts the disease.

Keywords:Support Vector Machine (SVM) algorithm, Fog Nodes

I. Introduction

Blockchain is one of the most innovative technologies and a digital wallet which retains track of transactions and events occurring across the network, and whose integrity is ensured via a peer-to-peer computing network, not by any centralized entity that might eliminate the risk of a single central point. It is composed of structured documents organized in a block structure that includes transaction batches and previous key hash. Every block is chronologically linked, and the data on the Blockchain network is unchallengeable Any users have individual access rights in a blockchain network to allow transactions that are modified throughout the framework, known as consensus protocol. For inserting transactions, a blockchain uses SHA256 hash. The NSA creates that, which is 64 characters large. All transactions are registered in a blockchain network though not modifying or manipulating the public ledger; Both transfers are distributed to various users across the network to transfer and update the data; a blockchain network may be duplicated toa separate venue, for example, within the same ability or healthcare distribution network, or as part of a regional or global data exchange system. A secure and privacy-conserving blockchain-based PHI networkingscheme was proposed for improving diagnosis in e-Health scheme. Private and consortium Blockchain is developed through the creation of their information structures and consensus mechanisms. The private ledger manages the PHI while the ledger community keeps a database of the robust indexes of the PHI.In recent years, healthcare practices across the country have accelerated their digital transformation efforts to modernize their operations, bake more efficiency into their workflows and processes, and deliver stronger patient experiences. While this digital evolution is a good and necessary thing, it also exposes practices to some significant challenges. As more of our healthcare processes transition to digital formats, providers need to be vigilant about security threats in healthcare.

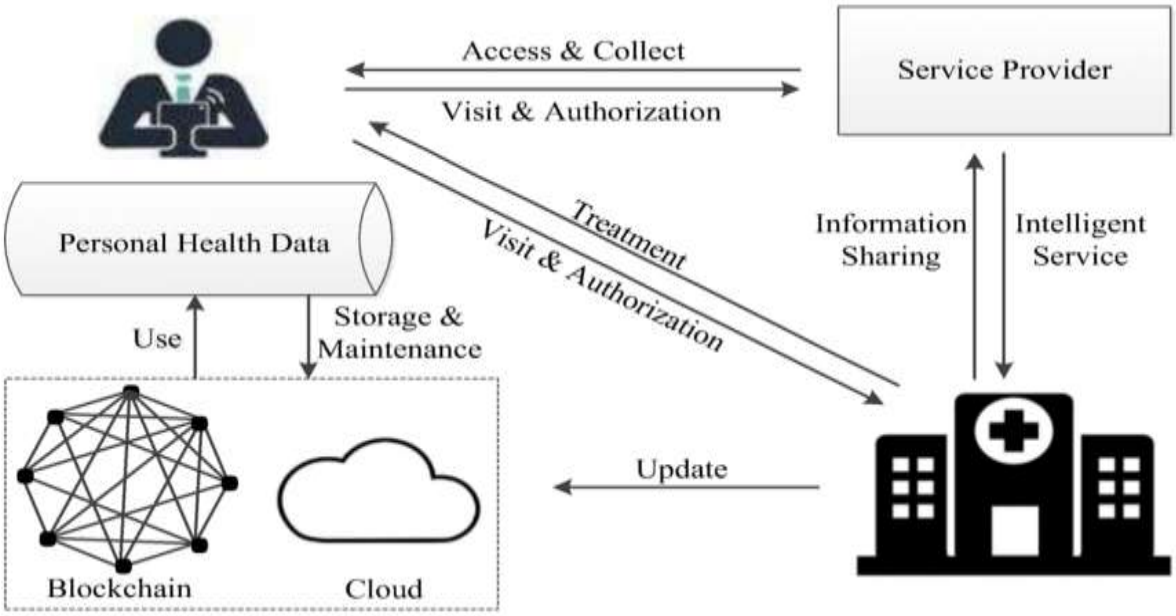

Figure 1: Architecture

Blockchain-based secure healthcare services for multiple disease prediction. In this project, the Diabetes and cardio diseases are considered for prediction. Initially, the patient health information is collected from Fog Nodes and stored on a Blockchain. The Support Vector Machine (SVM) algorithm is initially trained the patient health records. Firstly the patient used to register on this website and then login for the prediction. The patient has to select the disease to be predicted. The attributes are filled by the patient for a corresponding disease and then applied for the prediction. The SVM model is trained and provided a accurate result for a cardio and diabetic disease percentage. To evaluate the performance of the proposed work, an extensive experiment and analysis were conducted on data from the realworld healthcare. The experimental results show that the proposed work efficiently predict the disease.

1.1SVM algorithm

SVM has attracted a great deal of attention in the last decade. It also applied to various domains of applications. SVMs are used for learning classification, regression or ranking function. SVM is based on statistical learning theory and structural risk minimization principle. And have the aim of determining the location of decision boundaries. It is also known as a hyperplane. That produces the optimal separation of classes. Thereby creating the largest possible distance between the separating hyperplane. Further, the instances on either side of it have been proven. That is to reduce an upper bound on the expected generalization error. The efficiency of SVM based does not depend on the dimension of classified entities. Though, SVM is the most robust and accurate classification technique. Also, there are several problems. The data analysis in SVM is based on convex quadratic programming. Also, expensive, as solving quadratic programming methods. That need large matrix operations as well as time-consuming numerical computations. Training time for SVM scales in the number of examples. So researchers strive all the time for more efficient training algorithm. That resulting in several variant based algorithm. SVM can also extend to learn non-linear decision functions. That is by first projecting the input data onto a high-dimensional feature space. As by using kernel functions and formulating a linear classification problem. The resulting feature space is much larger than the size of a dataset. That is not possible to store on popular computers. Investigation of this issues leads to several decomposition based algorithms. The basic idea of decomposition method is to split the variables into two parts,a set of free variables called as a working set. That can update in each iteration and set of fixed variables. That are fix during a particular. Now, this procedure have to repeat until the termination conditions are met.The SVM was developed for binary classification. And it is not simple to extend it for multi-class classification problem. The basic idea to apply multi-classification to SVM. That is to decompose the multi-class problems into several two-class problems. That can address using several SVMs.SVM classification involves separation of data into classes by creating line or hyper plane. the main objective is to identify an ideal line or hyperplane that divides this dataset in two classes. Then, identify nearby line form both classes named support vectors. The distance between line to that support vectors called margin. The optimial hyper plane should be identified to maximize the margin values.

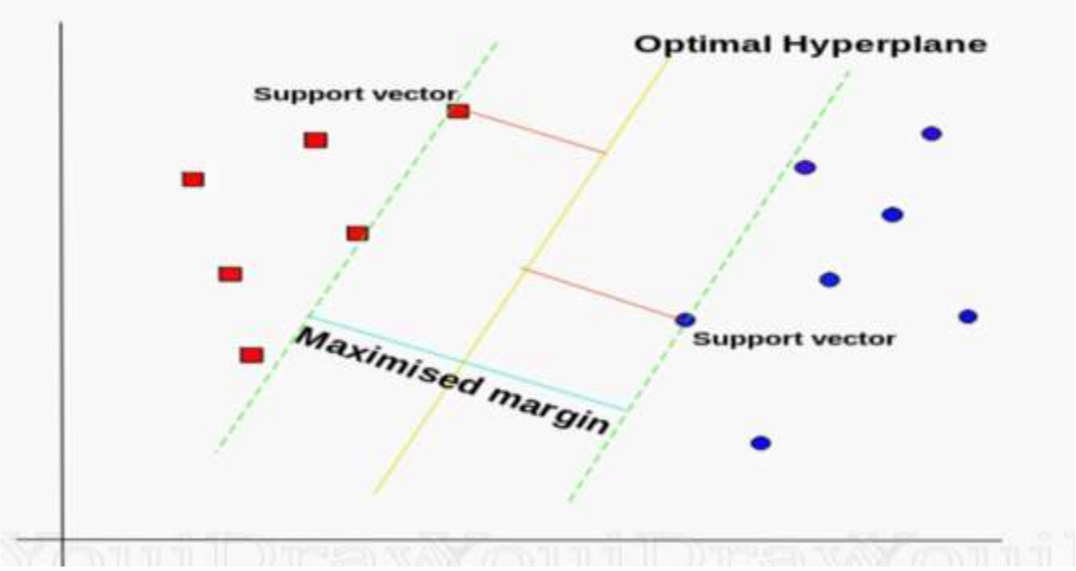

Fig 2: SVM Algorithm

By the training of labeled data set, SVM can classify the data with high speed and performance than other algorithms like neural networks. The performance of SVM depends upon the values of hyperparmeters . Before training , the values of hyperparameters adjusted to improve the classification accuracy of model. The SVM has following hyper parameters: kernel function, kernel function coefficient, penalty coefficient and polynomial degree. In this work, sailfish optimizer used to tune the hyper parameters of SVM in order to increase an accuracy level of classification.In short, support vector machine is a linear classifier. However, in some nonlinear problems, this model can also be used with some improvements [9], which are needed because not all data is linearly divided. This results in non-optimal results if linear SVM is still applied.The radial basis function (RBF) kernel was used to change the SVM modeling process from linear to non-linear. Generally, the RBF kernel is used for all types of data as a linear data separator. The RBF kernel has two parameters, namely Gamma and Cost.The Cost parameter is used for SVM optimization so that misclassification in the training dataset sample nghwaes not occur. Meanwhile, to measure the influence given by each training dataset sample, the Gamma parameter is used]. A low or high value is indicated by the use of this parameter. Low or high values are described as “far” and “near”. The formula below is the RBF Kernel equation:

K(x,z)=exp[−γ‖x−z‖2]Kxz=exp−γx−z2E4

II. Literature

Any users have individual access rights in a blockchain network toallow transactions that are modified throughout the framework, known as consensus protocol.For inserting transactions, a blockchain uses SHA256 hash.The NSA creates that, which is 64 characters large.All transactions are registered in a blockchain network though not modifying or manipulating the

public ledger; Both transfers are distributed to various users across the network to transfer and update the data; a blockchain network may be duplicated toa separate venue, for example, within the same ability or healthcare distribution network, or as part of a regional or global data exchange system.Ema, R. R et al proposed a new hybrid model based on Fuzzy C-means and Artificial Neural Networks (ANNs) with Principle Component Analysis that is capable to predict heart disease. The Principal Component Analysis is used to select the important features from the dataset. Then Fuzzy C-Means Clustering is used to cluster the extracted data from PCA and finally, Artificial Neural Network is used to predict Cardiovascular Disease. The simulation results confirm the effectiveness of the proposed method not only in terms of accuracy but also in terms of generalizability of the obtained models.Rahim, A et al proposed a MaLCaDD (Machine Learning based Cardiovascular Disease Diagnosis) framework for the effective prediction of cardiovascular diseases with high precision. Particularly, the framework
first deals with the missing values (via mean replacement technique) and data imbalance (via Synthetic Minority Over-sampling Technique - SMOTE). Subsequently, Feature Importance technique is utilized for feature selection. Finally, an ensemble of Logistic Regression and K-Nearest Neighbor (KNN) classifiers is proposed for prediction with higher accuracy. The validation of framework is performed through three benchmark datasets (i.e. Framingham, Heart Disease and Cleveland) and the accuracies of 99.1%, 98.0% and 95.5 % are achieved respectively. Finally, the comparative analysis proves that MaLCaDD predictions are more accurate (with reduced set of features) as compared to the existing state-of-the-art approaches. Therefore, MaLCaDD is highly reliable and can be applied in real environment for the early diagnosis of cardiovascular diseasesLi-Na Pu et al overviewed the eligible genome-wide association studies for CVD outcomes/traits . Clinical trials on CVD prediction using genetic information will be summarized from overall aspects. As yet, most of the single or multiple genetic markers, which have been evaluated in the follow-up clinical studies, did not significantly improve discrimination of CVD. However, the potential clinical utility of genetic information has been uncovered initially and is expected for further development.Pham, T. D. et al introduces a computational methodology for predicting such events in the context of robust computerized classification using mass spectrometry data of blood samples collected from patients in emergency departments. Applied the computational theories of statistical and geostatistical linear prediction models to extract effective features of the mass spectra and a simple decision logic to classify disease and control samples for the purpose of early detection. While the statistical and geostatistical techniques provide better results than those obtained from some other methods, the geostatistical approach yields superior results in terms of sensitivity and specificity in various designs of the data set for validation, training, and testing. The proposed computational strategies are very promising for predicting major adverse cardiac events within six months.Park, H. D et al propose a frequency-aware based Attentionbased LSTM (FA-Attn-LSTM) that weighs on important medical features using an attention mechanism that considers the frequency of each medical feature. Our model predicts the risk for cardiovascular disease using the ejection fraction as a prediction target and shows RMSE = 3.65 and MAE = 2.49.

2.1 Cardiovascular risk prediction

Xu, S et al focus on practical problem of Chinese hospital dealing with cardiovascular patients' data to make an early detection and risk prediction. To better understand the prescription and advice in Chinese, basic natural language processing method was used to synonym recognition and attribute extraction in Ultrasonic echocardiography. After data preprocessing, over 50 data mining techniques was tested for real patents dataset. Totake full advantage of multi-methods and reduce bias, top 6sub classifiers was selected to form an ensemble system, adjusted voting mechanism was used to make a final result, which consists of risk prediction and confidence. System has a high precision of 79.3% for 2628 cases of real patents in experiment. Therisk prediction confidence and algorithm accuracy shows great significance in practical use for doctors' diagnosing.Riggs et al. ,[26] propose smart, blockchain-based contractsto enable secure medical sensor research and management.The author built a network based on the Ethereum protocolusing a private blockchain where the sensorsconnect with a mobile computer that calls smart agreement and mark logs of every activity on the Blockchain.In [27], a blockchain-based system is introduced for safe, interoperable, and proficient access by patients,clinicians, and third parties to medical data while maintaining the confidentiality of personal details of patients. Through an Ethereum-based blockchain, it makes use of smart agreement to boostaccess control and code obfuscation, using advanced cryptographic methods for enhanced protection.

2.2 Clinical Implication of Machine Learning in Predicting

Joo, G et al assessed the effectiveness of various ML methods in predicting the 2-year and 10-year risk of CVD such as atrial fibrillation, coronary artery disease, heart failure, and strokes. To develop prediction models, we considered the usual medical examination data, questionnaire survey results, comorbidities, and past medication information available in the KNHSC data. We developed various ML-based prediction models using logistic regression, deep neural networks, random forests, and LightGBM, and validated them using various metrics such as receiver operating characteristic curves, precision-recall curves, sensitivity, specificity, and F1 score. Experimental results showed that all ML models outperformed the baseline method derived from the ACC/AHA guidelines for estimating the 10-year CVD risk, demonstrating the usefulness of ML methods. In addition, in our analysis, whether we included the past medication information as a feature or not, the prediction accuracy of all ML models was comparable to each other. Since the use of medications by the physicians provided important information on the occurrence of diseases, when we included it as a feature, all prediction models achieved a slightly higher prediction accuracy.presented as intervals of values for individual groups of surgical diseases and various age intervals. The total index is related to the determination of the risk of occurrence of cardiovascular complications of all levels of severity, and the lethal index to determination of the risk of lethal (severe) cardiovascular incidents. some common physiological attributes to identify a pattern among the people having a cardiovascular disease which, in further, has been used to distinguish whether a person has a risk of developing cardiovascular disease or not. To enhance the performance of the algorithm models, we have generated a secondary dataset based on the output of the classification model, pushing the accuracy of the model to 97.03%. We have also evaluated the correlation of the attributes to the chance of having cardiovascular disease and found some general observation. Producing a secondary dataset, the analysis leading to the observable patterns among the attributes and, defining general observation for cardiovascular disease using machine learning models make this study unique.

2.3 An explainable XGBoost–based approach

Athanasiou, M et al present study is to develop and evaluate an explainable personalized risk prediction model for the fatal or non-fatal CVD incidence in T2DM individuals. An explainable approach based on the eXtreme Gradient Boosting (XGBoost) and the Tree SHAP (SHapley Additive exPlanations) method is deployed for the calculation of the 5-year CVD risk and the generation of individual explanations on the model's decisions. Data from the 5- year follow up of 560 patients with T2DM are used for development and evaluation purposes. The obtained results (AUC=71.13%) indicate the potential of the proposed approach to handle the unbalanced nature of the used dataset, while providing clinically meaningful insights about the model's decision process

2.4 Early prediction of Cardiac Ailments

Bhatt, A et al focuses on analyzing cardiovascular health of rural and urban residents for early prediction of cardiac ailments through calcium score health indicator. Coronary Angiography is performed and Patients' Calcium Score results are taken randomly. Calcium score is also termed as Coronary Artery Calcium (CAC). This score is analyzed sex and age-wise in order to predict cardiovascular health issues at early stage. It is evident from the research study that males are affected more than twice of females by the cardiac health issues. The paper tries to figure out various factors affecting cardiac health among rural and urban residents of different age groups. The research outcomes motivates both rural and urban residents towards following a healthy routine and lifestyle in order to avoid such severity of cardiac health issues in future. A risk assessment model for patients, followed by the design and development of readmission risk assessment system for patients with cardiovascular disease. The risk assessment model includes three parts: risk prediction, clustering analysis and regression analysis of risk factors, which can automatically predicate the risk level and risk factors for the discharged patients in thirty days. The model was accurate 90.62% of the time. Combined the model assessment results with risk control knowledge base, a personalized health management and health guidance given by care workers can be put forward intelligently, which can not only help medical personnel in the rational allocation but also guide patients to carry out self-management better, resulting in the decrease of readmission rate

2.5 Neuro Fuzzy Inference System

Bhuvaneswari Amma N G et al proposed a medical diagnosis system to predict the risk of cardiovascular diseases with high prediction accuracy. This system is built using an intelligent approach based on Principal Component Analysis (PCA) and Adaptive Neuro Fuzzy Inference System (ANFIS). This system has two stages: In the first stage, dimension of heart disease dataset that has 13 attributes is reduced to 7 attributes using PCA. In the second stage, diagnosis of heart disease is conducted using ANFIS. In ANFIS, the learning capabilities of neural network and reasoning capabilities of fuzzy logic is combined inorder to give better prediction. The heart disease dataset used is Cleveland Heart Disease dataset provided by the University of California, Irvine (UCI) Machine Learning Repository. The obtained classification accuracy using this approach is 93.2%.A new algorithm, ModifiedBoostARoota, developed similar to BoostARoota, differing in the feature elimination process. Also, by choosing XGBoost and catboost as base models in both BoostARoota and ModifiedBoostARoota, a comparison of both the algorithms' performances are done. ModifiedBoostARoota algorithm has faster performance compared to BoostARoota, when catboost is chosen as the base model. Also, the XGBoost and CatBoost classifiers modelled on features selected by ModifiedBoostARoota gave better accuracy than that of BoostARoota.

III. Conclusion

In the current healthcare system, the use of Blockchain plays a crucial role. It can result in automated processes for collecting and verifying data, correcting and aggregating information from different resources that are indisputable, defiant to manipulation, and providing protected data, with condensed cybercrime chances and which also supports disseminated information, with system redundancy. This paper proposes efficient Blockchain-based secure healthcare services for disease prediction in fog computing. Diabetes and cardio diseases are considered for prediction. The proposed work efficiently cluster and predict the disease compared to other methods.

References

[1] Zhu, C.-Y., Chi, S.-Q., Li, R.-Z., Tong, D.-Y., Tian, Y., & Li, J.-S. (2016). Design and Development of a Readmission Risk Assessment System for Patients with Cardiovascular Disease. 2016 8th International Conference on Information Technology in Medicine and Education (ITME).

[2] Park, H. D., Han, Y., & Choi, J. H. (2018). Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease. 2018 International Conference on Information and Communication Technology Convergence (ICTC).

[3] Mostafa, N., Mostafa, N., Azim, M. A., Azim, M. A., Kabir, M. R., Kabir, M. R., … Ajwad,

R. (2020). Identifying the Risk of Cardiovascular Diseases from the Analysis of Physiological Attributes. 2020 IEEE Region 10 Symposium (TENSYMP).

[4] Pham, T. D., Honghui Wang, Xiaobo Zhou, Dominik Beck, Brandl, M., Hoehn, G., … Wong, S. T. C. (2008). Computational Prediction Models for Early Detection of Risk of Cardiovascular Events Using Mass Spectrometry Data. IEEE Transactions on Information Technology in Biomedicine, 12(5), 636–643.

[5] Li-Na Pu, Ze Zhao, & Yuan-Ting Zhang. (2012). Investigation on Cardiovascular Risk Prediction Using Genetic Information. IEEE Transactions on Information Technology in Biomedicine, 16(5), 795–808.

[6] Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access, 9, 106575–106588.

[7] Bhuvaneswari Amma N G. (2013). An intelligent approach based on Principal Component Analysis and Adaptive Neuro Fuzzy Inference System for predicting the risk of cardiovascular diseases. 2013 Fifth International Conference on Advanced Computing (ICoAC).

[8] Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020). Cardiovascular Disease Prediction Using Machine Learning Models. 2020 IEEE Pune Section International Conference (PuneCon).

[9] Loizou, C. P., Kyriacou, E., Griffin, M. B., Nicolaides, A. N., & Pattichis, C. S. (2021). Association of Intima-Media Texture with Prevalence of Clinical Cardiovascular Disease. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 68(9), 3017–3026.

[10] ` Bhatt, A., Kumar Dubey, S., & Kumar Bhatt, A. (2021). Systematic Cardiovascular Health Analysis of Rural and Urban Residents for Early prediction of Cardiac Ailments. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[11] Athanasiou, M., Sfrintzeri, K., Zarkogianni, K., Thanopoulou, A. C., & Nikita, K. S. (2020). An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus. 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE).

[12] Joo, G., Song, Y., Im, H., & Park, J. (2020). Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643–157653.

[13] Xu, S., Shi, H., Duan, X., Zhu, T., Wu, P., & Liu, D. (2016). Cardiovascular risk prediction method based on test analysis and data mining ensemble system. 2016 IEEE International Conference on Big Data Analysis (ICBDA).

[14] P, A., & Kalyani David, V. (2021). Feature selection using Modified Boost A Roota and prediction of heart diseases using Gradient Boosting algorithms. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).

[15] Ema, R. R., & Shill, P. C. (2020). Integration of Fuzzy C-Means and Artificial Neural Network with principle Component Analysis for Heart Disease Prediction. 2020 11thInternational conference on computing, communication and networking technologies(ICCCNT).

[16] Mendonca, F. manihar, R. Pal, & Prabhu, S. U. (2019). Intelligent Cardiovascular Disease Risk Estimation Prediction System. 2019 International Conference on Advances in Computing, Communication and Control (ICAC).