# An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction

Jan Carlo T. Arroyo
College of Computing Education, University of Mindanao, Davao City, Philippines
Email: jancarlo_arroyo@umindanao.edu.ph

Allemar Jhone P. Delima
Graduate School, Cebu Technological University-Barili Campus, Cagay, Barili, Cebu, Philippines
Email: allemarjhone.delima@ctu.edu.ph

***Abstract*—Cardiovascular disease prediction has gained the spotlight in research for the past years. Data-driven techniques for prediction through machine learning techniques have paved the way for an increased prediction accuracy in detecting the disease to people. In this paper, the cardiovascular disease dataset comprising 70,000 instances with 12 variables was used for training and testing using Artificial Neural Networks (ANN). However, the drawback of ANN when determining layers and neurons to be used persist. This paper aims to optimize the performance of ANN, leading to an increase in prediction accuracy. To realize, the use of the Genetic Algorithm (GA) was observed. Simulation results revealed that the use of GA had improved the performance of ANN by 5.08 percentage points against the prediction accuracy of the lone ANN. Further, the GA-ANN prediction model outperformed the other machine learning algorithms in the prediction of cardiovascular disease.**

***Index Terms*—artificial neural network, cardiovascular disease prediction, genetic algorithm, hybrid algorithm, parameter tuning**

## I. Introduction

One of the prominent strategies for modeling complex non-linear functions and systems is with the use of Artificial Neural Networks (ANN) [1]-[3]. ANNs work by simulating the neuronal architecture and function of the human brain in processing inputs of information [4]. One important property of artificial neural networks is their ability to learn from a training set even without any preliminary knowledge on the correlation between the data [5], [6]. With this, neural networks can make inferences by recognizing patterns in the data and learn through computational procedures [7].

The use of machine learning techniques has brought radical changes in the field of medicine, such as clinical decision making and precision diagnosis [8] undertakings, among others. Moreover, the use of ANNs in healthcare, such as in disease prediction and classification [9]-[17], is prevalent up to this day [18]. In this study, cardiovascular disease, being the leading cause of morbidity and mortality globally, is investigated wherein the resolve for optimal preventive cardiology may be attained by predicting this disease in people [19].

Further, this paper employs artificial neural networks for predicting cardiovascular disease. However, despite the wide application of ANNs in various fields, selecting its optimal parameters remains the primary concern for researchers [20]. The most common problem in implementing neural networks typically involves how many layers and neurons should be used [21]. To solve this, tuning method is integrated that finds the best parameters for a neural network [22]. Tuning is usually a case of trial and error, as there is no explicit method or general rule in defining these parameters [20]. Moreover, parameter tuning may be very tricky and can be computationally expensive [23]. Nevertheless, doing so improves the stability and accuracy of the neural network [24].

To improve the accuracy of ANN in cardiovascular disease prediction, this paper employs an optimization strategy in the neural network. The parameter tuning method is achieved with the use of the genetic algorithm. In general, this study is hoped to contribute to the two significant areas of literature. First is on the disease prediction mining on how medical dataset can be processed and mined, and second, as a basis for an improved prediction model in the quest to achieve the optimal prediction accuracy. The rest of the paper is arranged as follows: Section II discusses the literature review, Section III includes the design and methodology used in the study, Section IV discusses the results while Section V highlights the conclusion, and Section VI shows the recommendations.

## II. Literature Review

The use of machine learning has a gained broad interest across various domains, such as in health and medicine. As a result, studies have been published focusing on the development of medical applications using multiple machine learning methods and techniques. For instance, the study of [25] presented the use of machine learning in assessing developed smart home sensors and predicted the cognitive health of individuals.

Manuscript received July 29, 2021; revised December 2, 2021.


doi: 10.12720/jait.13.1.95-99

A total of 179 perceived volunteer participants as the respondents were instrumental. A statistically significant correlation (r=0.79) between automated assessment of task quality and direct observation scores was observed. An AUC value of 0.64 was revealed using machine learning techniques to predict the participants' cognitive health.

Moreover, a convolutional neural network was utilized to recognize Chinese herbal medicine images [26], which obtained a 71% overall accuracy. More so, the use of Artificial Neural Networks (ANN) was also observed for detecting lung cancer [27] with an outstanding accuracy of 96.7%. Furthermore, a neural network for predicting diabetes was proposed in the study of [28], showing a good prediction accuracy of 87.3%.

However, while much literature focuses on the development of classifiers, there are also studies that focus on improving the status quo. For instance, the study of [29] proposed the use of Natural Language Processing (NLP) to enhance the training and testing of the prediction model for detecting depression symptoms. Results reveal that the proposed method improved the neural network's performance at 91% and 85% train and test scores compared to its baseline of 89% and 81%, respectively.

The paper of [30] proposed a method of improving the accuracy of a neural network through the Particle Swarm Optimization (PSO) technique. The study utilized a publicly available dataset with 303 instances of healthy and sick individuals. The dataset contained 72 features, but only 13 were used to undergo feature selection using PSO. After preprocessing the dataset, the PSO was used for feature selection and ranking. Out of the 13 features, ranking results identified the eight best features for the feedforward backpropagation approach to increase the accuracy in the neural network training. The results of the study reveal a 91.94% accuracy in predicting heart diseases through the hybrid PSO-FFBP neural network.

Moreover, the use of an enhanced Deep Neural Network (DNN) using Grey Wolf Optimization (GWO) algorithm and Principal Components Analysis (PCA) was observed for predicting diabetic retinopathy [31]. The GWO algorithm was utilized to extract relevant features from the image dataset and select the optimal parameters for the neural network. The dataset was also subjected to dimensionality reduction using the PCA. The results of the study show that the enhanced DNN using GWO and PCA classifier yielded the highest accuracy at 97.3%, precision at 96.5%, recall at 97%, sensitivity at 91%, and specificity at 97%, outperforming all 12 other prediction models in every measure.

Furthermore, the study of [32] proposed a hybrid approach to predict heart disease among patients using K-Means and artificial neural networks. First, the dataset with 14 features is clustered and organized into data groups using K-Means. Further, the identified clusters were used as input to the neural network. Based on the results, improved prediction accuracy of 97% was attained against 93% and 88% for the KNN and Naïve Bayes methods, respectively.

## III. METHODOLOGY

In this section, the proposed model to predict cardiovascular disease using an optimized neural network is introduced. The simulation of the proposed model was executed using Python 3 and TensorFlow 2 in an i7-7700HQ 16GB RAM GTX1050 4GB VRAM laptop.

### A. Dataset

A publicly available dataset for cardiovascular disease obtained from [33] with 70,000 instances collected from medical examinations was utilized. The dataset comprises 12 variables where variables 1 to 11 are the input features, while variable 12 is the output feature. The dataset was divided into 70% training, and 30% testing sets used in the neural networks. The indexed description of the dataset is presented in Table I.

TABLE I. DATASET FEATURES

| No | Variables | Type |
|---|---|---|
| 1 | Age | Int - days |
| 2 | Height | Int - centimeters |
| 3 | Weight | Float - kilograms |
| 4 | Gender | Categorical - 1: female, 2: male |
| 5 | Systolic Blood Pressure | Int - continuous |
| 6 | Diastolic Blood Pressure | Int - continuous |
| 7 | Cholesterol | Categorical - 1: normal, 2: above normal, 3: well above normal |
| 8 | Glucose | Categorical - 1: normal, 2: above normal, 3: well above normal |
| 9 | Smoking | Binary |
| 10 | Alcohol Intake | Binary |
| 11 | Physical Activity | Binary |
| 12 | Presence (or absence) of Cardiovascular Disease | Binary |

### B. Data Preprocessing

Before using the dataset as input to the neural network, the preprocessing step is first initiated. The following steps were conducted as follows:

1) Removing duplicate instances
2) Removing instances with missing values
3) Converting age from days to years
4) Converting height from centimeters to feet
5) Removing instances with extreme outliers

A total of 4,860 instances were discarded from the dataset after the above processes. Thus, only 65,140 remaining instances were used for the training and evaluation of the neural network.

### C. Genetic Algorithm and Neural Network

To come up with a structurally efficient neural network, the genetic algorithm was used to identify the most suitable network parameters. The proposed GA-ANN process is presented in Fig. 1. For a detailed discussion on how a neural network and the genetic algorithm works, see [34] and [35], respectively.

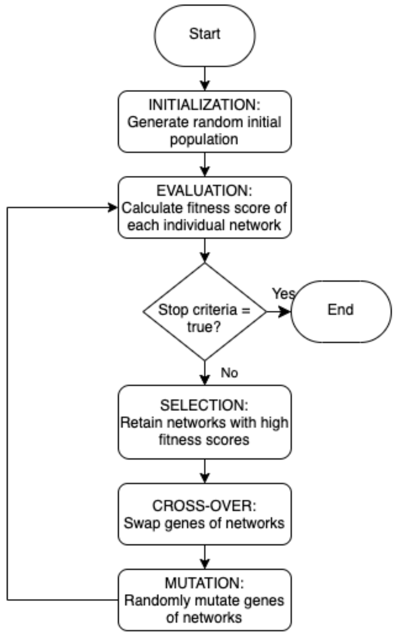


Figure 1. Proposed GA-ANN process.

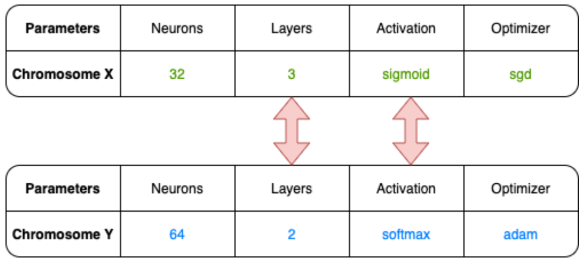


Figure 2. Crossover mating scheme.

## IV. Results and Discussion

In this study, a population of networks is evolved through 25 generations using the genetic algorithm. There are five primary phases for neural network optimization to wit:

- Initialization phase

The first generation starts with an initial population of randomly generated networks using a set of possible parameters. These sets of parameters characterize each network in the population. The parameters represent the genes of the chromosome, as presented in Table II.

TABLE II. Sample Chromosome

| **Parameters** | Neurons | Layers | Activation | Optimizer |
|---|---|---|---|---|
| **Chromosome X** | 32 | 3 | sigmoid | sgd |
| | Genes | | | |

- Evaluation phase

In this phase, each generated network from the previous stage is trained using the dataset. Further, each network is evaluated according to its accuracy identified as the fitness score.

- Selection phase

Moreover, the networks are sorted according to their fitness scores. Some of the networks are discarded, while others are retained. However, there is a certain probability that a rejected network is retained in the population.

- Cross-over phase

In the cross-over function, the selected parents undergo a mating process. The remaining networks are mated to produce new network offspring with varied genes. The probability that a particular network is selected for reproduction is based on their fitness score. The graphical representation of chromosomes X and Y being mated is shown in Fig. 2.

The layers and activation genes are exchanged, thus, forming new offspring with a new set of genes inherited from their parents, as presented in Table III. After offspring are produced, they are added as new members of the population, replacing discarded ones.

TABLE III. Generated Offspring

| Parameters | Neurons | Layers | Activation | Optimizer |
|---|---|---|---|---|
| Offspring 1 | 32 | ***2*** | ***softmax*** | sgd |
| | | | | |
| Offspring 2 | 64 | ***3*** | ***sigmoid*** | adam |

- Mutation phase

In this part, there is a certain probability that networks mutate, which means that some of their genes may be flipped to maintain diversity within the population. An example is shown in Table IV, where the optimizer gene of offspring one (1) is mutated from Stochastic Gradient Descent (SGD) to the Rectified Linear Unit (ReLU).

TABLE IV. Gene Mutation

| Parameters | Neurons | Layers | Activation | Optimizer |
|---|---|---|---|---|
| Offspring 1 | 32 | 2 | softmax | sgd |
| | | | | |
| Offspring 1 | 32 | 2 | softmax | ***relu*** |

With the integration of optimization techniques in ANN using the genetic algorithm, the possible set of parameters of the neural networks are revealed, as presented in Table V.

TABLE V. Network Parameters

| | |
|---|---|
| **Neurons** | 11, 24, 48, 64, 128, 256, 512, 1024 |
| **Layers** | 1, 2, 3, 4, 5 |
| **Activation function** | relu, elu, tanh, sigmoid, softmax |
| **Optimizer** | rmsprop, adam, sgd, adagrad, adadelta, adamax, nadam |

The genetic algorithm was executed for 25 generations, wherein a population of 20 networks per generation was depicted. Each neural network in every population was trained in batch sizes of 1,024 samples over 25 epochs or until the model stopped learning. The fitness scores of each individual in the population were identified based on the assigned parameters within the model generated by the neural networks.

Only 40% of the population with the best fitness scores were retained for every generation, and the other networks were rejected. There was a 10% probability that a rejected network was retained in the population during the selection process. The remaining networks in the population were selected as the parents of that generation for evolution. Moreover, a 20% probability for mutation was set in the study. After completing 25 generations, results were obtained, as shown in Table VI.

TABLE VI. INDEXED ACCURACY RESULTS PER GENERATION

| Generation | Average Accuracy |
|---|---|
| Generation 1 | 69.94% |
| Generation 5 | 72.95% |
| Generation 10 | 68.75% |
| Generation 15 | 73.04% |
| Generation 20 | 72.97% |
| Generation 25 | 73.137% |

After 25 generations, only the five best networks were selected, as shown in Table VII.

TABLE VII. SELECTED NETWORKS BASED ON ACCURACY

| Neurons | Layers | Activation | Optimizer | Accuracy |
|---|---|---|---|---|
| 64 | 3 | softmax | adagrad | 73.43% |
| 64 | 3 | softmax | adagrad | 73.38% |
| 128 | 3 | softmax | adagrad | 73.36% |
| 64 | 2 | softmax | adagrad | 73.35% |
| 64 | 2 | softmax | adagrad | 73.35% |

Results revealed the optimal parameters for training the neural network based on the features and instances of the dataset using the genetic algorithm. It can be observed that the prominent parameters are three (3) layers, 64 neurons, softmax function, and adagrad optimizer. It also shows that the average accuracy for the top 5 networks is at approximately 73.3%.

To test the effectiveness of the proposed model, the comparison of GA-ANN as against the other prediction algorithms, such as the Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, was undertaken. Table VIII shows the performance comparison between the proposed method and other machine learning algorithms using the same dataset. It can be observed that the GA-optimized neural network obtained the highest accuracy against the different prediction algorithms used. The study shows that the genetic algorithm is effective for optimization to increase the accuracy of a prediction model. Consequently, the proposed model is effective in predicting cardiovascular disease using the enormous patient dataset.

TABLE VIII. INDEXED SIMULATION ACCURACY RESULTS

| Prediction Model | Accuracy |
|---|---|
| GA-ANN | 73.43% |
| ANN | 68.35% |
| Logistic Regression | 72.35% |
| Decision Tree | 61.72% |
| Random Forest | 68.94% |
| Support Vector Machine | 72.16% |
| K-Nearest Neighbor | 68.34% |

## V. CONCLUSION

In this study, the use of the genetic algorithm to optimize the parameters for an ANN in predicting cardiovascular disease was implemented. Genetic algorithm, being one of the widely used optimization algorithms in the literature, is effective in selecting the best neural network parameters leading to an increased prediction accuracy of the ANN. Results revealed that the hybrid GA-ANN prediction model obtained the highest prediction accuracy as against the lone ANN, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and KNN algorithms with the prediction accuracy of 73.43%, 68.35%, 72.35%, 61.72%, 68.94%, 72.16%, and 68.34%, respectively. This shows that the ANN, when paired with an optimization algorithm, can be hybridized to increase its performance.

## VI. RECOMMENDATIONS

For future works, it is suggested to use the GA-ANN model in the different datasets and incorporate GA in other prediction models aside from ANN. Moreover, it is recommended that future researchers may use other optimization techniques to increase the performance of ANN extent on the accuracy of neural networks for prediction. It is also suggested to incorporate existing novel modified genetic algorithms in various prediction algorithms that need variable minimization or feature reduction processes for higher prediction accuracy.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

Jan Carlo T. Arroyo is responsible for the conceptualization, methodology, software utilization, investigation, and writing of the study's original draft and final manuscript. Allemar Jhone P. Delima directed, supervised, and provided suggestions for each stage of the research. He is also responsible for the writing, review, and editing of the paper. All authors had approved the final version of the manuscript.

## REFERENCES

[1] V. Khezri, E. Yasari, M. Panahi, and A. Khosravi, "Hybrid artificial neural network-genetic algorithm-based technique to optimize a steady-state gas-to-liquids plant," *Ind. Eng. Chem. Res.*, vol. 59, no. 18, pp. 8674-8687, 2020.

[2] L. Ohno-Machado and T. Rowland, "Neural network applications in physical medicine and rehabilitation," *Am. J. Phys. Med. Rehabil.*, vol. 78, no. 4, pp. 392-398, 1999.

[3] S. S. Vaezi, E. Poorazizi, A. Tahmourespour, and F. Aminsharei, "Application of artificial neural networks to describe the combined effect of pH, time, NaCl and ethanol concentrations on the biofilm formation of Staphylococcus aureus," *Microb. Pathog.*, vol. 141, p. 103986, 2020.

[4] E. M. Moreno, *et al.*, "Usefulness of an artificial neural network in the prediction of β-Lactam allergy," *J. Allergy Clin. Immunol. Pract.*, 2020.

[5] S. Savalia, E. Acosta, and V. Emamian, "Classification of cardiovascular disease using feature extraction and artificial neural networks," *J. Biosci. Med.*, vol. 5, no. 11, pp. 64-79, 2017.

[6] A. K. Shokhan and A. O. Sashcheko, "Artificial Neural Networks in Medicine," *Actual Problems of Modern Medicine and Pharmacy*, 2019.

[7] Z. Oner, M. K. Turan, S. Oner, Y. Secgin, and B. Sahin, "Sex estimation using sternum part lenghts by means of artificial neural networks," *Forensic Sci. Int.*, vol. 301, pp. 6-11, 2019.

[8] D. Itchhaporia, P. B. Snow, R. J. Almassy, and W. J. Oetgen, "Artificial neural networks: Current status in cardiovascular medicine," *J. Am. Coll. Cardiol.*, vol. 28, no. 2, pp. 515-521, 1996.

[9] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44-48, 2012.
[10] D. Wu, *et al.*, "Prediction of Parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization," *Int. J. Neural Syst.*, vol. 20, no. 2, pp. 109-116, 2010.
[11] A. Payan and G. Montana, "Predicting Alzheimer's disease a neuroimaging study with 3D convolutional neural networks," in *Proc. 4th Int. Conf. Pattern Recognit. Appl. Methods*, 2015, vol. 2, pp. 355-362.
[12] H. Wang, *et al.*, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J. Med. Imaging*, vol. 1, no. 3, p. 034003, 2014.
[13] P. M. Shakeel and G. Manogaran, "Prostate cancer classification from prostate biomedical data using ant rough set algorithm with radial trained extreme learning neural network," *Health Technol. (Berl)*, vol. 10, no. 1, pp. 157-165, 2020.
[14] K. Sujatha, B. C. Reddy, R. S. Ponmagal, and S. Q. Cao, "Analysis of CT images for detection of colorectal cancers using hybrid artificial neural networks and fire fly algorithm," *Procedia Comput. Sci.*, vol. 171, no. 2020, pp. 1517-1526, 2020.
[15] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Futur. Gener. Comput. Syst.*, vol. 111, no. 2020, pp. 17-26, 2020.
[16] N. B. Marya, *et al.*, "Application of artificial intelligence using a novel eus-based convolutional neural network model to identify and distinguish benign from malignant hepatic masses," *Gastrointest. Endosc.*, 2020.
[17] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," *Chaos, Solitons and Fractals*, vol. 140, 2020.
[18] J. Vargas, A. Spiotta, and A. R. Chatterjee, "Initial experiences with artificial neural networks in the detection of computed tomography perfusion deficits," *World Neurosurg.*, vol. 124, pp. e10-e16, 2019.
[19] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. V. D. Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423, 604 UK Biobank participants," *PLoS One*, vol. 14, no. 5, pp. 1-17, 2019.
[20] M. Bashiri and A. F. Geranmayeh, "Tuning the parameters of an artificial neural network using central composite design and genetic algorithm," *Sci. Iran.*, vol. 18, no. 6, pp. 1600-1608, 2011.
[21] D. Stathakis, "How many hidden layers and nodes?" *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 2133-2147, 2009.
[22] W. Y. Lee, S. M. Park, and K. B. Sim, "Optimal hyperparameter tuning of convolutional neural networks based on the parameter-setting-free harmony search algorithm," *Optik (Stuttg)*, vol. 172, pp. 359-367, 2018.
[23] Y. Wang, H. Zhang, and G. Zhang, "cPSO-CNN: An efficient PSO-based algorithm for fine-tuning hyper-parameters of convolutional neural networks," *Swarm Evol. Comput.*, vol. 49, pp. 114-123, 2019.
[24] K. G. Sheela and S. N. Deepa, "Review on methods to fix number of hidden neurons in neural networks," *Math. Probl. Eng.*, vol. 2013, p. 425740, 2013.
[25] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated cognitive health assessment using smart home monitoring of complex tasks," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 43, no. 6, pp. 1302-1313, Nov. 2013.
[26] X. Sun and H. Qian, "Chinese herbal medicine image recognition and retrieval by convolutional neural network," *PLoS One*, vol. 11, no. 6, pp. 1-19, 2016.
[27] I. M. Nasser and S. S. Abu-Naser, "Lung cancer detection using artificial neural network," *Int. J. Eng. Inf. Syst.*, vol. 3, no. 3, pp. 17-23, 2019.
[28] N. S. El_Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," *Int. J. Adv. Sci. Technol.*, vol. 121, pp. 55-64, 2018.
[29] U. Ahmed, S. K. Mukhiya, G. Srivastava, Y. Lamo, and J. C. W. Lin, "Attention-Based deep entropy active learning using lexical algorithm for mental health treatment," *Front. Psychol.*, vol. 12, p. 642347, Mar. 2021.
[30] M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and feed forward neural network," in *Proc. Artif. Intell. Robot.*, 2016, pp. 48-53.
[31] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," *J. Ambient Intell. Humaniz. Comput.*, 2020.
[32] A. Malav, K. Kadam, and P. Kamat, "Prediction of Heart disease using k-means and artificial neural network as hybrid approach to improve accuracy," *Int. J. Eng. Technol.*, vol. 9, no. 4, pp. 3081-3085, 2017.
[33] S. Ulianova. (2019). Cardiovascular Disease dataset. [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
[34] M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, Pearson Education Limited, 2011.
[35] O. Kramer, "Genetic algorithms," in *Genetic Algorithm Essentials*, Cham: Springer International Publishing, 2017, pp. 11-19.



**Jan Carlo T. Arroyo** is the chair of the Bachelor of Science in Information Technology program and an associate professor of the College of Computing Education at the University of Mindanao, Philippines. He graduated with a Bachelor of Science in Information Technology degree in 2012 and a Master in Information Systems degree in 2018 from the same institution. He is currently pursuing his Doctor in Information Technology degree at the Technological Institute of the Philippines.

**Allemar Jhone P. Delima** was born in Surigao City, province of Surigao del Norte, Philippines. He obtained his Doctor in Information Technology degree at Technological Institute of the Philippines – Quezon City. He earned his Master in Information Technology and Bachelor of Science in Information Technology degrees at Surigao State College of Technology, Surigao City, Philippines.
He is connected with Cebu Technological University (CTU) in Cebu, Philippines, as the University Director for Scientific Publications. He is also the Graduate School Director of CTU-Barili Campus. To date, he has presented several research papers at various international conferences. He has also published several research papers in indexed international journals. In addition to this, he is an active technical reviewer of several Scopus-indexed international journals. His research interests are in Data Mining, Machine Learning, and IT Security.
Dr. Delima is a recipient of the Commission on Higher Education (CHED) K-12 Scholarship for Graduate Studies-Local (SGS-L). He is an active member of different information technology and research associations in the Philippines.