International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 01-05, 2021, ISSN:-2319–8656

Research on Online Insulation Testing of Power Battery of New Energy Vehicles

Yuan Xu
College of Electrical Engineering, Southwest Minzu University,
Chengdu, China

Huazhang Wang
College of Electrical Engineering, Southwest Minzu University,
Chengdu, China

Jiacheng Li
College of Electrical Engineering, Southwest Minzu University,
Chengdu, China

Abstract: The insulation performance of new energy vehicles is an important factor in the normal operation of vehicles. This paper designs a voltage injection-type insulation detection based on the traditional detection.Based on the python language combined with the library provided by NI-visa, it can achieve high integration and meet the national GB/T 18384.1-2015 standard. Experimental results show that the insulation detection system can accurately test the insulation performance of new energy vehicles and meet the new energy vehicle offline detection standards.

Keywords: insulation test;new energy vehicles;power battery;insulation resistance;py-visa

1. INTRODUCTION

With the rapid development of the automobile manufacturing industry, domestic and foreign automobile manufacturers and major parts suppliers have shifted their business development focus to new energy vehicles, and the electrification of vehicles has become a development The inevitable trend [1].Whether it is a hybrid vehicle or a pure electric vehicle, the biggest feature is the use of high-voltage battery technology. The new energy vehicle power supply system is a typical AC/DC hybrid ungrounded system [2]. In order to meet high-power output, the output voltage of new energy vehicles is generally higher than 300V, and some even reach 720V.In severe cases, it may cause safety accidents such as leakage and fire, which will seriously threaten the personal safety of drivers and passengers [3].Strengthening the accurate detection of the insulation performance of the power battery of new energy vehicles is a very important technology to ensure the safety of drivers and passengers in new energy vehicles.

Python has the characteristics of object-oriented, concise and efficient, high portability, and good scalability, and is widely used in computer information processing[4].Py-visa is a Python package toolkit for NI-VISA, which provides support for the test system to implement different interfaces of similar instruments, accesses hardware attributes in the interpreter under the Python script environment, and uses the relevant information of the VISA driver library [5].This paper designs a new energy vehicle power battery online insulation detection system based on Python language and C language.

2. ON-LINE DETECTION METHOD OF INSULATION

The vehicle chassis is the main body of the connection. The insulated electrical system of the new energy vehicle chassis is shown in Figure1. The insulation performance of the vehicle high-voltage system is reflected by detecting the insulation of the positive and negative bus bars of the power battery of the new energy vehicle [6].The power of the high-voltage system of the new energy vehicle is provided by the battery pack, and the motor and the motor controller, air conditioning, brake and steering assist constitute the load electricity in the system. This article mainly introduces the insulation performance test of the power battery part. Common methods for testing the insulation performance of new energy vehicle power batteries include signal injection, balanced bridge, unbalanced bridge, and marginal insulation detection methods [7].

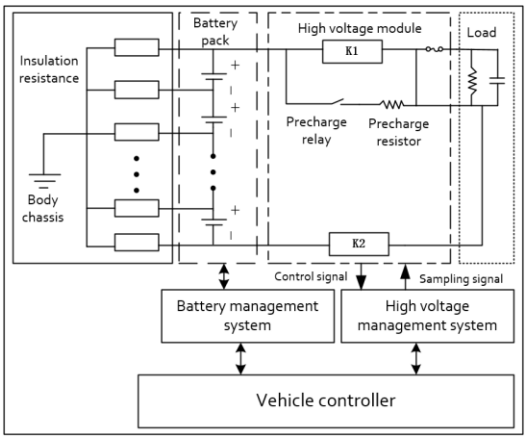

Figure. 1 Vehicle chassis insulation electrical system

The signal injection method of new energy vehicle power battery insulation detection is to transmit the signal to one end of the positive and negative charge and discharge interface of the power battery through a signal generator, load a Hall detection sensor [8] on the load, and connect the other end to an insulation detection device. In the state of good insulation performance, the load input current is equal to the detection output current, and the direction is opposite, and the Hall element has no signal output. When an insulation fault occurs inside the power battery, part of the AC current forms a current loop with the ground through the ground resistance of the insulation detection instrument, and the currents on both sides are no longer equal, and the Hall sensor sends out a voltage signal alarm. The injection of low-frequency signals will increase the DC voltage ripple of the vehicle system and affect the power supply quality of the DC high-voltage system.

The balanced bridge method means that the resistance values of the upper and lower bridge arms are artificially incorporated, so the circuit changes introduced into the vehicle system are also balanced. Although the structure of the balanced bridge method is simple, the upper and lower bridges may still maintain a balanced state when the

insulation resistance of the positive and negative bus change in the same proportion and grounded. At this time, errors are prone to occur when the insulation resistance is detected [9].

The unbalanced bridge method is to add the unbalanced resistance of the photoelectric control switch to the insulation detection circuit to detect the insulation resistance of the positive and negative bus bars to the ground [10]. The limitation of the unbalanced bridge method is that it can only be used when the bus is charged. It can't work normally when the bus is not charged.

The active insulation detection method uses the PWM signal to control the isolation transformer, and injects high-voltage DC signals between the positive and negative buses of the battery and the car body [11] to detect the insulation resistance. Although the insulation resistance can be detected when the positive and negative buses are not charged, however, the momentary high voltage during detection may have a great impact on the vehicle circuit itself.

3. DETECTION PRINCIPLE AND SYSTEM PLAN:

According to the GB/T 18384.1-2015 national standard, the insulation resistance of the power battery is usually defined as "if the power battery is short-circuited at a certain point between the stages, the resistance corresponding to the maximum leakage current". The standard states that the minimum insulation resistance of power batteries is 100Ω/V, and the safe value is 100-500 Ω/V. The insulation resistance value is divided by the nominal voltage of the DC system of the vehicle, and the result should be greater than 100V to meet the safety requirements. If the value is lower than this value, it is determined that the vehicle has an insulation failure [12].

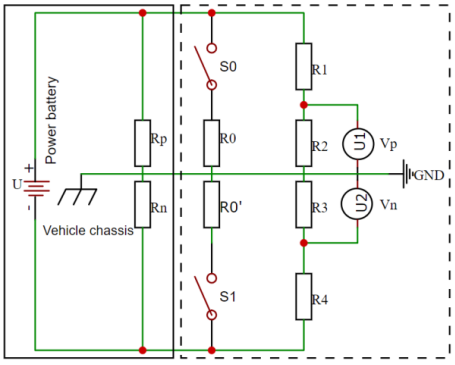

Fig. 2 Schematic diagram of equivalent circuit of vehicle chassis

The principle of the equivalent circuit for measuring the vehicle chassis is shown in Figure 2, where U is the power battery voltage, R_n and R_p represent the insulation resistance of the positive and negative bus bars to the ground, respectively. The dashed block diagram is a pure electric vehicle insulation resistance monitoring circuit model. Inside the dashed block diagram is a new energy vehicle power supply model, where R_0, R_0' are standard bias resistors, and R_0, S_1, S_2 form a bias resistor network. R_1 and R_2, R_3 and R_4 constitute a measuring voltage divider circuit, U_p is the positive voltage to ground, and U_n is the negative voltage to ground. When measuring, first disconnect S_1 and S_2 to obtain the voltage values U_p and U_n of the positive and negative bus to the ground, and then determine whether R0 is in parallel with R_p or R_n according to the magnitude of U_p and U_n.

If the measured value of U_p is greater than or equal to the value of U_n, close S_1 and open switch S_2, and measure a set of positive and negative bus-to-ground voltage values U_p and U_n'. The calculation of the insulation resistance value R_i of the DC high-voltage system can be obtained by the circuit principle The formula is as follows :

$$\frac{U_p}{R_p} = \frac{U_n}{R_n} \quad (1)$$

$$\frac{U'_p}{R_p // R_0} = \frac{U'_n}{R_n} \quad (2)$$

Solving the simultaneous formulas (1) and (2),

$$R_n = \left(\frac{U_p U'_n}{U'_p U_n} - 1 \right) R_0$$

Since $U_p \geq U_n$, then $R_p \geq R_n$, and R_i takes the smaller resistance value R_n.

When the total voltage of the battery pack is too low or the battery pack has an open circuit fault, the high-voltage DC signal is used to inject additional high voltage between the positive and negative buses and the car body through the isolation transformer, so that the insulation detection circuit can also be used when the positive and negative buses are not charged. Calculate the insulation resistance. Figure 3 is a schematic diagram of voltage injection insulation detection. The circuit in the dashed frame is the sampling circuit of the insulation detection device, and the injection signal generation part. R is the sampling resistor, and the appropriate sampling signal is obtained by adjusting the resistance of the rheostat. After obtaining the final selected resistance value in the experiment, it will be applied to actual working conditions in the form of fixed value resistance. In the injection signal generating circuit part, the single-chip microcomputer outputs PWM signal to control the on-off of the IGBT, and induces a 700V high-voltage signal on the secondary side of the transformer, and then obtains a DC high-voltage signal through RC filtering. The transformer uses a fly back converter with input and output isolation, and only one filter capacitor is needed for output filtering. At high voltage output, avoid high voltage inductors and high voltage freewheeling diodes. The power transistor is turned on at zero current and the turn-on loss is small. And the diode is turned off at zero current, and the reverse recovery problem can be ignored.

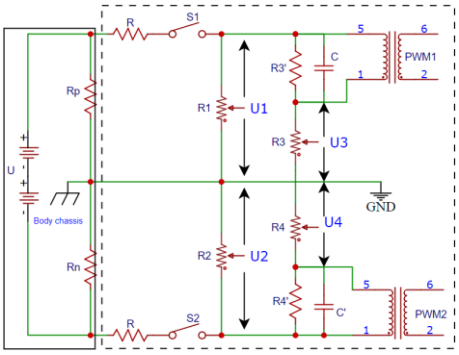

Fig. 3 Schematic diagram of voltage injection insulation detection

S_1 and S_2 are closed. Sampling the voltage of the resistors R_1 and R_2, after conversion, the voltages U_p and U_n on the ground insulation resistance of the positive and negative buses

can be obtained. Then the total voltage U of the DC system is obtained.

$$U = Up + Un \quad (3)$$

S_1 is closed and S_2 is open. Assuming that the voltage sampled at both ends of R_1 is U_1, combined with the current direction shown in the figure, the equation is obtained:

$$U = \frac{U_1(R_1 + R)}{R_1} + R_n\left[\frac{U_1}{R_1} + \frac{U_1(R_1 + R)}{R_1 + R_p}\right] \quad (4)$$

S_1 is closed and S_2 is open. Assuming that the voltage sampled at both ends of R_2 is U_2, the equation is obtained by combining the current direction in the figure:

$$U = \frac{U_2(R_2 + R)}{R_2} + R_p\left[\frac{U_2}{R_2} + \frac{U_2(R_2 + R)}{R_2 R_n}\right] \quad (5)$$

Combining equations (4) and (5) to solve the equations can obtain the insulation resistance of positive and negative buses to ground. Analyze the working process of insulation resistance detection with positive and negative buses with and without electricity. When the positive and negative buses are electrified, the switches S_4 and S_5 are controlled by the single-chip microcomputer, so that R_1 and R_2 are connected in parallel with the insulation resistance R_p and R_n of the positive and negative buses to be measured respectively [13]. The external measuring resistance $R_1=R_2=R_a$. Taking the negative pole of the battery as the reference ground, first sample the total battery voltage to obtain U. Then the S_4 switch is in the off state, and the measuring resistor R_1 is connected in parallel with R_P. At this time, the single-chip microcomputer collects the voltage of node 1 and obtains U_p. Open switch S4, close switch S_5, connect the external resistance R_2 and the resistance R_n to be measured in parallel to obtain U_n.

4 .HARDWARE DESIGN:

The insulation resistance monitoring system of pure electric vehicles mainly completes the three major functions of measurement, communication and cutting off the high-voltage circuit. In order to meet the system functional design requirements, reduce the interference caused by hardware circuit errors, and improve the convenience and high integration of user operations, the system hardware structure is shown in Figure 4. The main components include the main control system, measurement control instruments, and program-controlled power supply, High-voltage wiring harness, etc.

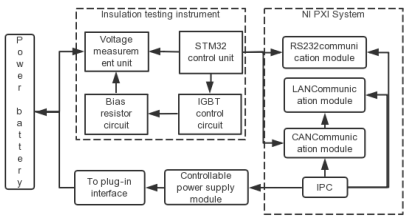

Fig. 4 hardware structure of insulation detection system

The main control system adopts the NI-based PXI expandable system to ensure the interoperability of instrument-level interface standards of different manufacturers' products. It can expand the control interfaces such as GPIB, LAN, RS232USB, etc. according to the difference of different instrument control interfaces [14]. In order to meet the characteristics of product diversification and various parameters, the hardware architecture of the debugging system adopts the "computer platform + control line" mode, which can realize interface standardization (including signal interface and hardware interface), test instruments can be interchanged, and test channels can be configured Wait.

The insulation detection instrument unit adopts the STM32 high-performance single-chip microcomputer introduced by ST, which itself has built-in 16-channel 12-bit high-speed ADC, which can simplify the hardware of the system [15]. The communication module reliably and efficiently realizes the internal and external communication between the detection system and the battery management system through the CAN bus. The main control chip drives a bridge unit composed of three elements K0, K1, and R to achieve the measurement of the voltage between the positive and negative bus to the ground; the bridge unit uses an IGBT switch with a withstand voltage of 1000V, which has a fast response speed. The voltage of the measured resistance is input to the external ADC after the resistance divider, and then digitally isolated into the single-chip microcomputer. The insulation resistance value is calculated by the collected voltage and the known standard resistance value. When the positive and negative bus of the power battery are not charged, the PWM drive circuit works to drive the transformer to generate high voltage and complete the insulation resistance detection when the positive and negative bus are not charged. The high-voltage cut-off module calculates according to the measured positive and negative bus-to-ground voltages against electric vehicle safety standards, and controls the high-voltage contactor to cut off the high-voltage circuit in time to ensure the safety of the insulation detection of electric vehicles.

5. SOFTWARE DESIGN AND TESTING:

Python is an easy-to-use scripting language. Based on the above principles, scripts are written in Python language through GPIB, RS232, Ethernet, USB and other interfaces to communicate with various measuring instruments and test equipment [16]. The entire software system adopts a modular design to achieve The integrity of the insulation detection cycle, the calculation of the insulation resistance value, the fault alarm and other functions. According to the overall function of the system, the whole set of software consists of the following modules: main program module; CAN and RS-232 communication module; data processing module;alarm program module; edit each parameter in order to be applicable to the protocol of different BMS manufacturers, such as The ID, byte position, scale factor and unit of the CAN message are edited and the XML file is finally generated [17]. The software flow chart is shown in Figure 5. Click the insulation safety detection module on the main interface to enter its sub-interface. The user first clicks the open button on the left to start the insulation detection hardware module, and then clicks the start detection button to display the current DC voltage and positive in the reading display column. Resistance of negative pole to ground. The main function modules of the system software platform include personnel login, product selection, debugging module, and report output data query. After the system is powered on, the system completes initialization and generates a square wave signal to control the push-pull circuit. In order to improve the sampling accuracy, the CPU judges the size of the feedback signal in real time. If the feedback signal is too small, the lift circuit is enabled. After the collection is completed, the insulation resistance

value is calculated, According to the calculated value, judge the insulation fault level alarm.

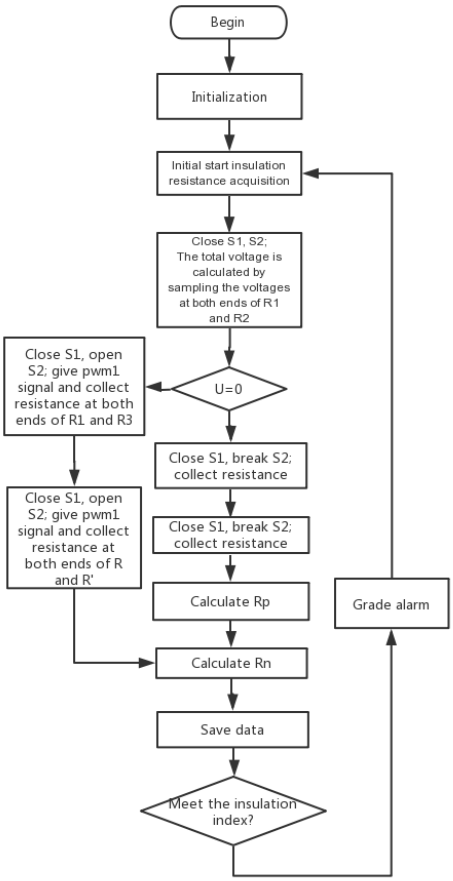

Fig.5 software flow chart

The insulation detection instrument is based on the C language programming of the STM32 single-chip microcomputer, carried out under the MDK integrated development environment, and mainly completes the AD measurement program. First, collect the total voltage U of the battery pack, and judge whether the bus is charged by the collected U. If U>0, use the balanced bridge method to detect the insulation resistance. After the insulation resistance Rp and Rn are calculated, the smaller one is compared with the threshold value. When the insulation resistance value is less than the threshold value, the insulation fault location is performed; when U= At 0 o'clock, the PWM signal is used to drive the DC source to boost injection to detect the insulation resistance.

6 CONCLUSION

Aiming at the shortcomings of traditional passive insulation measurement methods for new energy vehicles, this paper proposes a voltage injection insulation testing method. This method fundamentally solves the shortcomings that the symmetrical insulation fault of the power battery to the ground cannot be measured. The insulation performance between the circuit and the vehicle chassis. In the process of off-line testing of new energy vehicles, it can accurately conduct insulation testing of power TVs, improve off-line standards, and enhance vehicle driving safety. The next step of research will focus on real-time detection, improve its detection accuracy and reduce the size of the equipment, so that it can be integrated into the battery management system for secondary verification and real-time grasp of vehicle operating conditions.

7 .REFERENCES

[1] Zhu Minhui. Connecting the future helps the spare parts supplier behind China's new energy Weilai es8 [J]. Automobile and accessories, 2018, 000 (002): 50-51

[2] Liu Junfeng, Wu Jialei, Zhu Xiangkai, et al. A hybrid high voltage DC and high frequency AC pure electric vehicle electrical system: 2018

[3] Yang Kun, Yang Lin, Shi Yixin, et al. Single point insulation fault location method for electric vehicle power battery pack [J]. Automotive engineering, 2017, 039 (010): 1136-1140

[4] Huang Jianjun, Li Youmou, Liu Jing, et al. Design and implementation of automatic test system based on Python language [J]. Modern electronic technology, 2017 (04): 47-51

[5] Pu Tianhang. Research on instrument management and test system based on Python language [J]. China instruments, 2020, no.347 (02): 42-45

[6] Zhou Xingye, Zhu Jianxin, Chen Xiang, et al. Study on dynamic insulation resistance detection method of battery pack for hybrid electric vehicle [J]. Modern electronic technology, 2017, 40 (10): 121-124

[7] Zhang Xiangwen, Gao Guan. Real time online monitoring system for insulation resistance of electric vehicle power battery [J]. Journal of Jilin University (Engineering Edition), 2017 (05): 1395-1402

[8] Ni Hongjun, Chen Xiang, Zhou Xingye, et al. Detection method of dynamic insulation resistance of vehicle battery pack: 2016

[9] Mao Xingyu. Detection of short circuit or grounding short contact of phase line in fully enclosed bus duct by DC balanced bridge principle [J]. Installation, 1998, 000 (005): 31-32

[10] [2015.249. Electric current detection method of electric bridge based on Zhang Xingyang University [2015]

[11] Shi Wei, Jiang Jiuchun, Li SuoYu, et al. Study on SOC estimation method of lithium iron phosphate battery [J]. Journal of electronic measurement and instrumentation, 2010, 24 (008): 769-774

[12] Dong Haiyang, Yang Yuxin, Luo Yu, et al. Design of insulation resistance detection system for electric vehicles based on STM32 [J]. Electronic design engineering, 2019, v.27; no.417 (19): 186-189 + 194

[13] Zhu Feng. Design and development of instrument system software based on arm [D]. 2016

[14] Zhang Yue, Tao Linwei. Multichannel data acquisition system based on FPGA and STM32 [J]. Journal of Northwest Polytechnic University, 2020, v.38; No.182 (02): 128-135

[15] Chang Jiamin, Gao Fengmei, Wang Chongyang. Design of CAN bus communication between hybrid electric vehicle BMS and charger [J]. Industrial control computer, 2010, 23 (005): 45-46

[16] Zhou CF, Zhao Nan, Li Xin. An automated test system based on Python script language [J]. Journal of Nankai University (NATURAL SCIENCE EDITION), 2014 (05): 67-72

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 06-09, 2021, ISSN:-2319–8656

Product Ownership Management System (POMS) in the Post Supply Chain Using BlockChain

Hanmant D Magar
MIT World Peace University
School of Computer Engineering and Technology
Pune, India

Sandip Mane
Rajarambapu Institute of Technology
Islampur,India

Abstract – Nowadays, the authenticity of the RFID tags cannot be assured in the supply chain since these can be easily duplicated in the public space. We propose a novel Product Ownership Management System (POMS) of products for anti-counterfeits that can be used in the post supply chain by using the QR code. With the projected POMS, a consumer can reject the buying of counterfeits by scanning a QR code, if the seller does not have their proprietorship.
This paper gives an application of the system that will help to overcome the problems related with the presently functioning supply chain management system and runs the mechanism to show the ownership of the products.

Keywords: Product Ownership Management System (POMS), QR code, Anticounterfeits, Supply Chain Management

1. INTRODUCTION

Decentralized [3], the distributed system in which transactions are recorded in successive blocks making an immutable ledger is referred to as “blockchain”. A crypto currency network where anyone in the network can check the proof of possession of the balance or tokens is previously used in the blockchain. In our system, the concept of “proof of possession of balance” is replaced with an equivalent concept referred to as “proof of possession of products” within a supply chain.
Blockchain technology is supported by a distributed network consisting of a large number of interconnected nodes. Each of these nodes has its copy of the distributed ledger [1] that contains the full history of all transactions the network has processed. Blockchain in supply chain management is expected to boom over the next five years [2]. Blockchain will improve business for all global supply chain stakeholders by providing enhanced traceability, facilitating digitization, and securing chain-of-custody.
Supply chain traceability means corporations can handle the possession of products starting from the manufacturer to the current owner. The ability to trace a product throughout its life cycle supports risk management, fraud mitigation, quality assurance, worker rights, informed management decisions, and establishes direct responsibility for each link in the product life cycle.
Addressing the problem of supply chain traceability requires collaboration among stakeholders and deploying technical solutions to aid the transition. Blockchain is a nascent technology with a lot of hype that promises to disrupt status quo operations in many industries and supply chains. With the idea of this proposed system, counterfeits may be detected if a party cannot prove the possession of claimed products with the help of a quick response (QR) code [3].

2. LITERATURE REVIEW

Martin Westerkamp et al.[4] propose a blockchain-based supply chain traceability system using smart contracts. In such contracts, manufacturers define the composition of products in the form of recipes. Each ingredient of the recipe is a non-fungible token that corresponds to a batch of physical goods. When the recipe is applied, its ingredients are consumed and a new token is produced. The given mechanism preserves the traceability of product transformations. The system is implemented for the Ethereum Virtual Machine and it applies to any blockchain configuration that supports it.
The author of the paper [5] discusses how the traditional cloud storage model runs in a centralized manner, so a single point of failure might lead to the collapse of the system. The system is a combination of the decentralized storage system, IPFS, the Ethereum blockchain, and attribute-based encryption technology. Based on the Ethereum blockchain, the decentralized system has a keyword search function on the ciphertext solving the problem in traditional storage systems where cloud server returns wrong results.
A blockchain-based solution to address the problems of the supply chain such as Double Marginalization and Information Asymmetry etc is given in [6]. The SCM systems provide information sharing and analysis to companies and support their planning activities. The sharing of data between manufacturers, suppliers, and customers become very important to ensure reactivity towards market variability in Supply Chain Management (SCM).
Shanahan et al. [7] Suggested an RFID-based framework for beef traceability from farm to slaughter. By using RFID for the identification of individual cattle, this system was proposed as a solution to the inaccessibility of traceability records and fraudulent activities. To build an automated system this integrates online traceability data and chill chain condition monitoring information, Abad et al. [8] Tried to validate an RFID smart tag developed for real-time traceability and cold-chain monitoring of food under the case study of an intercontinental fresh fish logistics chain.
Christian Esposito [8] focuses on the various opportunities of blockchain for usage in the health-care sector. This paper proposes an Ethereum blockchain technology for a decentralized healthcare database.
From this paper, we get the general idea of how a decentralized network can be brought into effect for the Product Ownership Management System. This concept can be used to exchange data between various stakeholders in a supply chain.
Folinas et al [10] pointed out that the efficiency of a traceability system depends on the ability to track and trace individual product and logistics units, in a way that enables continuous monitoring from primary production until final disposal by the consumer.

3. PROPOSED SYSTEM

The proposed system provides a reliable and secure product transaction history using blockchain. Blockchain-based POMS in SCM that works in a decentralized environment based on consensus algorithm.

Figure 1 shows the architecture of the proposed system followed by the details working of the system.

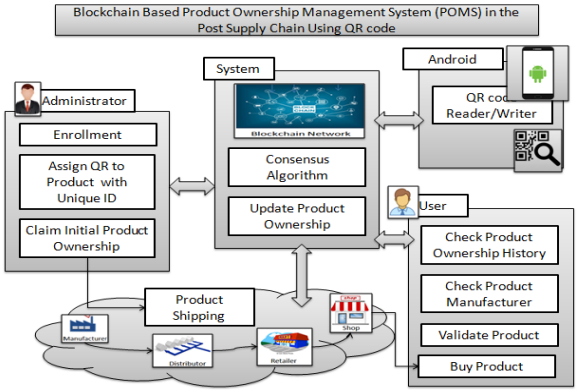

Figure: 1 System Architecture

3.1 Manufacturer Enrollment:

The manufacturer will enroll him with valid information on the blockchain with the company name and other details.

3.2 Assign QR code to the product:

The manufacturer will manufacture products and assigns a unique ID to each product with QR cod

3.3 Claim initial ownership of the product:

With the product, a unique ID admin can add a new transaction in blockchain to claim the initial ownership of the products. The detailed information of the product can fetch through the assigned QR code.

3.4 Shipping Product:

The manufacturer ships products to the Distributor where a distributor can check the manufacturer details and ownership of the products, etc. Distributor verifies the genuineness of the EPC using the assigned QR code and issues a transaction.

Ownership of the product will be transfer from the manufacturer to the distributor. Similarly, when any party receives products, a recipient follows the same procedure as above. At every time the product ownership transfer details are updated against the product's unique id through the QR code.

3.5 Check Product Ownership Using Blockchain:

Every product has its QR code, so through the QR code, we get the product history. The product history from manufacturing to shipping details is stored in the DB with the unique ID by using blockchain technology. So, after scanning the product QR code we get the unique identification number and through it, we get the overall product history.

3.6 Buying Product:

Customers should be able to buy products at the shop by validating the product information like the manufacturer of the product, Current owner of the product, etc using their assigned QR code. After that, the customer can buy and make a new transaction on the blockchain network if the product is valid or deny buying if fake product information is found in the product history.

4 .SYATEM FLOW

The below figure shows the overall flow of the system.

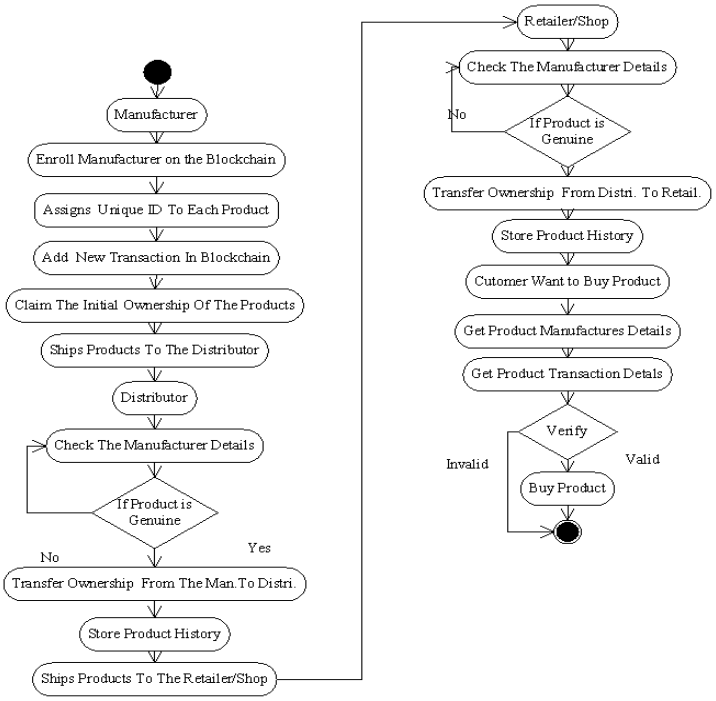

Figure: 2 System Flowchart

5. ALGORITHM USED

5.1 Consensus Algorithm:

In Consensus algorithm work in both decentralized and consortium way. Ethereum is used as a decentralized/permission-less consensus algorithm. Ethereum [1] [3] is a widely recognized and popular technology. Comparative protocols are likely also implemented on Ethereum, thus giving a better basis of comparison with protocols with similar goals.

5.1.1 Initialization Phase:

During the initialization of an election, a genesis contract must be placed on the blockchain (Algorithm 1).

Algorithm 1 Initialisation Phase

```
1: procedure ELECTIONGENESIS(_candidates,_pubk,_lengthPhaseOne,
   _lengthPhaseTwo,_cancelBallots)
2:     candidates ← _candidates
3:     pubk ← _pubk
4:     electionEndTime ← timeNow() + _lengthPhaseOne
5:     countEndTime ← electionEndTime + _lengthPhaseTwo
6:     cancelBallots ← _cancelBallots
```

5.1.2 Voting Phase (Initial Ballot):

To place a ballot on the blockchain, (Algorithm 2) the voter must have first communicated with the CA to receive a signed token authorizing the ballot.

Algorithm 2 Voting Phase - Initial Ballot

```
1: procedure PLACEBALLOT(vid,vote,msghashed, v,r,s)
2:     require((timeNow() < electionEndTime) And (verifyToken(msghashed,v,r,s))
3:     new InitialBallot(vid, vote)
4: procedure INITIALBALLOT(_vid,_vote)
5:     vid ← _vid
6:     vote ← _vote
7:     sealed ← true
8:     unsealedTimeStamp ← null
```

5.1.2 *Voting Phase (Altering Ballot):*

The process for pushing an altering ballot, (Algorithm 3) is similar to the process for creating an initial ballot.

Algorithm 3 Voting Phase - Altering Ballot

```
1: procedure PLACEALTERBALLOT(vid,vote,msghashed, v,r,s)
2:     require((timeNow() < electionEndTime) And (verifyToken(msghashed,v,r,s) And cancelBallots)
3:     new InitialBallot(vid, vote)
4: procedure ALTERINGBALLOT(_vid,_vote, _replacedBallot)
5:     vid ← _vid
6:     vote ← _vote
7:     replacedBallot ← _replacedBallot
8:     sealed ← true
9:     unsealedTimeStamp ← null
```

5.1.3 *Counting Phase:*

Once the election has concluded, votes will need to be counted (Algorithm 4).

Algorithm 4 Counting Phase

```
1: procedure RETRIEVEVOTE
2:     require(electionEndTime < timeNow())
3:     if isSealed then
4:         isSealed ← false
5:         unsealedTimeStamp ← timeNow()
6:     return(vote)
```

5.1.4 *Challenging Count:*

Nodes on the blockchain also have the functionality to examine the blockchain (Algorithm 5).

Algorithm 5 Challenging Count

```
1: procedure RETURNSEALED
2:     return(isSealed)
3: procedure RETURNTIMEUNSEALED
4:     return(unsealedTimeStamp)
```

5.2. EXPERIMENTAL RESULTS AND DISCUSSION

- USERS-

1. Company.
2. Manufacturer.
3. Distributor.
4. Retailer
5. End Users (Customer)

- Company sales their products to Manufacturer.

Following steps are that are taken to sale the product;

1. Selection of Product.
2. Changing the ownership

The ownership can be changed among;

- Company to Manufacturer
- Manufacturer to Distributor
- Distributor to Retailer
- Retailer to End Users

Example: -For transferring the ownership from Company to Manufacturer the following steps are performed;

- Transfer the generated OTP to the Manufacturer.
- Verify OTP of Manufacturer.
- The verified manufacturer will get the product's ownership.

But For transferring the ownership from Retailer to End Users the step changes as follows;

- View generated QR Code to User.
- Users can scan the QR code and get product ownership history.
- The user gets Ownership of the product.

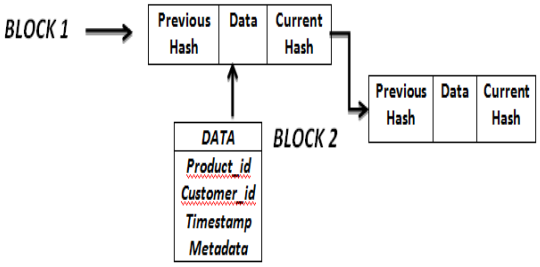

Figure 3: Structure of Block

- The above Figure elaborate the structure of blocks in the block chain.
- The block contains 3 parts **Previous Hash, DATA,** and **Current Hash**.
- The data in the block contain the following attributes:-

Product_id, Customer_id, Timestamp, Metadata.

- A separate block is generated for each transaction.
- For each product, a separate block chain is created.

6. CONCLUSION

Here we have developed a system that provides security to the product data and other details with blockchain technology in a manufacturing company. With the help of the system, we can also track the product ownership transaction history from the manufacturing stage to the end-user buying stage. With the use of this system, the customer can get valid product information when he wants to buy products. The proposed system is working for only verifying product ownership they do not give the transportation details. So, in the future, we implement a system that controls, monitors the product owner as well as the product transportation details. Also, we will use some hardware

for checking the product quantity while product transporting or shipping.

7. REFERENCE

[1] D. Johnson, A. Menezes, and S. Vanstone, “The elliptic curve digitalsignature algorithm (ecdsa),” International Journal of Information Security,vol. 1, no. 1, pp. 36–63, 2001.

[2] MELISSA J. RUSINEK, HAO ZHANG, AND NICOLE RADZIWILL,"Blockchain for a Traceable, Circular Textile Supply Chain: A Requirements Approach" SQP VOL. 21, NO. 1/© 2018, ASQ.

[3] Po-Yeuan Chang1, Min-Shiang Hwang, and Chao-Chen Yang1,"A Blockchain-Based Traceable Certification System",Springer International Publishing AG, part of Springer Nature 2018.

[4] Martin Westerkamp ,Friedhelm Victor,Axel Küpper,"Blockchain-based Supply Chain Traceability: Token Recipes model Manufacturing Processes", 2018 IEEE International Conference on Blockchain, At Halifax, Canada.

[5] UIGUO YU et al, “Authentication With Block-Chain Algorithm and Text Encryption Protocol in Calculation of Social Network, IEEE Access November 28, 2017.

[6] Mitsuaki Nakasumi, “Information Sharing for Supply Chain Management based on Block Chain Technology”, 19th Conference on Business Informatic, IEEE, 2017.

[7] Shanahan, C., Kernan, B., Ayalew, G., McDonnell, K., Butler, F., & Ward, S., A framework for beef traceability from farm to slaughter using global standards: an Irish perspective. Computer and Electronics in Agriculture. 2009. 66(1), 62-69.

[8] Abad, E., et al., RFID smart tag for traceability and cold chain monitoring of food: demonstration in an intercontinental fresh fish logistic chain. Journal of Food Engineering. 2009, 93(4), 394-399.

[9] Christian Esposito, Alfredo De Santis, GennyTortora, Henry Chang, Kim-Kwang Raymond Choo.” Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy”. IEEE Cloud Computing, January / February 2018.

[10] Folinas, D., Manikas, I., & Manos, B., Traceability data managementfor food chains. British Food Journal. 2006, 108(8),622-633. 11

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 10-13, 2021, ISSN:-2319–8656

4G LTE Network Coverage Optimization Using Metaheuristic Approach

Zikrie Pramudia Alfarhisi
Department of Electrical Engineering, Universitas Brawijaya
Malang, Indonesia

Hadi Suyono
Department of Electrical Engineering, Universitas Brawijaya
Malang, Indonesia

Fakhriy Hario Partiansyah
Department of Electrical Engineering, Universitas Brawijaya
Malang, Indonesia

Abstract: The main focus of this paper is to optimize the coverage of each 4G LTE network cell within the service area. There are many algorithms can be implemented to determine the optimal 4G LTE coverage area including the deterministic and heuristic approaches. The deterministic approach could solve accurately the optimization problem but need more resources and time consuming to determine the convergence parameters. Therefore, the heuristic approaches were introduced to improve the deterministic approach drawback. The methods used are the Differential Evolution Algorithm (DEA) and Adaptive Mutation Genetic Algorithm (AMGA), which are categorized as metaheuristic approach. The DEA and AMGA algorithms have been widely used to solve combinatorial problems, including for solving the network optimizations. In the network optimization, coverage is strongly related to 2 objectives, which are reducing the black spot area and decreasing the overlapping coverage areas. Coverage overlap is a condition when some cell sites in an area overlap. It implies in the occurrence of hand off and an inefficient network management. This research aims to obtain an optimal 4G LTE network coverage and reduce the overlapping coverage areas based on effective e-Node B arrangements by using the DEA and AMGA algorithms. The simulations results showed that the DEA algorithm's coverage effectiveness was 23,4%, and the AMGA Algorithm's was 16,32%.

Keywords: Adaptive Mutation Genetic Algorithm (AMGA), Differential Evolution Algorithm (DEA), network coverage, optimization, 4G LTE

1. INTRODUCTION

Coverage area is one of the main factors to maintain the communication of mobile station users in 4G LTE network. The adequate coverage area is indicated by the network's ability to cover the service area with good signal quality [1]. The modelling of cell sites in the service area is strongly related to the optimum level of coverage. In optimizations, coverage has two main objectives, which are the decrement of blank spot area and the reduction of the overlapping coverage areas [2]. Coverage overlap refers to a condition where the cell sites in one area overlap, which often causes handoffs and a poor network management.

Differential Evolution Algorithm (DEA) and Adaptive Mutation Genetic Algorithm (AMGA) have been widely used to solve combinatorial problems, such as those which are found in network optimizations. Mendes *et.al* analyzed the use of Differential Evolution Algorithm (DEA) for the active network mapping which resulted in the optimization level of 0.02 [3]. In other research, Lestandy optimized the mesh network routing by using the Adaptive Mutation Genetic Algorithm (AMGA) which obtained the MC/GA result of 2,3% [4]. Differential Evolution Algorithm (DEA) has unique characteristics in its mutation ratio and crossover ratio where the resulting probability will be adjusted to the inputs which contain the chromosome limitations [5]. Meanwhile, one of the benefits of the Adaptive Mutation Genetic Algorithm (AMGA) is it uses more complex mutation processes which are correlated with the desired fitness value [4].

The focus of this paper is to optimize the coverage of each cell of the 4G LTE network data in Mojokerto City, East Java, Indonesia, based on the e-Node B position coordinates. From the testing stage, optimization level of the network coverage in the service area will be obtained. The higher level of network coverage in the service area indicates the good performance of the methods used.

2. CELL SITES CONCEPTS

In cellular communications, information is interchanged between the Mobile Station (MS) and the Base Transceiver Station (BTS) via the radio signals. Each BTS can only communicate with the MSs within its coverage area. In other words, the radio signals deliveries are limited in a particular range of frequencies that several BTSs are needed to serve a wide area [6].

A BTS covers a certain area called a cell. The most common model of cells is some hexagonal with the same forms in the BTS's service area. Each cell provides some channels, that some of MSs can communicate with a BTS at the same time. A channel is usually defined according to time slots, the range of frequencies, encoding techniques, or a combination of TDMA, FDMA, and CDMA [7].

In cellular communication systems, as the amount of user traffic rises or the number of MSs increases, it is needed to add more channel's capacity. To add more channel's capacity, we can reduce the cell area's size (micro cell) or dynamically use the channels allocation and the re-use frequency. To plan the additional channel's capacity in a cellular system, we should consider the interference. There are two kinds of interference; those are the co-channel interference and the adjacent channel interference.

A group of adjacent cells which uses the whole frequency allocation is called a cluster size or a re-use frequency factor. According to the variation of their size and coverage, cells are categorized to femto, pico, micro, macro, and mega cells. Femto cells are usually used for connecting personal devices such as laptops. Pico cells usually cover a room or a part of a room in a building. The micro cells' coverage area is within

an urban area and the macro cells' is within a sub-urban area. Mega cells are usually used in satellite communications, as they cover an area of up to hundreds of kilometers.

3. RADIO PROPAGATION

Knowledges about radio propagation's characteristics are required to plan a cellular communication system design. Different from the regular communications, the environment profiles of cellular communication systems are hard to predict. Radio propagation is determined by the area's profiles, moving objects, radio frequency properties, the MS's speed, and interference sources.

The signal propagation mechanisms between a transmitter and a receiver vary depending on the area's profiles around the cellular communication environment. This causes the signals received by MS to fluctuate. Signal fluctuations can occur in three mechanisms: reflection, diffraction, and scatter.

4. METAHEURISTIC ALGORITHMS

The metaheuristic algorithms emerged as a new approach to solve the limitations of heuristic algorithms. Metaheuristic algorithms are inspired by events in nature, or better known as the nature inspired algorithms. This approach is simpler and easier to implement into the computer programming languages so it can provide faster solutions than the heuristic algorithms [8].

Some methods of the metaheuristic approach which can be used for solving various combinatorial optimization problems are Genetic Algorithm (GA), Cross Entropy (CE), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO).

5. DIFFERENTIAL EVOLUTION ALGORITHM (DEA)

Differential Evolution Algorithm (DEA) is included in the family of Evolutionary Algorithms (EA), which are evolutionary population-based algorithms. The algorithms' principles and philosophy simulate the biological evolution behaviors. The DEA was introduced by Storn and Price in 1995. DEA differs from the other type of Evolutionary Algorithms in the way to determine the distance and direction of the population's/solution's searching process.

According to Storn and Price, optimization techniques should in general meet 3 conditions [5]. First, the method must find a global optimum, despite the system's first parameters' values. Second, the convergence must be fast. Third, the program must have minimum limits of control parameters' values. Those conditions underlie the emergence of Differential Evolution Algorithms (DEA). DEA refines other types of Evolutionary Algorithms (EA) with a simple optimization strategy for a prompt optimization process (a shorter execution time with less literacy to find the global optimal solution).

6. ADAPTIVE MUTATION GENETIC ALGORITHM (DEA)

Genetic Algorithm (GA) is a method which is included in Evolutionary Algorithms (EA) family. In common, Evolutionary Algorithms (EA) imitates the natural process of evolution, where the main concept is the superior individuals will survive, while the inferior ones will become extinct [9]. The superiority of the individuals is measured through a mathematical function called the fitness function. The term of fitness in GA refers to the feasibility of a solution to the problem.

In GA, the mutation probability (Pm) value is constant, and this results in the lack of the optimization's efficiency. The Adaptive Mutation Genetic Algorithm (AMGA) uses adaptive mutations to prevent premature convergence [4].

7. RESEARCH METHODOLOGY

This research discusses the optimization of the 4G LTE's network coverage by using the metaheuristic approach. The testing stage was conducted through simulations in Python programming language by using the open-source Python notebook of Google Collaboratory.

The type of data used in this research was secondary data which were collected from books and research related to the topic. Secondary data used were the service area's geographical condition, cell's distribution of the cellular network, and the population density data. The functions of those data are:

1. The service area's geographical condition data were collected from the map of the service area. The data included special characteristics of the area, and those used to underlie the mappings of the cellular network.

2. Cells' distribution data included the position coordinates of communication provider's Base Transceiver Stations (BTSs) or e-Node Bs in the service area. From the preliminary data, we obtained the existing condition of the cellular communication's signal coverage. In this research, the cells' distribution would be repositioned according to the optimization results.

The cells' distribution data of Mojokerto City, East Java, Indonesia as well as the mutation and crossover ratio were used as inputs of the optimization models. The optimizations were conducted by using the Differential Evolution Algorithm (DEA) and the Adaptive Mutation Genetic Algorithm (AMGA). The inputs, processes, and outputs flow of this research is shown in the Figure. 1.

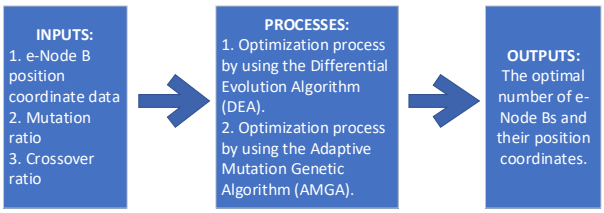

Figure. 1 Inputs, processes, and outputs flow diagram of the research

In general, the optimization processes in this research were started by collecting the input data, as follows:

1. The data of e-Node Bs' position coordinates in the service area. Those data contained the longitudes and latitudes of existing e-Node Bs which were the descriptions of the special data. Those position coordinates data were used in both Differential Evolution Algorithm (DEA) and Adaptive Mutation Genetic algorithm (AMGA) models.

2. Mutation ratio. This included the number of genes that should be used in the mutation steps in the Adaptive Mutation Genetic Algorithm (AMGA) optimization. On

implementation, Differential Evolution Algorithm (DEA) and Adaptive Mutation Genetic Algorithm (AMGA) have different mutation characteristics.

3. Crossover ratio. This was the ratio of offspring which are produced in crossover stage in every generation.

After defining the inputs, the optimization processes were conducted by using the Differential Evolution Algorithm (DEA) and Adaptive Mutation Genetic Algorithm (AMGA). From the optimization processes, we obtained the optimal numbers of e-Node B points which should be implemented in the service area as outputs. Those numbers described the effective levels of the coverage necessity within an area which were tabulated in a coordinate points table. From each algorithm which was implemented in this research, we obtained different outputs which would be compared and analyzed. Optimization parameters used in the simulations in this research for the DEA and AMGA are shown in the Table I and Table II, respectively.

Table 1. Optimization parameters of the Differential Evolution Algorithm (DEA)

Parameter	Parameter's value
Population's dimension	98
Population's size	10
Preliminary population's size	50
Maximum iteration	300
Mutation Ratio (F)	0.7
Crossover Ratio (CR)	0.6

Table 2. Optimization parameters of the Adaptive Mutation Genetic Algorithm (AMGA)

Parameter	Parameter's value
Population's dimension	98
Population's size	10
Preliminary population's size	50
Maximum iteration	300
Preliminary mutation ratios	0.1

8. RESULTS AND DISCUSSIONS

From the simulations, we obtained the number and the position coordinates of the e-Node Bs should be activated in the service area. The testing results of the Differential Evolution Algorithm (DEA) model is shown in the Figure. 2.

Figure. 2 The testing results of Differential Evolution (DE) method in Google Collaboratory

Fig 2. shows the number of e-Node Bs which are activated based on the optimization result. The Differential Evolution Algorithm (DEA) model generated 23 inactivated e-Node Bs and 17 activated e-Node Bs of 98 existing e-Node Bs. Moreover, we also get the activation status of each e-Node B based on the optimization results.

On the other hand, the Adaptive Mutation Genetic Algorithm (AMGA) model generated 16 inactivated e-Node Bs and 24 activated e-Node B of existing 98 e-Node Bs. We also obtained the position coordinates and the activation status of each e-Node B in the service area. The testing results of the Adaptive Mutation Genetic Algorithm (AMGA) model can be seen in Fig 3.

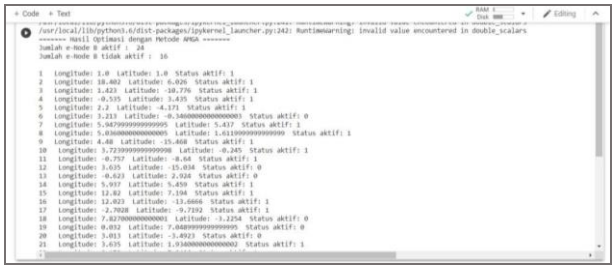

Figure. 3 The testing results of Differential Evolution (DE) method in Google Collaboratory

In existing condition, there are 98 e-Node B points to meet the coverage necessity of a 36.56 km2 area with 165.362 population. From the testing results we can see that the necessity of network coverage in the service area can be meet with less e-Node B points.

From the testing stage, the Differential Evolution Algorithm (DEA) model generated 23 inactivated e-Node B points, those are e-Node B number 2, 8, 20, 21, 34, 38, 39, 41, 47, 54, 59, 60, 65, 69, 70, 74, 80, 81, 83, 86, 87, 90, dan 94. This means there were overlapping coverage areas between some cell sites in the existing condition. Moreover, there were 17 activated e-Node B points to satisfy the necessities of network coverage in some areas that had poor coverage. Differential Evolution Algorithm (DEA) model resulted in 23.4% effectivity rate compared with the existing condition with evenly spread coverage.

On the other hand, the Adaptive Mutation Genetic Algorithm (AMGA) model resulted in 16 inactivated e-Node B points, those are e-Node B number 4, 8, 9, 21, 24, 34, 38, 39, 41, 47, 58, 67, 83, 90 dan 94. Then, there were 24 activated e-Node B points to meet the coverage necessities of some areas that had poor coverage. According to the testing results, the Adaptive Mutation Genetic Algorithm (AMGA) model got 16,8% effectivity rate compared with the existing condition. The comparison between the existing condition and the optimization result with the metaheuristic approach is shown in Fig 4 and Fig 5.

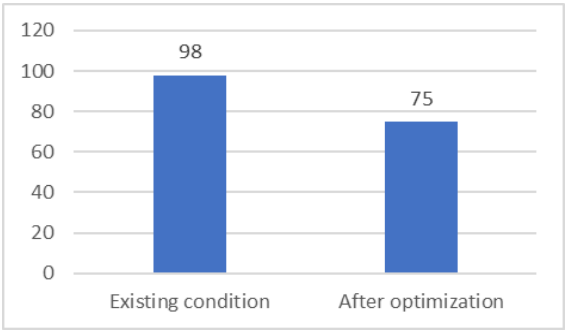

Figure. 4 Optimization's result of the Differential Evolution Algorithm (DEA)

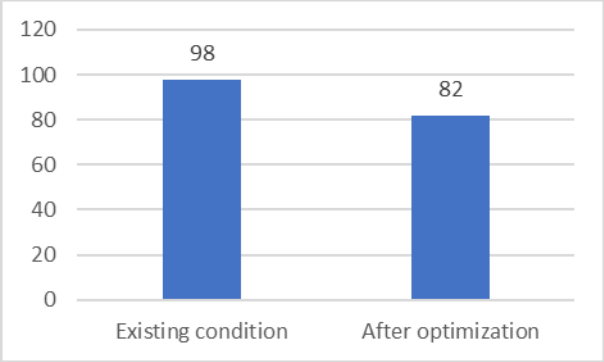

Figure. 5 Optimization's result of the Adaptive Mutation Genetic Algorithm (AMGA)

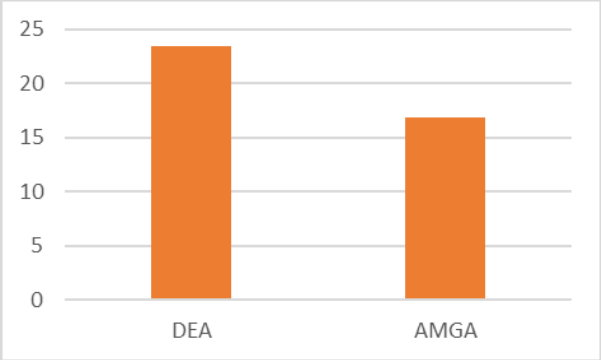

Figure. 6 Comparison between the optimization result of DEA and AMGA

According to the testing results, the effectivity rate of the Differential Evolution Algorithm (DEA) model is 23,4% dan Adaptive Mutation Genetic Algorithm (AMGA)'s is 16,8%. We can conclude that both metaheuristic methods have better results than the existing condition.

9. CONCLUSION

This research focuses on optimizing the 4G LTE's network coverage by using metaheuristic approach. The Differential Evolution Algorithm (DEA) and the Adaptive Mutation Genetic Algorithm (AMGA) have been proposed to determine the optimal coverage of the 4G LTE network. The Differential Evolution Algorithm (DEA) and the Adaptive Mutation Genetic Algorithm (AMGA) could be implemented to solve the network coverage optimization problem, especially for reducing overlapping coverage areas. The Differential Evolution Algorithm (DEA) model resulted in 23,4% of coverage effectivity rate compared with the existing condition. The Adaptive Mutation Genetic Algorithm (AMGA) model resulted in 16,32% of coverage effectivity rate. The future research can be conducted by implementing non-heuristic methods to solve the network coverage optimization problem of 4G LTE network. Moreover, we can add the other QoS parameter, such as traffic capacity.

10. REFERENCES

[1] Joonas Säe, Jukka Lempiäinen, "Maintaining Mobile Network Coverage Availability in Disturbance Scenarios", Mobile Information Systems, vol. 2016, Article ID 4816325, 10 pages, 2016. https://doi.org/10.1155/2016/4816325

[2] J. Wiley, Small Cell Optimization, Hoboken, NJ, USA: Wiley Library, 2015.

[3] S. P. Mendes, J. A. Gomez Pulido, M. A. Vega Rodriguez, M. D. Jaraiz Simon and J. M. Sanchez Perez, "A Differential Evolution Based Algorithm to Optimize the Radio Network Design Problem," 2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), Amsterdam, The Netherlands, 2006, pp. 119-119, doi: 10.1109/E-SCIENCE.2006.261052.

[4] Lestandy, M., Pramono S.H,, Aswin M., "Routing Optimization on Metropolitan Mesh Network Using Adaptive Mutation Genetic Algorithm" (Indonesian Version), November 2017, Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) 6(4), DOI: 10.22146/jnteti.v6i4.355

[5] D. Chaudhary, A. K. Tailor, V. P. Sharma and S. Chaturvedi, "HyGADE: Hybrid of Genetic Algorithm and Differential Evolution Algorithm," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-4, doi: 10.1109/ICCCNT45670.2019.8944822.

[6] B. Zhuang, D. Guo and M. L. Honig, "Energy-Efficient Cell Activation, User Association, and Spectrum Allocation in Heterogeneous Networks," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 823-831, April 2016, doi: 10.1109/JSAC.2016.2544478.

[7] Allen W. Scott; Rex Frobenius, "Multiple Access Techniques: FDMA, TDMA, AND CDMA," in RF Measurements for Cellular Phones and Wireless Data Systems , IEEE, 2008, pp.413-429, doi: 10.1002/9780470378014.ch30.

[8] X.-S. Yang, "Nature-Inspired Metaheuristic Algorithms: Success and New Challenges," J Comput. Eng. Inf. Technol., vol. 1, no. 1, pp. 1-3, 2012.

[9] A. P. Engelbrecht, Fundamentals of Computational Swarm Intelligence, Hoboken, NJ, USA: John Wiley & Sons, 2006.

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 14-21, 2021, ISSN:-2319–8656

Analyzing Brain Activities in Response To Music and Video Stimulants

Mohamed Fakhredine
PHD Student in VFU
Bulgaria

Zahra Jaish
PHD Student in Lebanese University, Lebanon

Houssein Ibrahim
PHD Student in VFU
Bulgaria

Abstract: Studying the effects of different stimulants on the brain is an ongoing research aiming to discover the activities and reactions of the brain. In this sake, we propose a study to show the impact of external stimuli on the human brain. The implementation includes 50 subjects exposed to four various stimuli while recording EEG. First, a control experiment is done where EEG is recorded in the absence of any stimulation. Then, EEG is recorded while the subjects are watching two consecutive sections from nature and violence videos, listening to sections of classical music and then heavy music.

The recorded signals are analysed through EEGLAB. The power spectral densities are computed and an analysed. The results showed a decrease in alpha absolute power was observed in 68.1% subjects when watching the nature movie, and in 72.34% subjects when watching the violence video. An increase in alpha and beta absolute power was detected in 70.45% of subjects when listening to the heavy music. Relative power spectral density showed a significant increase of delta rhythm upon watching the video of natural scenes, and a significant increase of alpha rhythm when listening to classical music. ANOVA test is also used to verify if any effect of the stimuli can be observed on each frequency bands. The results showed that the delta rhythm's highest values were obtained in reaction to the natural scenes video and that there is a significant relationship between alpha rhythms and the stimulus due to Classical music.

Keywords: *Data Analysis, Machine learning, ANOVA, EEG: Electroencephalography, δ: Delta wave, θ: Theta wave, α: Alpha wave, β: beta wave, γ: Gamma wave*

1. INTRODUCTION

For several medical purposes, there is a crucial need to study the effects of external stimulants, mainly auditory and visual, on the brain of human beings. These studies will help in detecting changes in the brain activity due to these stimulants. This in turn will lead to more and more discoveries of the behavior and functioning of the brain as well as different disorders of the brain.

Being the most interesting and complicated organ in the human body, discovering brain behavior and activity represents a major concern for the worldwide medical society. Therefore, studying brain activity is a necessary need for which many research activities are being performed in perspective of finding methods for measuring and representing brain activities.

An effective widely used method for studying the electrical activity of the brain is the "Electroencephalography" (EEG). EEG measures and displays the electrical activity of the brain. It has the ability to detect wide variety of mental activities and brain disorders and diseases including autism, epilepsy, and dementia [1]. EEG is a non-invasive, painless, and safe technique.

Many research works were carried out attempting to understand different brain disorders by using the EEG. Moreover, several studies have been designed and implemented to test the effect of various external stimulations on the brain via EEG.

One of the most commonly used stimuli in previous research studies is music. Music is a universal mode of auditory communication. In the beginnings of modern science, music has been studied as sounds and vibrations. The scientific research seeking an understanding of the cognitive aspects of music has grown almost recently. Music engages much of the brain processes; it evokes emotions whether positive or negative, connects to memory, and may trigger physical activities like dancing.

In this paper, we tend to examine the effect of music (auditory) as well as audio-visual (video) stimuli on the brain using EEG. A new approach using music stimulation is proposed. The main objective of this paper is to detect any type of change in the brain activity due to these stimulants.

On another hand, till now, there is still poor scientific understating of the biological role of music and its connection to certain brain disorders. New insights concerning the connection between music and brain disorders have been provided by clinical neuroscience [2]. However, in this paper we go deeper to prove on a scientific level the biological role of music in modulating brain processes. Moreover, the paper proposes a new approach in selecting the music and video stimuli. The music and video criteria are chosen to be as much opposite as possible in order to be able to detect variety of changes within the brain. Thus, music was taken from two distinct genres, and the videos were taken from two diverging styles.

2. LITERATURE REVIEW

Studying the brain's function and process has been aa long research journey. In particular, experimenting the response of the brain to certain external stimuli, via EEG or other acquisition technologies, is an essential part of this research. In this section, we present the most relevant articles researching the effect of music on the brain, the special effect of Mozart's music, and the effect of visual and combined audio-visual stimuli.

2.1 Music Effect on EEG

Using the EEG technology, Geethanjali et al. (2012) studied the impact of three different types of music, Carnatic, Hard Rock and Jazz, on the brain function with and without performing mental task activity. The study finds that during mental task performance, listening to Carnatic music showed high a significant increase in beta activity compared to listening to Jazz and Hard Rock. However, when listening to Jazz music without any mental task, the power spectrum of theta was significantly high compared to Carnatic and Hard Rock [4].

Anilesh et al. (2013) examined the impact of music on the central nervous system by comparing the EEG of various subjects while listening to instrumental music to their EEG recorded in normal conditions. Both the results of quantification of EEG signals and the obtained topological maps has led to similar conclusion, that music affected mainly the frontal lobe of the brain. The central lobe showed some changes but not as significant as the frontal one [5].

In the same context, Hurless et al. (2013) studied the effect of music genre and tempo on brain waves, specifically alpha and beta. EEG was recorded for 10 non-musician participants while listening to songs belonging to two music genres: rock and jazz, in which the tempo of each song was varied artificially three times: slow (100 beats per minute BPM), normal (120 BPM), and fast tempo (140 BPM). Music preference was also taken into consideration. Hurless et al. demonstrated that alpha waves amplitude increased upon listening to preferred music, yet no impact was seen on alpha when the tempo varied. On the other hand, as tempo increased beta waves amplitude showed a huge increment, whereas no change was observed on beta when changing genres [6].

2.2 Mozart's Effect

Mozart's music has gained a lot of attention since the first article published in Nature in 1993 studying the effect of listening to Mozart on human's spatial reasoning. In their experiment done to study and compare the effect of classical and heavy metal music on the brain and cardiovascular system, Kalinowska et al. (2013) recorded EEG signals for 33 participants before, during, and after listening to Mozart's sonata K.448 (classical) and Iron Maiden (heavy metal). They've seen that no noteworthy contrasts of power spectra between measurements of before, during, and after listening to both types of music. Only the amplitude of alpha rhythm showed a significant decrease after listening to the Mozart's music [7].

Yang et al. (2014) tried to prove Mozart effect scientifically by designing an experiment where the EEG is measured before, during, and after listening to Mozart's K.448 for 29 college students. The study indicates a decrease in alpha, theta, and beta power spectra in healthy adults due to listening to Mozart [8].

In 2015, Verrusio et al. (2015) studied Mozart's effect on the brain activity of healthy young adults, healthy elderly, and elderly with Mild Cognitive Impairment (MCI) through quantitative EEG. After listening to Mozart K448, an increase in alpha band and median frequency index of background alpha rhythm activity was observed, whereas no such change was detected after listening to Beethoven's "Fur Elise". The study argues that Mozart's music is capable of activating neuronal cortical circuits related to attentive and cognitive functions in young subjects, as well as in the healthy elderly [9].

2.3 Visual and Audio-Visual Stimulants Effect on EEG

In 2012, Ahirwal and Londhe (2012) implemented a study to test the effect of visual attention on the brain's electrical activity through power spectrum analysis of EEG signals. The power spectrum of the signals while focusing attention on a visual stimulus was 10-12 dB, with frequency range of 18-22 Hz, which lies in the beta frequency band. The study concludes that in the beta frequency band corresponds to cognitive brain process or visual attention [10]. Along with investigations of the brain's reaction in response to auditory or visual stimulations separately, the effect of simultaneous combination of these two types of stimulations has been also studied by few researchers and for various objectives.

Christos et al. (2010) examined the impact of audio-visual stimulation on alpha brain oscillations using EEG technology. Subjects were exposed to binaural auditory beats with different frequencies combined with flickering lights of 4 different colors (RGBY), and their effect on upper and lower alpha bands has been analyzed. The study showed that this combination can significantly synchronize alpha 1 and alpha 2 bands [11].

In a very different setup experiment, Lee et al. (2016) studied the EEG signals of subjects watching disgust-eliciting videos of disgusting creatures and body mutilation but with varying the auditory music. The videos were played two times; first with their original soundtrack, and then with relaxing music or exciting music and the original soundtrack muted. Extracting the relative power spectra from the EEG data showed that alpha, theta, and delta frequency bands were lower with disgust-eliciting videos with external music stimuli than with the original soundtrack. This study argues that participants experienced less disgust when watching disgust eliciting videos while listening to music rather than the original soundtrack [12].

After summarizing some of the previous researches done studying the effect of different stimulations on the human brain using EEG technology, the methods we have adopted in our research project for EEG signal treatment and time frequency analysis, are presented next.

3. STUDY DESIGN AND METHODOLOGY

3.1 Database

EEG signals are recorded for 56 healthy students using KT88 EEG Machine. The recording is done while applying different auditory and visual stimulations. The 21 males and 35 females' participants are aged between 18 years and 26 years (mean age 20.82 years). Subjects suffering from any neurological disorders or complications, or taking central nervous system depressants are excluded from undergoing the experiment. After obtaining the EEG recordings of all participants, 50 recordings among them are considered for further analysis. The recordings of 6 subjects (2 males and 4 females) are discarded due to errors and huge artifacts. The percentages of males and females remaining in the study, and whether they're right or left handed are given as follows; 38.0% males, 62.0% females, 86.0% right-handed, 14.0% left-handed.

3.2 Pre-experimental Conditions

Subjects are asked to wash their hair the day before, and not to expose it to any kind of oils, gels, and chemical reagents. Subjects are forbidden to drink any source of caffeine (tea, or coffee) and alcohol 12 hours before the experiment. Subjects are also asked to get enough sleep on the day before.

Prior to starting the experiment, the participant is asked to read and sign a consent in order to make sure that he accepts the whole experimental process and he is voluntarily participating. To preserve his confidentiality, a number is assigned for each participant, instead of name, that is used on the EEG recording and all research documents.

3.3 Positioning

The experiment is done in a dark room. The participant is asked to sit down on an armchair and to extend his feet on a facing chair, so that he sits while his body is lying and relaxed. This helps us avoid any unnecessary movement.

3.4 Electrodes Attachment

After ensuring the right positioning, 19 reusable cup gold-plated electrodes) are attached to the subject's scalp using a gel named "Ten20". Electrodes are positioned according to the standard methodology named "The International 10-20 System" that is recommended by the International Federation of Societies of Electroencephalography and Clinical Neurophysiology [1]. According to the 10-20 International System, the electrodes cover all cortical lobes including Frontal (Fp1, F3, F7, Fp2, F4, F8), Temporal (T3, T5, T4, and T6), Parietal (P3 and P4), Central (C3 and C4), and Occipital (O1 and O2). Two electrodes are attached to left and right earlobes, named A1 and A2 respectively. An additional electrode is added on the forehead of the subject to filter the signal. After placing electrodes, the participant wears the earphones that are connected to a portable hp-PC where music and videos are played. In and two students appear in the dark room, well positioned, and with the electrodes attached to their scalp.

3.5 Montage

The referential/monopolar montage is used, in which the 16 channels represent potential difference between one active and one inactive reference electrode. The inactive electrodes are the one's attached to earlobes A1 and A2. The left hemisphere active electrodes are defined by their odd index (Fp1, F3, F7, T3, T5, O1, P3 and C3) and their reference electrode is A1. The right hemisphere active electrodes are defined by their even index (Fp2, F4, F4, T4, T6, O2, P4, and C4) and their reference electrode is A2. In, the board of KT88 EEG machine appears, where the names of electrodes and their position are drawn on its surface.

3.6 Experiment Procedure

For each subject, five separate EEG signals are recorded; each for 2 minutes. The total time is 10 minutes for each subject.

The 1st EEG recording is done while the subject is closing his eyes and not exposed to any auditory or visual stimulation. After that, stimulations are applied according to the following steps:

Videos part: The 2nd and 3rd EEG recordings are done while the participant watches two consecutive 2-minutes videos on a PC screen in front of him. The first video is of scenes of nature accompanied with Piano Instrumental music (Spring Flowers Inspiration, YouTube). The second video is of violence scenes mainly killing and gun shooting (from "Headshot" movie, 2016). Thirty seconds break was taken between them.

Music part: The 4th and 5th EEG signals are recorded simultaneously while the participant listens to two consecutive 2-minutes music sections with a 30 seconds break between them. First, a section of Dubstep, Electronic Dance Music ("Bangarang" by Skrillex) and then a section of Classical Music (Mozart's Sonata Nb. 16). The participant is asked to close his eyes during listening to avoid eye blinking artifacts. A resting time, two minutes, is taken between videos and music part, where the participant is allowed to distract himself by any other activity.

Table 3-1 represents the stimuli applied while recording EEGs and summarizes different stimuli applied during the 5-EEG recordings and the URL of the music and videos used.

4. SIGNAL PROCESSING PHASE

Multiplying the number of subjects with the number of recorded signals for each subject, i.e. 50*5, gives us a total of 250 datasets. These EEG datasets are saved on the EEG software.

4.1 Pre-processing

Pre-processing of each signal was done before exporting them from the software. The signals are filtered with a band-pass filter, which is included within the software, of frequency range 1-40 Hz. This allows the exclusion of the noise of very low and very high frequencies. After filtering, the signals are exported in BDF format, and then introduced, one by one, into EEGLAB for the sake of analysis and interpretation.

4.2 Artifact Rejection

Artifacts in EEG, due to eye movements, blinks, muscle activity, etc. can greatly mislead EEG results and thus must be rejected. Standardly, visual inspection is used which allows users, mainly professional experts, to differentiate between artifact and non-artifact components. However, some artifacts features can be ambiguous to notice leading to difference in making decisions between users. Moreover, visual inspection requires a lot of experience and consumes time.

To avoid its complications, researchers have developed several algorithms capable of detecting artefactual ICs automatically, such as IC-Label. In these procedures, objective statistical measures from ICs are computed and used therefore to decide automatically whether this component is artefactual or not [13]. "IC Label" is a novel EEGlab-plugin which allows Automatic Artifact Rejection: It classifies ICs into brain and non-brain components determining specifically the contents of each component.

Using EEGLAB, Independent Component Analysis (ICA) decomposition is applied on each EEG dataset, giving rise to independent components of the same number of channels. IC-Label is then applied on these datasets to detect which ICs are mainly composed of artifacts.

IC-Label classifies artefactual components into 6 categories: Eye, Muscle, Heart, Line Noise, Channel Noise, and Other. And the components containing neural information from the brain are called Brain components. After visual inspection of the classification given by IC-Label, the index of each artefactual component is inserted manually in order to be rejected, i.e. excluded from the dataset. The number of rejected components may differ from one dataset to another depending on how much artifacts appear in each one.

4.3 Power Spectral Density

After removing the components of non-neural origin, Fast Fourier transform is applied to the dataset. Then the power spectral density PSD of each waveform, delta, theta, alpha and 40 beta, is calculated for each component using Welch method. That is, for each component in a dataset four numbers, PSDs, are obtained.

Since we have 250 datasets, and each dataset has up to 16 components (depending on how much components are removed after artifact rejection), and in order to avoid the complications of high number of obtained PSD values, the average of PSD of each dataset is calculated. In other words, for each dataset the mean PSD of delta, theta, alpha, and beta is computed, i.e. 4 PSD values are obtained for each dataset.

Then, and as stated in [12], the relative power spectrum was computed. Then ANOVA test was applied to obtain if there a relationship between the stimuli and the brain rhythms. SPSS was used in order to perform the required ANOVA test.

5. RESULTS

In this section the results of our experiment will be presented and discussed.

A brief glance on the effect of filtering and artifact rejection is illustrated first.

5.1 Filtering EEG Signals

The EEG signals are first filtered in order to remove all very low and very high frequencies.

5.2 ICA and IC-Labelling

Rejecting artifacts by running the ICA decomposition, and then applying IC-Label has cleaned the EEG signals from data of non-neural origin. The EEG signal is first transported to EEGLAB where ICS is applied first and independent components appear instead of channels. Then, running the IC-Label in the EEGLAB allows the classification of the artefactual components. In this dataset, 3 components are contaminated: IC2 (eye component), IC8 (eye component), and IC16 (other). Via EEGLAB, the contaminated components are removed, and therefore 13 ICs are only kept.

After being sure that each EEG dataset is clean of artifacts, absolute PSD of delta, theta, alpha, and beta rhythms is calculated for each component in the dataset.

5.3 Results from the Power Spectral Density Calculation

Recall that the power spectral density, or power spectrum, of a signal illustrates the power existing in a signal as function of frequency. The average PSD over the whole dataset is computed. The following are the results of absolute PSD variation for each stimulation ANOVA test is then performed for relative PSD values.

5.3.1 Absolute PSD variation after each stimulation using SPSS techniques

Consider first this notation: A denotes 1st recording (no stimulus), B denotes the 2nd (nature video), C denotes the 3rd (violence video), D denotes the 4th (heavy music), and E denotes the 5th (classical music). For each subject, the value of PSD, for all bands, was observed and compared between (A and B), (A and C), (A and D), and (A and E), not to detect the difference between values but to observe whether it increased or decreased. After that, the number of subjects in which PSD increased or decreased is determined. Finally, the percentage of subjects is computed.

Figure 5-1, Figure 5-2, Figure 5-3, and Figure 5-4 present the percentage of subjects in which an increase or a decrease was observed in their delta, theta, alpha, and beta PSD values between experiment “A” (no stimulus) and the other cases of different stimuli (B, C, D, E).

Table 5-1: The stimuli applied while recording EEGs

EEG Recording	Type of Stimulus	Name of Stimulus	URL
1st		*No Stimulus*	
2nd	Nature Video	Spring Flowers Inspiration	https://www.youtube.com/watch?v=FTBnAdrqZU4
3rd	Violence Video	Action scene from "Headshot" movie, 2016	https://www.youtube.com/watch?v=GlbojT7u7HA
4th	Dubstep Electronic Music	"Bangarang" by Skrillex	https://soundcloud.com/skrillex/skrillex-bangarang-feat-sirah
5th	Classical Music	"Sonata Nb. 16" by Mozart	https://soundcloud.com/moozar/mozart-the-piano-sonata-no-16

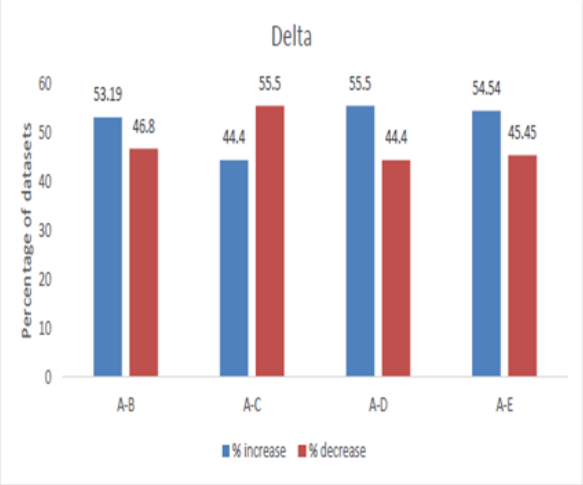

Figure 5-1: Variation of Delta PSD in all datasets

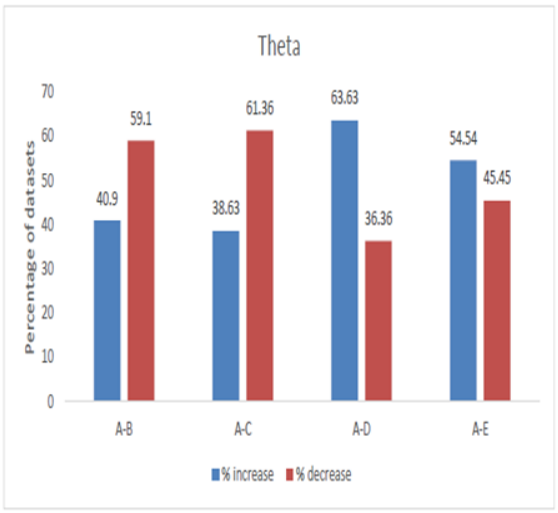

Figure 5-2: Variation of Theta PSD in all datasets

Figure 5-3: Variation of Alpha PSD in all datasets

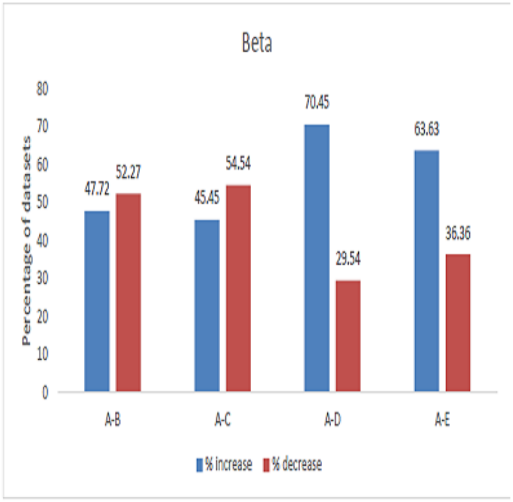

Figure 5-4: Variation of Beta PSD in all datasets

The results presented in Figure 5-1 and Figure 5-2 showed no significant difference between the percentages of datasets in which an increase or decrease was obtained in their PSD value of both Delta and Theta. Whereas in the case of Alpha and Beta, in some cases the percentage of people with an increase (or decrease) in PSD value due the application of stimuli is higher than that where a decrease (or increase) is obtained.

For alpha frequency band (Figure 5-3) in the case of A-B, the percentage of subjects with decreasing PSD is high **(68.1%)** compared to that of increasing PSD **(31.9%)**. Also, in the case of A-C, the percentage of subjects with decreasing PSD is high **(72.34%)** compared to that of increasing PSD **(27.65%)**. Moreover, in the case of A-D the percentage of subjects with increasing PSD is high **(70.45%)** compared to that of increasing PSD **(29.54%)**.

For beta frequency band (Figure 5-4), in the case of A-D the percentage of subjects with increasing PSD is high **(70.45%)** compared to that of increasing PSD **(29.54%)**.

5.4 ANOVA Test of Relative PSD

ANOVA test is used in order to obtain if any effect of the stimuli can be observed on each frequency bands. Table 5-2, Table 5-3, Table 5-4, and Table 5-5 show the results of the ANOVA test of relative PSDs of each of the brain rhythm: delta, theta, alpha, and beta. The same notation of the previous section is followed here for A, B, C, D, and E.

Table 5-2: ANNOVA test between Delta rhythm and the stimuli

Wave	Stimulus	Mean	Std. Deviation	Minimum	Maximum	p-value	Result
Delta (1-4 Hz)	A	0.678	0.144	0.250	0.886	0.049	Significant
	B	**0.709**	0.156	0.250	0.935		
	C	0.699	0.140	0.250	0.939		
	D	0.654	0.163	0.250	0.937		
	E	0.626	0.169	0.185	0.941		
	Total	0.673	0.157	0.185	0.941		

Table 5-3: ANNOVA test between Theta rhythm and the stimuli

Wave	Stimulus	Mean	Std. Deviation	Minimum	Maximum	p-value	Results
Theta (4-8 Hz)	A	0.166	0.054	0.046	0.289	0.864	Not Significant
	B	0.176	0.092	0.033	0.568		
	C	0.177	0.063	0.038	0.353		
	D	0.175	0.060	0.024	0.335		
	E	0.181	0.056	0.036	0.299		
	Total	0.175	0.066	0.024	0.568		

Table 5-4: ANNOVA test between Alpha rhythm and the stimuli

Wave	Stimulus	Mean	Std. Deviation	Minimum	Maximum	p-value	Results
Alpha (8-13 Hz)	A	0.129	0.089	0.046	0.509	0.000	Significant
	B	0.085	0.057	0.024	0.250		
	C	0.092	0.063	0.017	0.278		
	D	0.142	0.115	0.031	0.563		
	E	**0.162**	0.122	0.018	0.522		
	Total	0.122	0.097	0.017	0.563		

Table 5-5: ANNOVA test between Beta rhythm and the stimuli

Wave	Stimulus	Mean	Std. Deviation	Minimum	Maximum	p-value	Results
Beta (13-30 Hz)	A	0.028	0.047	0.006	0.250	0.987	Not Significant
	B	0.030	0.048	0.007	0.250		
	C	0.032	0.048	0.006	0.250		
	D	0.029	0.047	0.006	0.250		
	E	0.032	0.052	0.004	0.308		
	Total	0.030	0.048	0.004	0.308		

In order to obtain a significant relationship between the variables, p-value must be less than 5% (0.05). For Delta rhythm, Table 5-2 shows that p-value equals 0.049 which is less than 0.05 (5%) indicating that there is a significant relationship between them. The mean of stimulus B is the highest (**Mean = 0.709**) amongst all other stimuli.

Moreover, for Alpha rhythm, Table 5-4 shows that p-value is 0.00 which is also less than 0.05 (5%). There is also a significant relationship between alpha rhythm and the stimuli, and the highest mean is for stimulus E (**Mean = 0.162**).

On the other hand, for Theta brain rhythm, Table 5-3 shows a p-value of **0.864**, which not less than 5% (0.05). Consequently, the results are not significant, and none of these variations can be generalized. The same goes for Beta, where the p-value in is **0.987** (greater than 0.05) as shown in Table 5-5.

6. DISCUSSION AND LIMITATION

6.1 Discussion

Referring to Figure 5-4, the increase of PSD of beta rhythms in the case D, i.e. while listening to heavy music (Electronic Dubstep music, “Bangarang” by Skrillex), can be explained by the fact that beta is known as expression of alertness. The high percentage of subjects who showed an increase in PSD of beta when listening to the electronic

dubstep music is logical. Heavy music may have induced alertness and attentiveness of the subject.

For alpha frequency band (Figure 5-3), the percentage of subjects with decreasing PSD is high compared to that of increasing PSD in both cases of A-B and A-C. Alpha appears in the state of relaxed awareness with no attention or concentration, and appears more during eye closure. That is why a decrease in alpha is observed when comparing A-B and A-C, since in both cases of B and C, the subject opens his eyes to watch the videos (nature and violence) and therefore his attention increased compared to the case A where he is closing his eyes with no stimulus around.

In the case of A-D, the percentage of subjects with increasing alpha PSD is higher than that of decreasing PSD. This indicates that an increase of alpha PSD occurs at the case of Electronic Dubstep music. In fact, this results are counter common sense, since the heavy electronic music because it induces some alertness. In [6], Hurless et al. demonstrated that alpha waves amplitude increased upon listening to preferred music. This may suggest that alpha power spectrum increases also upon listening to a preferred type of music, and the electronic music may be a preferred genre for the subjects whose alpha showed an increase.

Referring to ANOVA test performed between relative PSDs of brain rhythms and the various stimuli, the results of Table 5-2 has shown that the highest mean was of stimulus B, the video of natural scenes accompanied with Piano Instrumental music. This implies that delta rhythm's highest values were in the reaction to the natural scenes video. This result is logical, since delta appears in the sleeping on drowsiness state. This may indicate that watching this video has relaxed the subjects and made them a bit sleepy due to its relaxing scenes of nature and flowers and the calm music accompanied with it.

Additionally, the results of alpha rhythm, displayed in Table 5-4, has shown a significant relationship between alpha rhythm and the stimulus E, the Classical music (Mozart's Sonata Nb. 16). This result is also logical since alpha represents the state of relaxed awareness. This means that the classical music may have increased the relative power spectral density of alpha rhythm due to its relaxing effect.

6.2 LIMITATIONS

In this study, certain limitations have stood in the face of obtaining more specific results. Technically, the subject was not completely lying on his back, and although the positioning was good to reduce movement, it was not sufficient to prevent it. Thus, artifacts appeared frequently in the data.

Although the stimulations were as divergent as possible to elicit a change in the brain, but in fact the time of the stimulation, 2 minutes for each, was relatively short. Two minutes were not enough to trigger a detectable change in brain rhythms power spectra. Also, the time separating the exposure to each stimulus was not sufficient (30 s between the 2 videos, 2 minutes between video and music part, 30 s between the 2 music stimuli). In fact, extending the time of the experiment was difficult, since the volunteers were students who cannot tolerate length experiments.

7. CONCLUSION

The brain is the most complex and undiscovered organ in the human body. Therefore, studying the effects of various stimulants on the brain can help grasp some of knowledge about it. This work has studied the impact of listening to heavy and classical music on the brain activity using EEG technology. A control experiment was done by recording EEG with no stimulus. The absolute and relative power spectral density of the four brain waves was studied.

ANOVA test was done between relative power spectral density and the various stimuli. High delta power implies that watching this video has made the participants sleepy or triggered their feelings of drowsiness, and this agrees with the usual appearance of delta in the states of sleep or drowsiness.

ANOVA test also showed a significant increase of alpha relative power when listening to classical music ("Sonata Nb. 16" by Mozart). Since alpha appears frequently in the state of relaxed awareness and in less frequency during alertness or tense, high alpha power indicates that the classical music has relaxed the participants and decreased their tense.

Moreover, observing the variations of absolute power spectral density of the four brain rhythms in the EEG signals showed a high percentage of subjects whose alpha decreased from the case of no stimulus to both cases of watching nature and violence videos.

Additionally, a high percentage of subjects whose alpha and beta increased when listening to heavy music (Electronic Dubstep music, "Bangarang" by Skrillex) is detected.

The study aims to collect the different values of ANOVA signals after a various stimulus, in order to predict the patient situation in a normal case, that will help the medical teams to get direct notification about the real diagnostics or unexpected measures.

Several limitations have been detected in our study and may have affected our results, including the short time of applying the stimulations, the large number of artifacts in the EEG signals

8. FUTURE WORK

Further work can be done using EEG machine in order to know the impact of stimulations of the brain activity. Future work can have a more specific design for the experiment. It can consider studying only music with adding more than two genres. Preference may be taken into consideration too. It can also include healthy subjects in addition to subjects suffering from certain brain disorders, in order to study the effect of music on both healthy and non-healthy subjects. Additionally, using feature extraction tools like Wavelet Transform and Approximate Entropy is essential in a further research about EEG, due to their high accuracy in analyzing time varying signals

9. REFERENCES

[1] S. Sanei and J. A. Chambers, EEG Signal Processing, England: John Wiley & Sons Ltd, 2007.

[2] C. Clark, L. Downey and J. Warren, "Brain disorders and the biological role of music," 2015. [Online]. Available: www.ncbi.nlm.nih.gov.

[4] B. Geethanjali, K. Adalarasu and R. Rajsekaran, "Impact of music on brain function during mental task using Electroencephalography," World Academy of Science, Enginering and Technology, 2012.

[5] D. Anilesh, P. K. Sanjay, D. K. Bhattacharya, D. N. Tibarewala and D. Derbaj, "Study of the effect of music on central nervous system through long term analysis of EEG signals in time domain," International Journal of Engineering Sciences & Emerging Technologies, vol. 5, no. 1, pp. 59-67, April 2013.

[6] N. Hurless, A. Mekic, S. Pena, E. Humphries, H. Gentry and D. F. Nichols, "Music genre preference and tempo alter apha and beta waves in human non-musicians," The Premier Undergraduate Neuroscience Journal, 2013.

[7] A. Kalinowska, A. Kulakowska, W. Kulak and B. Okurowska-Zawada, "Effects of classical and heavy metal music on the cardiovascular system and brain activity in healthy students," 2013.

[8] L.-C. Lin, C.-S. Ouyang, C.-T. Chiang, R.-C. Wu, H.-C. Wu and R.-C. Yang, "Listening to Mozart K.448 decreases electroencephalography oscillatory power associated with an increase in a sympathetic tone in adults: a post-intervention study," The Royal Society of Medicine, 2014.

[9] W. Verrusio, E. Ettorre, E. Vicenzini, N. Vanacore, M. Cacciafesta and O. Mecarelli, "The Mozart Effect: A quantitative EEG study," ELSEVIER, Conciousness and Cognition, pp. 150-155, 2015.

[10] M. K. Ahirwal and N. D. londhe, "Power Spectrum Analysis of EEG Signals for Estimating Visual Attention," International Journal of Computer Applications, 2012.

[11] C. Moridis, M. Klados, I. Kokkinakis, V. Terzis, A. Economides, A. Karlovasitou, P. Bamidis and V. Karabatakis, "The Impact of Audio-Visual Stimulation on Alpha Brain," in IEEE International Conference on Information Technology and Applications in Biomedicine, 2010.

[12] M.-J. Lee, H.-L. Kim and H.-B. Kang, "The effects of visual and auditory stimulation on EEG power spectra during the viewing of disgust-eliciting videos," in Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods ICPRAM, Korea, 2016.

[13] M. Chaumon, D. V. Bishopc and N. A. Buscha, "A practical guide to the selection of independent components of the," Journal of Neuroscience Methods, ELSEVIER, 2015.

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 22-29, 2021, ISSN:-2319–8656

Implementation of Simple Additive Weighting (SAW) Method and Profile Matching for Employee Selection

Andi Pratomo Wiyono
Dept. of Electrical Engineering
Universitas Brawijaya
Malang, Indonesia

Muhammad Aziz Muslim
Dept. of Electrical Engineering
Universitas Brawijaya
Malang, Indonesia

Muhammad Aswin
Dept. of Electrical Engineering
Universitas Brawijaya
Malang, Indonesia

Abstract: Employees are an important element in a company that determines the progress of a company. With good quality employees in a company, it is easier to achieve desired goals of a company. Conventional (manual) recruitment method is vulnerable to non-technical factors such as frequent duplicate data or invalid data. In such condition, a Decision Support System (DSS) will be helpful in making decision process valid and reliable. In this paper, a Simple Addictive Weighting (SAW) method and Profile Matching were proposed to solve employee selection problem. This research was conducted at UPT Career Development and Entrepreneurship Universitas Brawijaya Malang, using data collected from written test selection in 2019. The effectiveness of both methods is analyzed by means of confusion matrix. SAW method give Accuracy rate of 94.7%, Precision rate of 87.5%, Recall rate of 91.3% and F-measure rate of 89.4%. On the other hand, Profile Matching method obtained the Accuracy rate of 90.4.7%, Precision rate of 81.4%, Recall rate of 81.4% and F-measure rate of 81.4%. From these results, it can be concluded that both methods have a high accuracy value accompanied by a high precision value when used for the selection process. This system can also reduce the bias of the same data very well, as can be seen from the high Recall and F-measure rates.

Keywords: decision support system, employee selection, simple additive weighting method, profile matching, confusion matrix.

1. INTRODUCTION

Employees are an important element in a company in determining the progress of a company. With good quality employees in a company, it makes easier for the company to achieve the goals of a company. Selection of effective applicants or employee candidates to assess technical abilities, education, work experience as well as psychological assessments of applicants, psychological tests will generally show a person's emotional state, in addition, a technical ability test will show a person's competence to work. However, someone with good technical skills, if not supported by sufficient emotional intelligence, will experience difficulties in his work environment [1].

At present, the method used in the employee selection process at the Career Center of Universitas Brawijaya Malang (UPKK) is still using conventional methods, by using human labor in the process of determining whether or not applicants will qualify. This method is vulnerable to non-technical factors such as frequent duplicate data or invalid data. To solve this problem, the right Decision Support System is needed in determining decision making. There are various kinds of decision support system methods, namely: AHP, WP, TOPSIS, Simple Addictive Weighting (SAW), Profile Matching, expert systems and simple linear regression. Of all the decision support system methods above, the method chosen in determining the decision to acquire new employees is Profile Matching and SAW [2].

The SAW method is a systematic method of decision making that is able to show assessing the competence of applicant according to the criteria set by the company or decision maker based on systematic data analysis [3] while the Profile Matching Method is a method that compares competencies owned by the candidate and the competency of the position. So that it can be seen that the difference in competence is also often referred to as a gap. The smaller the gap (difference) a candidate gets, the candidate has a greater final score and is very close to the required qualifications [4].

Based on the description of the above problems, regarding the needs of the UPKK regarding a decision support system to assist in the selection of recruitment for employees of a company in recruiting, comparing the results of the process using the Simple Addictive Weighting (SAW) method and the Profile Matching method is an interesting thing. The application of this method is in the employee candidate selection system in UPKK so that it can help to see the potential of prospective employees to occupy a certain position in a certain institution in the company.

2. LITERATURE REVIEW

In 2016, M. Isman conducted research using SAW to support employee selection decisions at PT Philips Seafood Indonesia. The results of this study indicate the highest value is 77.5 with a range of 0-100. Manual calculations and calculations using a decision support system are claimed to get the similar results so that the system has high validity [5]. In other studies, using a similar method, it is said that the results of the 30-applicant data used get the comparison between manual and system calculations that have an accuracy of 81% [6].

Several other studies that have been carried out using the Profile Matching method, namely supporting sorting decisions based on the type of voice of the new members of the BIOS choir division studied by Syah in 2017. The results show that the system performance he designed can be used to make member admission decisions with the output in the form order based on the highest to the lowest end with the number of test data as much as 61, has a validity percentage of 77.04%. In fact, other studies have shown an accuracy of 96.2% [7] [8].

Based on some of the studies that have been described, it can be seen that the use of the Simple Additive Weighting and Profile

Matching methods has satisfactory results in each method. This research will deal with the application and accuracy comparison of the Simple Additive Weighting and Profile Matching methods, with the case study of selection of prospective employees based on data held by UPKK Universitas Brawijaya.

2.1. Decision Support System

Michael S. Scott Morton (1970) first articulated the important concept of a Decision Support System (DSS). Michael S. Scott Morton defines DSS as an interactive computer-based system, which helps decision makers to use data and various models to solve unstructured problems.

The concept of DSS is characterized by a computer-based interactive system that helps decision making utilizing data and models to solve unstructured problems. Basically, the DSS is designed to support all stages of decision making starting from identifying problems, selecting relevant data, determining the approach used in the decision-making process, to evaluating alternative choices [9].

2.2. Simple Additive Weighting (SAW)

Simple Additive Weighting is a method that is often used for decision making because this method is more efficient and has a fairly high accuracy. This method uses the largest (selected) result as its output. In the Simple Additive Weighting method, there are 2 types of criteria, namely the criteria that are beneficial (benefit) and criteria that are detrimental (cost). The advantages of this method in the form of the ability to assess more accurately because it is based on the value of the criteria and weighting preferences are predetermined and can choose the best alternative from a number of alternatives, other than that due to the increase in the after determining the weight values for each attribute [10].

2.3. Profile Matching

Profile Matching is a method where this method first determines the competency value (ability) required for a position. The competence of these abilities must be met by the holder or the candidate whose performance will be assessed. Broadly speaking, Profile Matching is a comparison process between individual competencies and job competencies so that the difference in competence is known as a gap, and the smaller the gap resulting from the comparison process above, the greater the weight value. This means that they have a greater chance of becoming an employee candidate to occupy the position [11].

In other literature, it is stated that the Profile Matching method is a decision-making mechanism by assuming that there is an ideal predictor variable level that must be met or passed. In Profile Matching, identification of good or bad groups of employees or job applicants. The employees in the group are measured using several assessment criteria. In Profile Matching, the job applicants who are appointed are those who are closest to the ideal profile of a successful employee [4].

2.4. The application of DSS uses SAW and Profile Matching

The concept of DSS (as shown in Figure 1) is characterized by a computer-based interactive system that helps decision making utilizing data and models to solve existing problems.

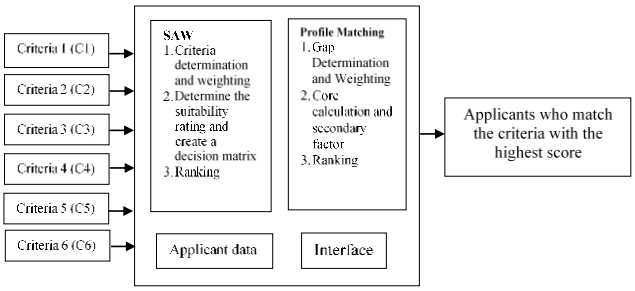

Figure 1. Application of SAW and Profile Matching in DSS

3. STUDY DESIGN AND METHODOLOGY

The method used in this research is data collection, design, implementation, testing and analysis as well as drawing conclusions and suggestions. Figure 2 shows the research methodology carried out in this research.

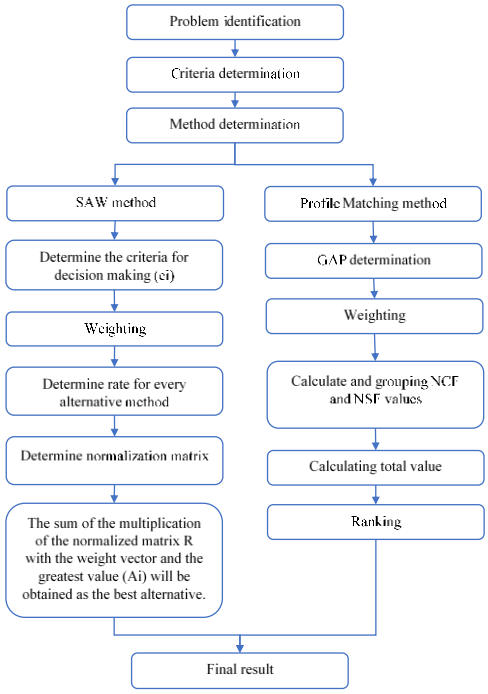

Figure 2. Research method.

3.1. Data Source

The data source used as research material is the primary data source. The data required is the result of selection of employee interviews with variables of educational suitability, GPA, which comes from the University, technical abilities, work experience, proficiency test results, biographical information, and data of all applicants who register at one particular company with the same position in accordance with company needs. In this study, the company under study was PT Kayaba Indonesia. This company opens vacancies for 4 positions, namely: Production Foreman & Warehouse, Foreman Production Planning & Control, Foreman PCE & Maintenance, and Supervisor Management System Information. Each vacancy has its own qualifications. In this study, the Foreman position requires qualifications in the form of

male gender, D3 Department of Mechanical / Electrical / Industrial Engineering, minimum GPA of 2.75, single, maximum age 24 years, while for supervisor positions requires qualifications such as S1 Informatics Engineering Department, minimum GPA 2.75, single, maximum age 26 years.

The data sources obtained are stored in the form of CSV (Comma Separated Values) files. This data will then be loaded through the application, and output in the form of a CSV file as well.

3.2. System Planning

The system to be used is a computer with hardware specifications an Intel Core i3 processor, with 4 GB of RAM. The software used is the Ubuntu 18.04 LTS operating system and the Python 3.6 programming language.

In the initial step of the SAW method, the data used is a CSV file that will be inputted through the application. The data will later take values from 5 to 11 criteria selected as a reference. After that, the ranking calculation will be carried out using the SAW method. After the calculation is complete, the system will issue the name and point of the result and can be saved as a CSV file. The system process flow image can be seen in Figure 3.

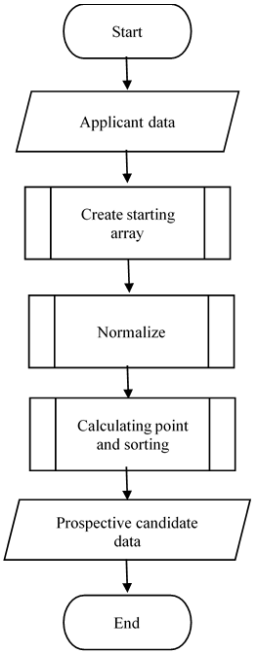

Figure 3. System design with SAW method.

In working on the Profile Matching method applied by the researcher, initially the CSV data is entered by the user then the system will run. When the system is started, a preprocessing process will first run to prepare the data so that it is ready to be processed. Then the calculation process will be carried out using the Profile Matching method to find out which candidate is closest to the predetermined criteria. An overview of the process flow is shown in Figure 4.

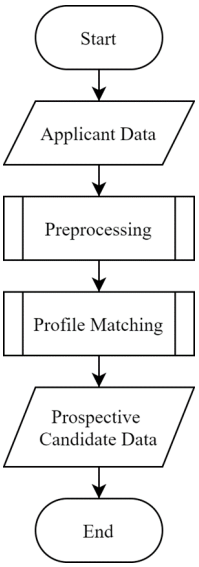

Figure 4. System design with Profile Matching method.

Each candidate will be given a score according to the conditions they have. The process of assigning candidate competency scores in the SAW method is depicted in Figure 5.

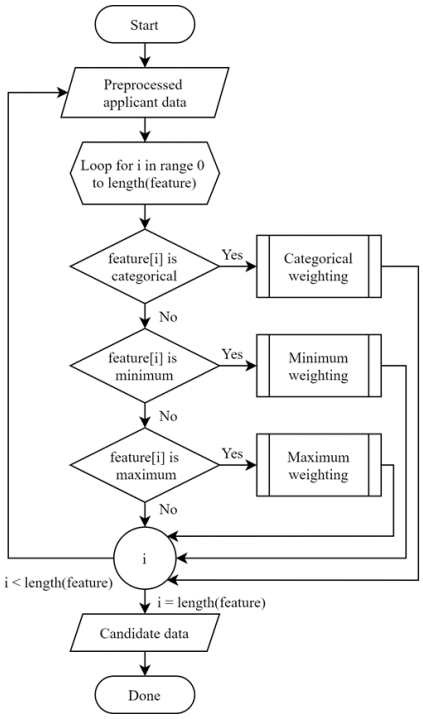

Figure 5. Flowchart of candidate scoring with SAW.

Scoring to each candidate will be divided into 3 ways, the first is to assign scores to features that are categorical. Categorical features include Study Program, Gender, Department, Faculty and Civil Status. Then the candidate value which has a value equal to the ideal profile value that has been determined by the user will get a value of 2, otherwise it will have a value of 1. Second is the value of features based on the minimum limit. Features that use a minimum score in determining the desired

conditions include GPA, TOEFL / TOEIC scores, Height, Year of Graduation, and Year of Entry. Candidates who have a value more than the same as the ideal profile value that has been determined by the user will get a score of 2, if not then it will have a value of 1. Then the third is a feature that uses the maximum value in determining the desired conditions, including weight and age. Candidates who have a value less than equal to the ideal profile value defined by the user will get a score of 2, otherwise it will have a value of 1.

The calculation process in Profile Matching that used in system development is shown in Figure 6.

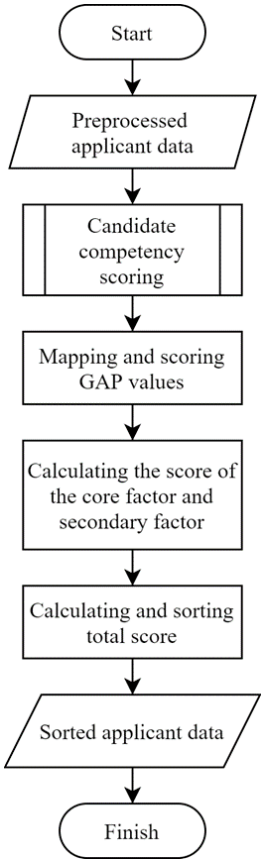

Figure 6. Flowchart of candidate scoring with Profile Matching.

These systems are implemented on computer using programming software with Python programming.

4. RESULTS AND DISCUSSION

4.1. Data Preparation

The data used are from PT XX's applicants (real name of the company is keep confidentially) with the criteria for the applicant's initial data in the form are full name, position, gender, civil status, place of birth, date of birth, age, height, weight, home address, cellphone number, email, type of English test (TOEIC/TOEFL), TOEIC/TOEFL test score, University, educational stage, year of university entry, year of university graduation, date of graduation trial, semester, GPA, study program, department, faculty, interest, completeness of transcript, completeness of Certificates/SKL, work experience (company, position and length of work), and organizational experience (organization, position, period). In minimizing misunderstandings in university input, normalization is carried out for university features. The number of applicants was 564 applicants with the desired positions, namely: Production Foreman & Warehouse, Foreman Production Planning & Control, Foreman PCE & Maintenance and Supervisor Management Information System. Foreman positions have specific criteria such as male gender, D3 Department of Mechanical / Electrical / Industrial Engineering, minimum GPA of 2.75, single, maximum age 24 years, while supervisor positions require qualifications such as S1 Department of Informatics, minimum GPA of 2.75, Single, maximum age 26 years

4.2. Results of SAW Method

The calculating steps to get the scores using the SAW method are as follows (Tables 1-9 are the corresponding results of each steps).

4.2.1. Determine the criteria (Ci) set by PT XX which will be used as a reference in making decisions.

Table 1. Criteria used in the Company

No	Criteria number	Criteria	Information	Weight
1	C1	Benefit	Educational stage	0.6
2	C2	Benefit	Department	0.5
3	C3	Benefit	GPA	0.4
4	C4	Benefit	Civil status	0.3
5	C5	Cost	Age	0.2
6	C6	Benefit	Gender	0.1

4.2.2. Provide the value of each alternative on each predetermined criterion. Each criterion must be assigned.

Table 2. Weighting of C1

No	User	Educational level	Score
1	User1	SMK	1
2	User2	D3	2
3	User3	S1	1
4	User4	S1	1
5	User5	D4	1

Table 3. Weighting of C2

No	User	Department	Category	Score
1	User1	Accountant	Not available	0
2	User2	Mechanical Engineering	Available	1
3	User3	Industrial Engineering	Available	1
4	User4	Informatics	Not available	0
5	User5	Electrical Engineering	Available	1

Table 4. Weighting of C3

No	User	GPA	Score
1	User1	3	3
2	User2	3,05	3,05
3	User3	3,32	3,32
4	User4	3,19	3,19
5	User5	3,29	3,29

Table 5. Weighting of C4

No	User	Civil status	Score
1	User1	Single	1
2	User2	Single	1
3	User3	Single	1
4	User4	Single	1
5	User5	Single	1

Table 6. Weighting of C5

No	User	Birth date	Score
1	User1	18/11/2001	18
2	User2	22/03/1998	21
3	User3	01/01/1994	26
4	User4	18/09/1996	23
5	User5	13/09/1996	23

Table 7. Weighting of C6

No	User	Gender	Score
1	User1	Woman	0
2	User2	Man	1
3	User3	Man	1
4	User4	Man	1
5	User5	Man	1

4.2.3. Determine the suitability rating of each alternative on each criterion

Table 8. Table of Ratings in each Criterion

No	User	Criterion					
		C1	C2	C3	C4	C5	C6
1	User1	0,5	0	0,90361	1	1	0
2	User2	1	1	0,91867	1	0,857143	1
3	User3	0,5	1	1	1	0,692308	1
4	User4	0,5	0	0,96084	1	0,782609	1
5	User5	0,5	1	0,99096	1	0,782609	1

4.2.4. Decision matrix based on criteria (Ci), then performed the matrix normalization

$$X = \begin{bmatrix} 0,5 & 0 & 0,903614458 & 1 & 1 & 0 \\ 1 & 1 & 0,918674699 & 1 & 0,857142857 & 1 \\ 0,5 & 1 & 1 & 1 & 0,692307692 & 1 \\ 0,5 & 0 & 0,960843373 & 1 & 0,782608696 & 1 \\ 0,5 & 1 & 0,990963855 & 1 & 0,782608696 & 1 \end{bmatrix}$$

4.2.5. Normalization in each criterion

Criteria of level, including benefit:

$R1.1 = \frac{1}{2} = 0.5$

$R1.2 = \frac{2}{2} = 1$

$R1.3 = \frac{1}{2} = = 0.5$

$R1.4 = \frac{1}{2} = = 0.5$

$R1.5 = \frac{1}{2} = = 0.5$

Criteria of major, including benefit:

$R2.1 = \frac{0}{1} = 0$

$R2.2 = \frac{1}{1} = 1$

$R2.3 = \frac{1}{1} = 1$

$R2.4 = \frac{0}{1} = 0$

$R2.5 = \frac{1}{1} = 1$

Criteria of GPA, including benefit

$R3.1 = \frac{3}{3.32} = 0.9036144$

$R3.2 = \frac{3.05}{3.32} = 0.9186746$

$R3.3 = \frac{3.32}{3.32} = 1$

$R3.4 = \frac{3.19}{3.32} = 0.9608433$

$R3.5 = \frac{3.29}{3.32} = 0.990963$

Criteria of status, including benefit

$R4.1 = \frac{1}{1} = 1$

$R4.2 = \frac{1}{1} = 1$

$R4.3 = \frac{1}{1} = 1$

$R4.4 = \frac{1}{1} = 1$

$R4.5 = \frac{1}{1} = 1$

Criteria of Age, including benefit:

$R5.1 = \frac{18}{18} = 1$

$R5.2 = \frac{18}{21} = 0,85714285$

$R5.3 = \frac{18}{26} = 0,69230769$

$R5.4 = \frac{18}{23} = 0,78260869$

$R5.5 = \frac{18}{23} = 0,7826087$

Criteria of sex, including benefit

$R6.1 = \frac{0}{1} = 0$

$R6.2 = \frac{1}{1} = 1$

$R6.3 = \frac{1}{1} = 1$

$R6.4 = \frac{1}{1} = 1$

$R6.5 = \frac{1}{1} = 1$

4.2.6. Final Result of SAW Method

Table 9. Final Result of SAW Method

No	Alternative	Criterion						Sum
		Benefit				Cost	Benefit	
		C1	C2	C3	C4	C5	C6	
1	User 2	0,6	0,5	0,36746988	0,3	0,171428571	0,1	2,038898451
2	User5	0,3	0,5	0,396385542	0,3	0,156521739	0,1	1,752907281
3	User3	0,3	0,5	0,4	0,3	0,138461538	0,1	1,738461538
4	User4	0,3	0	0,384337349	0,3	0,156521739	0,1	1,240859089
5	User1	0,3	0	0,361445783	0,3	0,2	0	1,161445783

From 564 initial data, researchers processed by using the SAW method and filtered 144 data of prospective employees. From the 144 data of prospective employees, the researchers got some recommendations based on the highest ranking (rank 1- 5) of these prospective employees, namely:

1. User 2 with total points 2.038898451
2. User 5 with total points 1.752907281
3. User 3 with total points 1.738461538
4. User 4 with total points 1.240859089
5. User 1 with total points 1.161445783

The data will be tested for the accuracy and specificity by using the Confusion Matrix method. The results of the Confusion Matrix test are given in Table 10.

Table 10. Confusion Matrix SAW Table

Initial Data		**564**
Result	SAW	144
Confusion Matrix	TP	126
	FP	18
	FN	12
	TN	408
Accuracy		94.7%
Precision		87.5%
Recall		91.3%
F - Measure		89.4%
Specificity		95,8%

From the table 10, there were 564 initial participants and screened into 154 participants by using the SAW method. It was shown that the SAW method had an accuracy of 94.7%, a precision of 87.5%, a recall of 91.3%, and an F-Measure of 89.4%.

The accuracy of 94.7% in the SAW method is greater than the research hypothesis which stated that the accuracy rate of SAW method was 80-90%. This indicates that the SAW method has a very good level of accuracy to be applied in the new employee candidate selection system. This high level of accuracy is also supported by a high number of high precision (87.5%). This shows that the SAW method is very specific to be used in selecting employee candidate recommendations according to predetermined criteria. This is supported by the high recall rate of 91% and a F-Measure value of 89.4%.

4.3. Result of Profile Matching Method

The calculating steps to get the scores using the SAW method are as follows (Tables 11-20 are the corresponding results of each steps).

4.3.1. GAP mapping

Table 11. C1 GAP calculation

Alternative	C1 (Age)		
	Employee profile	Position profile	GAP
User1	24	24	0
User2	26	24	-2
User3	24	24	0
User4	24	24	0
User5	23	24	1

Table 12. C2 GAP calculation

Alternative	C2 (Status)		
	Employee profile	Position profile	GAP
User1	Single	Single	0
User2	Single	Single	0
User3	Single	Single	0
User4	Single	Single	0
User5	Single	Single	0

Table 13. C3 GAP calculation

Alternative	C3 (Education)		
	Employee profile	Position profile	Employee profile
User1	SMK	User1	SMK
User2	D3	User2	D3
User3	S1	User3	S1
User4	S1	User4	S1
User5	D4	User5	D4

Table 14. C4 GAP calculation

Alternative	C4 (Major)		
	Employee profile	Position profile	GAP
User1	Business	Mechanical Eng.	1
User2	Mechanical Eng.	Mechanical Eng.	0
User3	Industrial Eng.	Mechanical Eng.	1
User4	Informatics	Mechanical Eng.	1
User5	Telecommunication Eng.	Mechanical Eng.	1

Table 15. C5 GAP calculation

Alternative	C5 (GPA)		
	Employee profile	Position profile	GAP
User1	2,82	3	0
User2	3,26	3,05	0
User3	3,33	3,32	0
User4	3,16	3,19	0
User5	3,46	3,29	0

Table 16. C6 GAP calculation

Alternative	C6 (Gender)		
	Employee profile	Position profile	GAP
User1	Woman	Man	1
User2	Man	Man	0
User3	Man	Man	0
User4	Man	Man	0
User5	Man	Man	0

4.3.2. Weighting

Table 17. Weighting results

Alternative	Weight					
	C1	C2	C3	C4	C5	C6
User1	2	2	1	1	2	1
User2	1	2	2	2	2	2
User3	2	2	1	1	2	2
User4	2	2	1	1	2	2
User5	2	2	1	1	2	2

4.3.3. Calculating and grouping of core and secondary factor

Table 17. Grouping of core and secondary factor

No	Category	Information	Factor
1	C1	Age	Secondary
2	C2	Graduation year	Core
3	C3	Education stage	Core
4	C4	Major	Core
5	C5	GPA	Core
6	C6	Gender	Core

Table 18. Core and secondary factor calculation

Alternative	Weight					
	C1	C2	C3	C4	C5	C6
User1	0,4	1,6	0,8	0,8	1,6	0,8
User2	0,2	1,6	1,6	1,6	1,6	1,6
User3	0,4	1,6	0,8	0,8	1,6	1,6
User4	0,4	1,6	0,8	0,8	1,6	1,6
User5	0,4	1,6	0,8	0,8	1,6	1,6
Core factor 80% secondary factor 20%						

4.3.4. Total calculation of score

Table 19. Total calculation score

User	NSF	NCF	NCI	Rank
User1	0,4	1,4	1,8	5
User2	0,2	2	2,2	1
User3	0,4	1,6	2	3
User4	0,4	1,6	2	4
User5	0,4	1,6	2	2

4.3.5. Final ranking result

Table 20. Total rank

User	NSF	NCF	NCI	Rank
User2	0,2	2	2,2	1
User5	0,4	1,6	2	2
User3	0,4	1,6	2	3
User4	0,4	1,6	2	4
User1	0,4	1,4	1,8	5

From the initial 564 data, after the Profile Matching Method was carried out, the filtered data was obtained for 140 prospective employees. From the 140 data on prospective employees, recommendations for prospective employees are obtained based on the ranking of the highest. Prospective employees include:

1. User 2 with total points 2.8
2. User 5 with total points 2.7
3. User 3 with total points 2.7
4. User 4 with total points 2.7
5. User 1 with total points 2

The data will be tested for accuracy and specificity using the Confusion Matrix method. The results of the Confusion Matrix test are given in Table 21.

Table 21. Confusion Matrix of Profile Matching

Initial Data		564
Result	PM	140
Confusion Matrix	TP	114
	FP	26
	FN	26
	TN	377
Accuracy		90,4 %
Precision		81,4 %
Recall		81,4 %
F - Measure		81,4 %
Specificity		93,5 %

5. CONCLUSION

From the research that had been done, it could be concluded that:

1. The Simple Addictive Weighting (SAW) Method and the Profile Matching Method are proven to have the equal level of accuracy, namely 80-90% in the process of recruiting new employees
2. The Simple Addictive Weighting (SAW) Method and the Profile Matching Method are proven to have the equal level of Sensitivity to recall, namely 80-90% in the process of recruiting new employees.
3. The Simple Addictive Weighting (SAW) Method and the Profile Matching Method are proven to have the equal level of Precision, namely 80-90% in the process of recruiting new employees.

6. REFERENCES

[1] S. Abadi, F. Latifah and K. Kunci, "Sistem Pendukung Keputusan & Kinerja Karyawan Pada Perusahaan Menggunakan Metode Simple Additive Weighting," *Technology Acceptance Model*, vol. 6, no. 4312, p. 37, 2016.

[2] Sunarti and S. Jenie, "Perbandingan Metode SAW dan Profile Matching pada Pemilihan Rumah Tinggal Studi Kasus: Perumahan Depok," *INTENSIF*, vol. 2, no. 2580-409X, 2018.

[3] S. S. Sundari and Y. F. Taufik, "Pegawai Baru dengan menggunakan Metode Simple Additive Weighting (SAW)," *Sisfotenika*, vol. 4, pp. 140-151, 2014.

[4] Kusrini, Konsep dan Aplikasi Sistem Pendukung Keputusan, C.V. Andi Offset, 2007.

[5] M. I. T., A. P. N. and Sulton, "Sistem Pendukung Keputusan Seleksi Karyawan Menggunakan Metode Simple Additive Weighting Pada PT. Philips Seafood Indonesia," *Jurnal Informatika Merdeka Pasuruan*, vol. 1, no. 2503-1945, p. 3, 2016.

[6] N. Nuraeni, "Penerapan Metode Simple Additive Weigting (SAW) Dalam Seleksi Calon Karyawan," *SWABUMI*, vol. 6, no. 2355-990X, pp. 63-71, 2018.

[7] D. W. Syah, E. Santoso and R. S. Perdana, "Sistem Pendukung Keputusan Pengurutan Berdasarkan Jenis Suara Anggota Baru Divisi Paduan Suara BIOS Menggunakan Metode Profile Matching (Studi Kasus : Logicio Choir FILKOM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1(12), no. ISSN 2548-964X, pp. 1678-1686, 2017.

[8] K. Giovani, R. Rekyan and M. A. F, "Sistem Pendukung Keputusan Seleksi Tenaga Pengajar Musik Menggunakan Metode Profile Matching," *Infortech*, vol. 7(13), no. 3115-8170, 2015.

[9] E. Turban, R. Sharda and D. Delen, Decision Support and Business Intelligence Systems, Pearson Education Inc, 2011.

[10] S. Kusumadewi, S. Hartati, A. Harjoko and R. Wardoyo, Fuzzy MultiAtribute Decision Making (MADM), Yogyakarta: Graha Ilmu, 2006.

[11] A. Sudarmadi and E. Santoso, "Sistem Pendukung Keputusan Pemilihan Personel Homeband Universitas Brawijaya Menggunakan Metode Profile Matching," *SISFOTENIKA*, vol. 1(12), pp. 1788-1796, 2017.

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 30-34, 2021, ISSN:-2319–8656

System Design of MPPT Incremental Conductance on Zeta Converter Connected Solar Panel

Jendra Sesoca
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Bambang Siswojo
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Ponco Siwindarto
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Abstract: MPPT Incremental Conductance algorithm has a function to obtain maximum power points on a solar panel. This MPPT Incremental Conductance works based on the P-V curve of the solar panel. In order to obtain better power results, the MPPT Incremental Conductance system will be connected to the zeta converter. The zeta converter is a DC-DC converter that is a development of the SEPIC converter. This converter can also produce good efficiency. In this research will compare the power generated with 2 different methods, solar panels connected with zeta converter without using MPPT Incremental Conductance and solar panels connected zeta converter using MPPT Incremental Conductance. The result of the research obtained is that solar panels connected to zeta converter using MPPT Incremental Conductance can produce better power than not using MPPT Incremental Conductance**.**

Keywords: Incremental Conductance, MPPT, Zeta, Converter, Solar Panel

1. INTRODUCTION

Global warming has increased and created global concern. Greenhouse gas emissions such as carbon dioxide and carbon monoxide are the main causes of global warming. The contribution of these gases comes from the use of coal used as a conventional source of electricity. Conventional power sources are still used as electricity providers for electricity use. The demand for electricity usage that is currently increasing rapidly certainly raises concerns about global warming. Renewable energy sources play an important role as a provider of electrical energy.

Solar energy is one of the renewable energy sources that can be utilized. At this time solar energy has an important role in the world because the energy is unlimited, clean and environmentally friendly. The solar energy can be converted into electricity using solar panels. Solar panels are a technology that can convert sunlight into DC electricity. The electricity generated by solar panels is affected by the large intensity of sunlight.

The main problem of solar panels is the output produced has a small power efficiency value. The impact of the efficiency of small solar panels affects the output power of these solar panels. Therefore, it is necessary to develop and research to improve the efficiency of solar panel power.

MPPT Incremental Conductance method is a simple method but has the advantage to find a good maximum power point despite changes in conditions on solar panels. Changes in conditions such as changes in environmental conditions, such as irradiance and temperature entering the solar panels [1] (Tekeshwar, 2014). The MPPT Incremental Conductance method will then be used in the DC-DC converter. It is enabled to control the work of solar panels in order to maintain the maximum power point that has been obtained.

The Zeta converter is a development of the DC-DC converter, where the Zeta converter generates low

voltage ripples and low current ripples. The Zeta converter can also raise and lower the output voltage at a certain level. The output in the Zeta converter produces the same voltage polarity as its input voltage [2] (Soedibyo, 2015).

Research on comparison analysis on several converters include Cuk, SEPIC and Zeta converters. In this research showed that Zeta converter has a good performance from other converters. In the Zeta converter it produces an output voltage that is not reversed, rise-time and settling-time faster than the Cuk and SEPIC converters. [3] (N. Karthick, 2015).

The next research is to conduct a research study on SEPIC, Luo and zeta converters. The purpose of this researcg was to find out the efficiency produced in each converter. The results showed that Zeta converter produce the best efficiency compared to SEPIC and Luo converters [4] (Niranjana, 2019).

The next research is research on Buck Boost, Zeta and SEPIC converters. The purpose of this research is to find out the exact type of converter used for MPPT systems. The results showed that Zeta converter produce a stable response and produce small ripples of voltage and ripple current. Small voltage ripple and current ripple values will be well used in the process of determining the maximum power point by using MPPT [5] (Prashanth, 2020).

The next research is research on duty cycle control in Cuk converter with several different MPPT. MPPT used is Peturb and Observe with Incremental Conductance. The results showed that the output on the converter using Incremental Conductance provides a better accurate rate for tracking maximum power points than using MPPT Peturb and Observe [6] (Takeshwar, 2014).

Based on the description of previous research, a research converter zeta conducted using MPPT Incremental Conductance connected solar panels.

2. RELATED WORK

Incremental Conductance is an MPPT algorithm that can track the maximum power point value of solar panels. This MPPT system obtains the maximum power point of the solar panel instead of working dynamically following the motion of the direction of the sun, but the MPPT system is working through the approach characteristic of solar panels. In this research using Zeta converter because it produces good power efficiency. The data to be researched are the intensity of light and temperature as solar panel inputs, as well as the value of batteries attached to the Zeta converter . Furthermore, the data that has been obtained in this research will be analyzed. The data observed is the output of zeta converter in the form of voltage, current, and power.

3. METHOD

3.1 Photovoltaic

Photovoltaic is a technology that function to convert solar radiation into electrical energy directly. The PV is usually packaged in a unit called a module. In a solar module consists of many solar cells arranged in series or parallel. While solar cell is a semiconducting element that can convert solar energy into electrical energy on the basis of photovoltaic effects
[7] (Nelly Safitri, Teuku Rihayat, 2019).

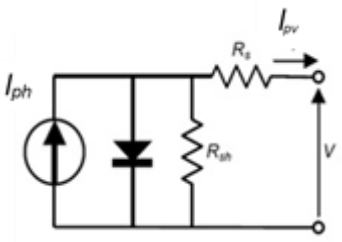

Figure 1. Equivalent Circuit of Photovoltaic

Model mathematic of PV module as shown below Equation. 1.

$$I = I_{ph} - I_s\left(exp\frac{q(V+IR_s)}{NKT} - 1\right) - \frac{(V+IR_s)}{R_{sh}} \qquad (1)$$

Where, I_{ph} is the photocurrent, I_s saturation current, q is the electronic charge, N is the diode factor, K is the Boltzmann's constant, T is the junction temperature in Kelvin, V is the voltage across the diode, R_{sh} is the shunt resistance and R_s is the series resistance [8] (Nema, S. and Nema.). The output characteristic of P-V module with the different irradiance and temperature value shown in Figure 2.

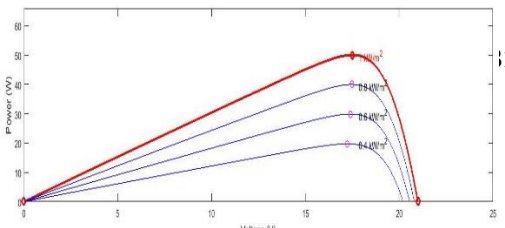

Figure 2. Characteristic of P-V Module

3.2 Zeta

The Zeta converter is a development of a Buck-Boost converter that can produce low output voltage ripples. The Zeta converter works like a Buck-Boost converter that can raise and lower incoming DC voltage. In addition, the output voltage polarity of the Zeta converter is not reversed. The Zeta converter consists of diode, inductor, capacitors and MOSFET. MOSFET component serves as a reasoning that is influenced based on PWM duty cycle value that enters at the foot gate MOSFET [9] (N. Sowmya Smitha Raj & B. Urmila, 2013)

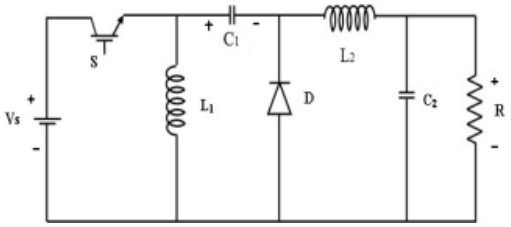

Figure 3. Zeta Converter

The design of the Zeta converter can be obtained from the following formula. The value duty cycle is given as Equation 2.

$$D = \frac{V_O}{V_{in}+V_o} \quad (2)$$

The inductor value is calculated as Equation 3.

$$L_1 = L_2 = L = \frac{1}{2}\left(\frac{V_{in}D}{\Delta I_L \times f_{sw}}\right) \quad (3)$$

The capacitor C1 and C2 value are calculated as Equation 4 and Equation 5.

$$C_1 = \frac{D}{\Delta I_L \times V_{in} \times f_{sw}} \quad (4)$$

$$C_2 = \frac{D}{8 \times \Delta V_{C2} \times f_{sw}} \quad (5)$$

3.3 Incremental Conductance

The Incremental Conductance method works based on the gradient of the P-V curve or the characteristic P-I curve of the solar cell. The maximum working point of solar cells is at different voltage values in each different condition called V_{MPP}. The P-V characteristic of solar cells is the function of power to voltage. Reaches the maximum point when the gradient is zero. [10] (Bharti, M., Kumar, U. 2017). Incremental Conductance (IC) works based on the gradient of the P-V curve or the characteristic P-I curve of a solar cell in search of a Maximum Power Point (MPP) value. Maximum Power Point (MPP) condition is a position where maximum power is obtained in MPPT system. The flowcharts of Incremental Conductance shown in Figure 4. The current and the voltage of PV module are read by MPPT controller. Duty cycle of the converter is increased if Equation 7 is satisfied and duty cycle of the converter is decreased if Equation 8 is satisfied. Duty cycle will no change if Equation 6 is satisfied and MPP has been achieved.

$$\frac{dI}{dV} = -\frac{I}{V} \quad (6)$$

$$\frac{dI}{dV} < -\frac{I}{V} \quad (7)$$

$$\frac{dI}{dV} > -\frac{I}{V} \quad (8)$$

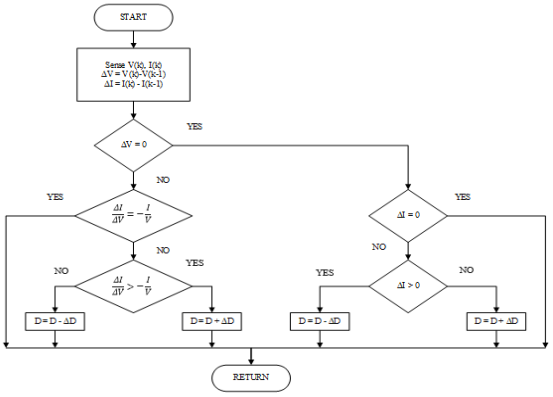

Figure 4. Flowchart Incremental Conductance

3.4 Proposed Zeta and Incremental Conductance

In Zeta converters are connected solar panels using MPPT Incremental Conductance system. MPPT Incremental Conductance serves to find the value of voltage point (V_{MP}) and maximum current (I_{MP}) of solar panels.

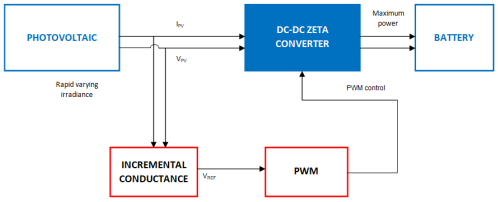

Figure 5. Zeta and MPPT Incremental Conductance

4. RESULT AND DISCUSSION

The Zeta converter generates maximum solar panel power by using MPPT Incremental Conductance. The output of the Zeta converter can be shown in Figure 6. The results showed that zeta converter using MPPT Incremental Conductance can increase output power.

Table 1 shows the comparison of power output generated by Zeta converter using MPPT Incremental Conductance and without using MPPT Incremental Conductance.

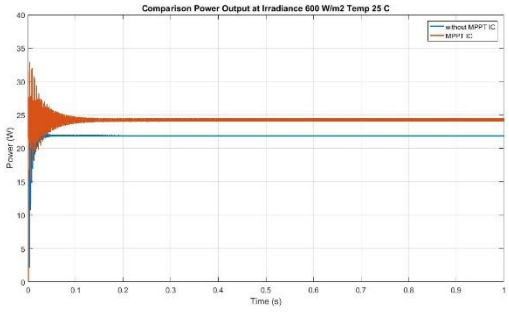

Figure 6. Output Power System at Irradiance 600 W/m^2 and Temperature 25^{0}C

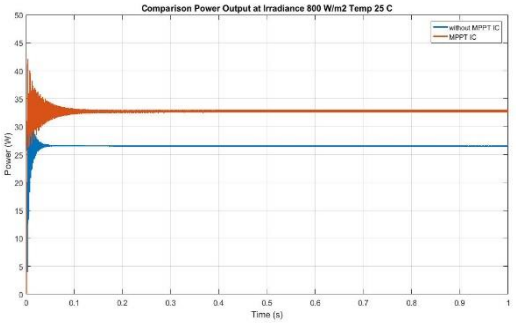

Figure 7. Output Power System at Irradiance 800 W/m^2 and Temperature 25^{0}C

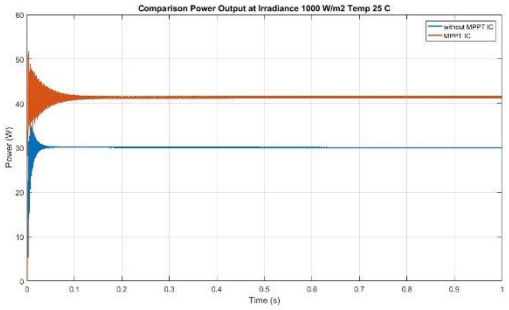

Figure 8. Output Power System at Irradiance 1000 W/m^2 and Temperature 25^{0}C

Table 1. Result Comparison

MPPT	Average Output Power (W)		
	Irradiance 600 W/m^2	Irradiance 800 W/m^2	Irradiance 1000 W/m^2
Non MPPT	21.76	26.40	29.97
Mppt	24.20	32.68	41.29

5. CONCLUSION

Design of Zeta converter using MPPT Incremental Conductance for maximum power point tracking in photovoltaic has been presented.

The results of this research showed that Zeta converters using MPPT Incremental Conductance can find a good maximum power point. The power generated by Zeta converter with MPPT Incremental Conductance is better than without using MPPT Incremental Conductance.

6. ACKNOWLEDGMENTS

This research was technically supported by Brawijaya University 2020. The first author would like to thanks to Dr. Ir. Bambang Siswojo, MT and Dr. Ir. Ponco Siwindarto,M.Eng.Sc from Brawijaya University for encouraging and giving best effort to finish in this research.

7. REFERENCES

[1] Tekeshwar, P.S., Dixit, T.V. 2014. Modelling and Analysis of Peturb and Observe and Incremental Conductance MPPT Algorithm for PV Array Using Cuk Converter. IEEE.

[2] Soedibyo, Budi Amri, Mochamad Ashari. 2015. The Comparative Study of Buck Boost, Cuk,SEPIC and Zeta Converter for Maximum Power Point Tracking Photovoltaic Using P&O Method. Int. Conference on Information Technology, Computer and Electrical Engineering.

[3] N. Karthick, I.Manoj, and K.V.Kandasamy, 2015. Performance Characeteristic of Various DC-DC Converter for Efficient Solar Energy Conversion for Automobile Applications. Journal of Chemical and Pharmaceutical Sciences.

[4] Niranjana Siddharthan, Baskaran Balasubramarian, 2019. Performance Evaluation of SEPIC, Luo and Zeta Converter. International Journal of Power Electronics and Drive System.

[5] Prasanth N.A., Anirudha K.S., B.Manjunath. 2020. Comparative Study of Buck Boost, Zeta and SEPIC DC-DC Converters for Maximum Power Point Tracking Application in PV System. Journal of Emerging Technologies and Innovative Research.

[6] Tekeshwar, P.S., Dixit, T.V. 2014. Modelling and Analysis of Peturb and Observe and Incremental Conductance MPPT Algorithm for PV Array Using Cuk Converter. IEEE.

[7] Nelly Safitri, Teuku Rihayat, 2019. Buku Teknologi Photovoltaic. Yayasan Puga Aceh Riset.

[8] Nema, S., Nema, R.K., Agnihotri, G. 2010. Matlab Simulink based Study of Photovoltaic Cells Modules Array and Their Experimental Verification. International Journal of Energy and Environment (IJEE), Vol 1, Issue 3, pp.487-500.

[9] N. Sowmya Smitha Raj & B. Urmila, 2013. Zeta Converter Simulation For Continuous Current Mode Operation. International Journal of Advanced Research in Engineering and Technology (IJARET), Volume 10, Issue 1.

[10] Bharti, M., Kumar, U. 2017. Virtualization and Simulation of Incremental Conductance MPPT Based Two Phase Interleaved Boost Converter using Simulink in MATLAB. International Journal for Technological Research in Engineering (IJTRE), Volume 4, Issue 9.

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 35-41, 2021, ISSN:-2319–8656

Data Mining Approach for Cyber Security

Varsha P.Desai
Assistant Professor
Department of Computer Studies
VPIMSR, Sangli
India

Dr.K.S.Oza
Assistant Professor
Department of Computer Science
Shivaji University, Kolhapur
India

Dr.P.G.Naik
Professor
Department of Computer Studies,
CSIBER, Kolhapur
India

Abstract: Use of internet and communication technologies plays significant role in our day to day life. Data mining capability is leveraged by cybercriminals as well as security experts. Data mining applications can be used to detect future cyber-attacks by analysis, program behavior, browsing habits and so on. Number of internet users are gradually increasing so there is huge challenges of security while working in the cyber world. Malware, Denial of Service, Sniffing, Spoofing, cyber stalking these are the major cyber threats. Data mining techniques are provides intelligent approach for threat detections by monitoring abnormal system activities, behavioral and signatures patterns. This paper highlights data mining applications for threat analysis and detection with special approach for malware and denial of service attack detection with high precision and less time.

Keywords: Malware, Data Mining, Cyber-attack, Cyber Threat, Ransomware.

1 INTRODUCTION

Data mining techniques are implemented to extract hidden patterns from data. It is scientific research method for analysis, prediction and determine complex relationship between hidden patterns from large volume of data. Knowledge discovery in databases (KDD) process consist of data preprocessing, data cleaning, transformation, mining and pattern evaluation. In data mining classification of data into predefined labeled classes called as supervised leaning. Extracting similar behavioral patterns into different clusters form huge dataset called as unsupervised learning. The gaming technique of data mining where machine learning model is trained to take sequence of complex decisions in uncertain environment as per reward or punishments for specific moves called as reinforcement learning. Association, classification, clustering, regressions, decision tree, Naïve Bayes, Support vector machine, sequence mining, time series analysis are the basics techniques of data mining. Appropriate selection and implementation of data mining technique is depends on the type of data, size of data, complexity and outcome of prediction etc. Artificial intelligence based methods like neural network, fuzzy logic, genetic algorithms, deep learning are used for complex data analysis and prediction of hidden interesting patterns from complex real time database.

Data mining techniques provide systematic approach for discovering vulnerabilities, detection of threats, system loopholes, monitoring intruder's behavior and pattern. Passive attack signatures like scanning open network ports, eavesdropping, phishing, sniffing these passive attacks can be identified by using data mining algorithms. Whereas the active attack signatures like Denial of service attack, malware detection, ransomware detection is possible through data mining and artificial intelligent techniques. Machine learning technique potentially implemented for intrusion prevention system for identifying tricks and methods used by intruder as well as finding vulnerabilities, recording footprints of attack on specific network.

In supervised approach of data mining target variables can be determined according to IP address location, frequencies of web requests and time of requests. Machine learning model used to predict particular IP address is a part of which

attack signature. Implementation of liner and logistic regression, decision tree, support vector machine algorithms are used in supervised learning.

In unsupervised approach of machine learning there is no prediction of target variables while finding association between different patterns in datasets. Computer programs such as malware having similar operating behavioral pattern using clustering & association algorithm.

2 RESEARCH DESIGN

2.1 Type of the research: In the backdrop of above discussion the present research is an attempt to explore certain key aspects of cyber security. Hence the type of the research adopted in this present endeavor is descriptive research.

2.2. Objective of study:

To study data mining techniques for malware detection.

2.3 Scope of the study: The research work is focuses on study of cyber security, types of attacks, network vulnerability, cyber threats and mechanism for malware detection using data mining techniques.

3 RESULT AND DISCUSSION:

3.1 Malware detection using Data Mining:

Malicious computer program which causes abnormal behavior of computer applications though Virus, Trojan's, Warms called as malware. Using classification techniques in data mining malware can be detected and reported to the system administrator. Malware attack on system due to surfing infected websites, games or free apps download, download infected music files, installation of software application extensions, plugins or toolbar and so on. It is important to read warning messages before downloading any application, especially permissions while accessing email or personal data.

3.2 Malware Statistics: As per the research it is found that 80% damage to the system is due to malware attacks [3]. It is found that 92% malware delivered through email attachments. Mobile malware infection increase 54% from year 2018. Overall 98% malware targeted android devices. 99% malware entered through third party app downloads. Out of 10 payloads 7 are ransomware. Overall 18 million websites are infected by malware in one week. 90% financial institutions are targeted of malware from 2018. 40% ransomware victim paid the ransom. More than 50% ransomware attacks demands for bitcoin[17]

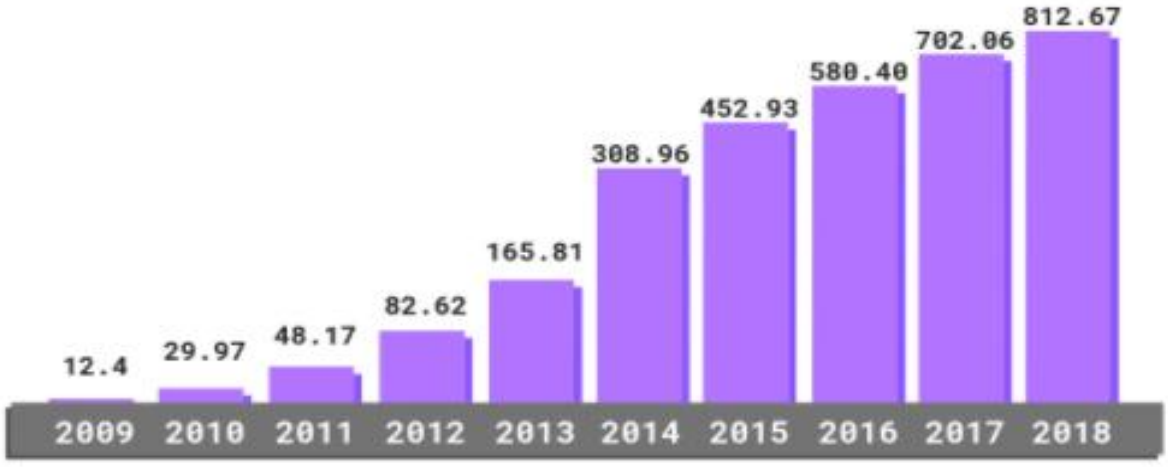

Fig.1 Total Malware Infection Growth Rate (In Millions)[17]

Now a days Malware detection is an important challenge to maintain integrity, confidentially and authentication, non-repudiation of data communicated over the internet. Data mining algorithms helps for early detection of malware as per their behavior and signature stored in database.

3.3 Malware Detection:

In behavioral based malware detection both static and dynamic analysis techniques are used for classification of program as malware. Static analysis for malware detection works on binary code which is complex to analyze and detect malware. Dynamic analysis consist of runtime code execution for testing infected files through virtual machine.[1] Malware are the malicious software code that enters into system through spam mails, email attachments, vulnerable services on internet, downloading process and browser extensions. This causes compromising computer system, unauthorized access of personal data, crippling critical infrastructure, bringing down servers, stealing system as well as network configuration information and so on. Implementation of Future extraction, classification/ clustering techniques of data mining are significant methods for malware detection [2]. Following diagram shows process of malware detection using data mining.

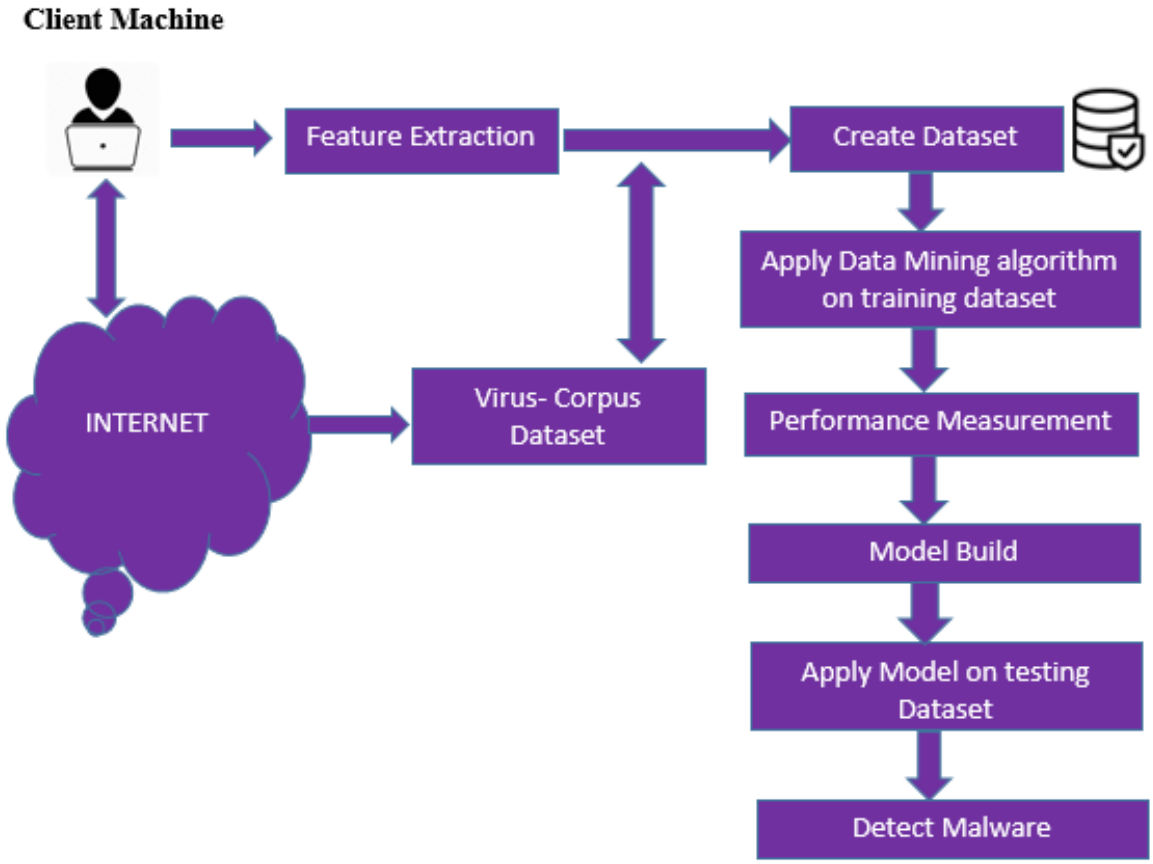

Fig.2. Malware Detection using Data Mining

When client machine is connected to internet during the scanning process machine data fetch to antimalware program. This program start future extraction process by extracting attributes from different files to create dataset. Virus files from corpus dataset is used to store virus definitions. Static analysis, dynamic analysis as well as hybrid analysis techniques are used for extracting features or patterns from data. IDA pro disassembler used to generate assembly files. Abstract assembly files are generated by eliminating operands from assembly code for better results. Extract frequent instruction association from training dataset. Data mining techniques like classification, association rule mining mechanism using Apriori algorithm are applied on training dataset based on their behavior or signature to generate frequent instructions from assembly.[5] Malware detection performance of algorithm is checked using statistical tools. Algorithm is trained until we get expected performance and finally build the model. This trained model is applied on the testing dataset to detect and report malware type and status information.

In static analysis technique of feature extraction PE files are analyses without actual execution. Detection pattern of statistical analysis in the form of windows API, N-grams, string, Opcodes or Control Flow Graph (CFG) techniques. It is one of the useful technique to investigate or explore all possible execution methods paths in malware samples [2]. Artificial neural network techniques is used to detect boot sector virus using N-gram method. Hidden dependencies between code sequences in the malware can be detected using API call method.

In Dynamic analysis debugging or profiling the code by actual execution of code at runtime. This process depend on variable value, program input, system configuration. This analysis mechanism is used for detecting new malware definitions. Detection pattern of statistical analysis in the form of debugger, simulator, emulator and virtual server based environment [2].

Hybrid analysis techniques combines benefits of static and dynamic analysis where packed malware first analyze using dynamic method and the hidden code of packed malware are extracted by comparing runtime execution of malware and its instance is analyze through static model. Hidden files are detected by dynamic analyzer while unpacked file monitor by static model [2]

3.4 Techniques of Malware Detections:

3.4.1 Signature based malware detection:

Signature database store malware footprints of previous attacks. When susceptible code is found it is tested by extracting unique bytes sequence of code as a malware signature. If it matched with existing signature the report as malware and pack malicious code file by anti-malware program. Here anti-malware program need to wait for signature until any device is victim of attack [4]. Data mining techniques like classification, regression are implemented for categorization of threat as a malware using supervised learning approach saves the time and improves the accuracy of prediction than traditional method. This method is easy to run, comprehensive malware information, search and broadly acceptable. [5] Signature database may bypass the threat using some obfuscation, cryptography methods. [4] It fails to detect the polymorphic malware that replicating information in the huge database. [5]

3.4.2 Behavior based malware detection:

Program behavior, speed of execution, response time, browsing habits, cookies information, and kinds of attachments as well as statistical properties helps to detect abnormal behavior or malicious code. In behavior based detection assembly features and API calls methods are applied using data mining algorithm. Unsupervised techniques like clustering, SVM, nearest neighboring algorithms can be implemented for behavior analysis and detection of hidden malware. This method helps to detect polymorphic malwares as well as detect data flow dependencies in the malicious software program. More time and storage space is required to detect complex behavioral pattern. Following table depicts data mining techniques for malware detection:

Type of Malware	Data Mining Techniques	Data Analysis Method
Polymorphic Malware Detection[6]	K-means	Dynamic
Android Malware Detection[7][14]	SVM, J48, Naïve Bayes	Dynamic
API Malware Detection[8]	Naïve Bays, SVM, Decision Tree, Random Forest	Dynamic
N-gram Malware Detection[9]	SVM, ANN	Dynamic
Service Oriented Mobile Malware Detection[10]	Naïve Bayes, Decision Tree	Hybrid
Sequential Pattern Malware Detection[11]	All-Nearest-Neighbor, KNN, SVM	Hybrid
Multi-objective evolutionary Malware Detection[12]	Genetic Algorithm	Static
Frequent Pattern Malware Detection[13]	Graph Mining	Static
Behavioral Malware Detection	Regression, SVM, J48	Dynamic

Table 1: Data Mining Techniques for Malware Detection

Above table depicts different data mining techniques used for malware detection according their signature and behavioral aspects. To extract hidden patterns from the data static, dynamic and hybrid data analysis techniques are used for improving accuracy of malware detection. It is the challenge for cyber security experts to select best algorithm and data analysis techniques for finding the hidden threats and provide alerts to provide data from further attacks.

4. CONCLUSION

Due to globalization usage of internet and communication technology is drastically increase. Data leakage, insecure Wi-Fi connections, lack of security awareness, hardware, software, network vulnerability are the major reasons for cybercrime. To mitigate major risk of cyber-attacks like data benches, ransomware attack, DDos attacks it is necessary to implement efficient as well as intelligent techniques for early detection of cyber threats as a proper security solution. Malware detection is one of challenge for security experts. Data mining techniques like classification, SVM, regression, decision tree, graph mining, KNN algorithms can be integrated with anti-threat system helps to detect malware before enters into system that leads to protect your IT infrastructure form further attack. Artificial neural network, genetic algorithm, deep learning mechanism provides intelligent malware detection from behavior and signature database.

REFERENCES

[1] Monire Norouzi, Alireza Souri, and Majid Samad Zamini (2016),"A Data Mining Classification Approach for Behavioral Malware Detection", Volume 2016, Journal of Computer network and Communications.

[2] Yanfang, Donald Adjeroh, et.al, (2017) "A Survey on Malware Detection Using Data Mining Techniques", ACM Computing Surveys, Vol. 50, No. 3, Article 41.

[3] Rieck. K, Willems.T, et.al (2008), Learning and classification of malware behavior, 5th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: Springer-Verlag, pp. 108–12.

[4] Sara Najari, Iman Lotfi, (2014) "Malware Detection Using Data Mining Techniques". International Journal of Intelligent Information Systems. Special Issue: Research and

Practices in Information Systems and Technologies in Developing Countries. Vol. 3, No. 6-1, pp. 33-37.

[5] Raviraj Choudhary, Ravi Saharan (2012), "Malware Detection Using Data Mining Techniques" International Journal of Information Technology and Knowledge Management, Volume 5, No. 1, pp. 85-88.

[6] Fraley JB, Figueroa M (2016) Polymorphic malware detection using topological feature extraction with data mining. In: SoutheastCon 2016, pp 1–7

[7] Sun L, Li Z, Yan Q, Srisa-an W, Pan Y (2016) SigPID: significant permission identification for android malware detection. In: 2016 11th international conference on malicious and unwanted software (MALWARE), pp 1–8

[8] Fan CI, Hsiao HW, Chou CH, Tseng YF (2015) Malware detection systems based on API log data mining. In: 2015 IEEE 39th annual computer software and applications conference, pp 255–260.

[9] Boujnouni ME, Jedra M, Zahid N (2015) New malware detection framework based on N-grams and support vector domain description. In: 2015 11th international conference on information assurance and security (IAS), pp 123–128

[10] Cui B, Jin H, Carullo G, Liu Z (2015) Service-oriented mobile malware detection system based on mining strategies. Pervasive Mob Comput 24:101–116.

[11] Fan Y, Ye Y, Chen L (2016) Malicious sequential pattern mining for automatic malware detection. Expert System Application 52:16–25.

[12] Martín A, Menéndez HD, Camacho D (2016) MOCDroid: multi-objective evolutionary classifier for Android malware detection. Soft Comput 21:7405–7415.

[13] Hellal A, Romdhane LB (2016) Minimal contrast frequent pattern mining for malware detection. Comput Secur 62:19–32.

[14] Bhattacharya A, Goswami RT (2017) DMDAM: data mining based detection of android malware. In: Mandal JK, Satapathy SC, Sanyal MK, Bhateja V (eds) Proceedings of the first international conference on intelligent computing and communication springer Singapore, Singapore, pp 187–194.

[15] Norouzi M, Souri A, Samad Zamini M (2016) A data mining classification approach for behavioral malware detection. J Comput Netw Commun 2016:9.

[16] Galal HS, Mahdy YB, Atiea MA (2016) Behavior-based features model for malware detection. J Comput Virol Hacking Tech 12:59–67. https://doi.org/10.1007/s11416-015-0244-0.

[17] Retrieved From: https://purplesec.us/resources/cyber-security-statistics/ 22 Dec 2020, 1.30pm.

International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 42-50, 2021, ISSN:-2319–8656

SSH-Brute Force Attack Detection Model based on Deep Learning

Stephen Kahara Wanjau,
School of Computing and Information Technology,
Murang'a University of Technology,
Murang'a, Kenya.

Geoffrey Mariga Wambugu,
School of Computing and Information Technology,
Murang'a University of Technology,
Murang'a, Kenya.

Gabriel Ndung'u Kamau,
School of Computing and Information Technology,
Murang'a University of Technology,
Murang'a, Kenya.

Abstract: The rising number of malicious threats on computer networks and Internet services owing to a large number of attacks makes the network security be at incessant risk. One of the predominant network attacks that poses distressing threats to networks security are the brute force attacks. A brute force attack uses a trial and error algorithm to decode encrypted data such as passwords or Data Encryption Standard keys, through exhaustive effort (using brute force) rather than using intellectual strategies. Brute force attacks resemble legitimate network traffic, making it difficult to defend an organization that rely mainly on perimeter-based security solutions a major challenge. For stopping the occurrence of such attacks, several curable steps must be taken. This paper proposes an efficient mechanism for SSH-Brute force network attacks detection based on a supervised deep learning algorithm, Convolutional Neural Network. The model performance was compared with experimental results from 5 classical machine learning algorithms including Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbour, and Support Vector Machine. Four standard metrics namely, Accuracy, Precision, Recall, and the F-measure were used. Results show that the CNN-based model is superior to the traditional machine learning methods with 94.3% accuracy, a precision rate of 92.5%, recall rate of 97.8% and F1-score of 91.8% in terms of the ability to detect SSH-Brute force attacks..

Keywords: Convolutional Neural Network, Deep Learning, Feature Selection, Network Security, Occam's razor principle, SSH Brute force

1. INTRODUCTION

Computer network attackers have acquire advanced skills and are exploiting unknown vulnerabilities to bypass security solutions. Among the leading network attacks are the brute force attacks [7]. Brute force attacks are becoming harder to successfully detect on a network level due to the growing ubiquity of high-speed networks and increasing volume and encryption of network traffic [25]. A brute force attacking application proceeds through all possible combinations of legal characters in sequence until they find the correct input. The longer the password, the more time it will typically take to find the correct input. Most common brute force attacks use a password dictionary that contains millions of words to test. Successful brute force attacks not only give hackers access to data, applications, and resources, but can also serve as an entry point for further attacks.

Several signs can be construed to be indicators of a brute force attack. Among them include, several failed login attempts from the same IP address; logins with multiple username attempts from the same IP address; logins for a single account from many different IP addresses; failed login attempts from alphabetically sequential usernames and passwords; logins with a referring URL of someone's mail or IRC client; excessive bandwidth consumption over the course of a single session and a large number of authentication failures [24].

Secure Shell (SSH) is one of the most popular communication protocols on the Internet widely used by developers, webmasters, and system administrators. SSH permits one to gain remote access to a new cloud service or a dedicated box in just seconds using an encrypted communication channel. SSH-Brute force attacks tries to gain access to a remote machine by performing authentication attempts, systematically checking all the possible passwords until the correct one is found [26] on the Secure Shell protocol. Normally, attackers may use applications and scripts as brute force tools. These tools try out numerous password combinations to bypass authentication processes. If the host is exposed directly to the Internet or Wide Area Network (WAN) and the SSH service is running on the host, it becomes a subject of constant brute force attacks performed by automated scripts such as hydra. In other cases, attackers try to access web-based applications by searching for the right session ID. Normally, human-selected passwords are characteristically weak as users tend to choose simple passwords which are easier to remember. Sometimes, they don't change the machine's default password or simply use the user name as the password. This makes such machines prone to successful brute force attacks.

In the 2018 report by Verizon [38], brute force attacks were ranked top among attack types detected by IPS (pg.51). Microsoft Office 365 was also a target of massive brute force attacks [32]. The attackers tried logging in with different versions of employees' Office 365 usernames, suggesting they may already possess some combination of employee names and passwords and were seeking valid Office 365 usernames for data access or spear phishing campaigns. The password data could have been obtained in a database breach of a service like Yahoo or a phishing attack, given that password reuse across accounts remains rampant. Alibaba, one of the world's largest retailer and e-commerce Company suffered a massive brute-force attack on its e-commerce site, TaoBao. Using a database of about 99 million usernames and passwords, the attackers managed to compromise approximately 21 million accounts [31].

GitHub, a source code management system, fell victim to massive brute force attacks in 2013, which successfully

compromised some accounts emanating from about 40,000 unique IP addresses [3]. Similarly, on the same year, WordPress, one of the most high-profile open source content management system in use today was a target of massive brute force attacks. A large botnet with more than 90,000 servers attempted to log in by cycling through different usernames and passwords [17]. In 2016, a British national pleaded guilty in German court to launching an automated botnet brute-force attack designed to infect 1.25 million German routers with Mirai malware and causing €2 million ($2.33 million) in damage [30].

In general, to prevent against brute force attacks, network administrators can employ several measures that include (i) adding to password complexity, thereby making any process of guessing a password take significantly longer [28]. For example, some websites will require passwords of 8-16 characters, with at least one letter and number with special characters (such as "$"), as well as not allowing a user to have their name, username or ID in their password, (ii) login attempts that focus on predefined time and number of attempts a user will make to input passwords/usernames [19], (iii) use of Captchas that show a box with warped text and asks the user what the text in the box is. This prevents bots from executing automated scripts that appear in brute force attacks, and (iv) two-factor authentication (a type of multi-factor authentication) that adds a layer of security to the primary form of authentication. Usually, two-factor security requires two forms of authentication (for example, to sign in to a new Apple device, users need to put in their Apple ID along with a six-digit code that is displayed on another one of their devices previously marked as trusted).

Existing tools and methods to detect and avoid these attacks have mainly remained static over the years and are built on common data models such SSH tunneling (also called SSH port forwarding), firewalls and common pairs of username and password [5]. Thus, an efficient mechanism for SSH-Brute force network attacks detection is required in order to organically grow with the ever-expanding attack structures of today's cyber environment.

In recent years, machine-learning techniques have been actively employed to network security and are becoming a prevalent way of detecting advanced attacks with unexpected patterns [26], [29], [37]. More recently, deep neural networks have been applied in studies involving network security [6], [16] since they have a powerful mechanism for supervised learning. They can represent functions of increasing complexity, by including more layers and more units per layer in a neural network [9]. In network intrusion detection, deep neural networks can be used to discover patterns of malicious and benign traffic hidden within huge amounts of structured data [6]. However, its efficacy in the context of SSH-Brute force attacks detection has not been systematically investigated despite its tremendous success in other application domains such as malware detection and spam mail detection. This paper proposes an efficient mechanism for SSH-Brute force network attack detection based on a supervised deep learning algorithm.

2. RELATED WORKS

Several studies have investigated detection of SSH brute force attacks. The study by [14] proposed an approach of detecting attacks in individually sly activities, which operates in unsuspected manner in a SSH Brute-Force attack. The study depended on two elements; Site Aggregates Analyser (to observe the activities and attacks which occur in the sites and detect it) and Attack Participants Classifier (to analyze and classify the attack's participant). Another study [41] proposed a protocol called Password Guessing Resistant Protocol (PGRP), derived upon revisiting proposals previously designed to avoid such attacks. The system was divided into three parts namely: User & Password Authentication, IP Authentication, and Cookie Authentication.

The study conducted by [1] examined brute force attacks based on SSH log files to discover unsuccessful logins and then establish if these unsuccessful authorized IP's belong to attackers or to trusted users. The study suggested a technique, Detecting Distributed Brute Force Attack on a Single Target (DBFST), a strategy to determine who should transact with the IP addresses based on the IP kind, and prevent the attackers IP's from attempting to login to the system. The technique was able to detect the attacks from the same subnet or network, and block the attackers' networks.

In another study by [26], the researchers aggregated network flow data along with a machine learning approach for the detection of SSH brute force attacks at the network level. Classification algorithms (k-Nearest Neighbor, decision trees, artificial neural network and Naïve Bayes) were used to build models for extracting discriminative features for the detection of brute force attacks, collecting real SSH traffic from a campus network. The dataset also contained data similar to attack network traffic (failed login data produced by legitimate users that have forgotten their passwords). They analyzed brute force versus normal SSH traffic in order to define the features that can be discriminative enough to discriminate between normal and attack traffic.

In a recent study, [6] used deep learning for both supervised network intrusion detection and unsupervised network anomaly detection. In their study, a feedforward fully connected Deep Neural Network (DNN) was used to train a network intrusion detection system through supervised learning. An auto encoder was also used to detect and classify attack traffic via unsupervised learning in the absence of labelled malicious traffic. The study evaluated the models using two recent network intrusion detection datasets with known ground truth of malicious versus benign traffic, the CIC IDS 2017 dataset [33] and ISCX IDS 2012 [34] dataset. The study results demonstrated that the DNN outperform other machine learning based intrusion detection systems.

In their study, [16] utilized deep-learning techniques to develop a convolutional neural network (CNN) model to detect network intrusions. The study used CIC IDS 2018 dataset where the numerical data was converted into images for training. The CIC-2018 dataset consist of 10 days of sub-datasets collected on different days through injecting 16 types of attacks generated using CICFlowMeter-V3 [2] containing about 80 types of features. The model performance was evaluated by comparing experimental results with that of a recurrent neural network (RNN) model.

With widespread adoption of cloud computing, coupled with extensive deployment of plenty of Web applications, the need for anomaly detection from the packet payloads is becoming a challenge. Researchers, [22] proposed a feature engineering method that constructs block-based features of the packet payload to adaptively detect anomalies through block

sequence extraction and block embedding. In addition, a deep neural network was designed to learn the representation of the packet payload based on Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) and evaluated the framework using three public dataset namely, CSIC 2010 HTTP dataset [8] which contains generated traffic targeted to an e-commerce Web application, ISCX 2012 dataset that contains network traffic which aims to describe network behaviours and intrusion patterns, and the CIC IDS 2017 dataset that contains both normal traffic and up-to-date attacks which resemble the true data.

Our study builds upon the works by [6] and [16]. The study by [6] showed that deep neural networks can outperform other machine learning based intrusion detection systems, while being robust in the presence of dynamic IP addresses. The study by [16] employed deep-learning techniques and developed a convolutional neural network (CNN) model for network anomaly detection using CSE-CIC-IDS 2018 dataset. Their model performance was compared with a recurrent neural network (RNN) model. The CNN model performance was higher than that of the RNN Model.

This work propose to extend these two studies by proposing an SSH-Brute force attack detection model based on a supervised deep learning algorithm, Convolutional Neural Network. The choice of CNN is because the algorithm can cope with tabular data that contains categorical variables of high cardinality, which are exhibited by the dataset used [6]. The study trained the images based on the proposed model and evaluate its performance by comparing experimental results with that of 5 classical machine learning algorithms namely Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbor, and Support Vector Machine.

3. PROPOSED APPROACH

The study adopted the design science research methodology. The study proposed a method that comprises of feature selection and a deep learning algorithm. Feature selection extract the most relevant features or attributes to identify the instance to a particular group or class. The deep learning algorithm builds the necessary intelligence or knowledge using the results obtained from the feature selection [35] using a dataset. Figure 1 shows the proposed deep learning classifier for SSH-Brute force attack detection.

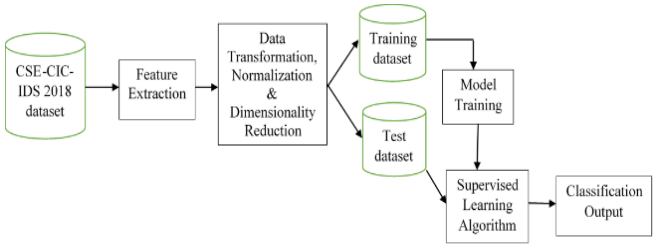

Figure 1: Proposed Deep Learning Classifier for SSH-Brute force attack detection.

3.1 Dataset

This study used, the CIC- IDS 2018 benchmark dataset [16] that includes the contemporary activities of benign and malicious attacks which depicts the real-time network traffic. In this dataset, benign background traffic was collected using B-profile system and contains the characteristics of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols. The network traffic was collected for five days and dumped with normal activity traffic on one day, and attacks injected on other days. The various injected attacks include Brute Force FTP, Brute Force SSH, Denial of Service, Heartbleed, Web Attack, Infiltration, Botnet and Distributed Denial of Service [39].

The dataset was generated using CICFlowMeter-V3 [2] and contains 83 types of features that provide forward and backward directions of network flow and packets, including one column for the label, and another Column for the FlowID. Five of the features are categorical including, 'SourceIP', 'DestinationIP', 'SourcePort', 'DestinationPort' and 'Protocol'.

The remaining 78 features are continuous. This pre-processed network flow data is provided as CSV files that can be fed into the machine learning pipeline. The dataset not only contain modern-day attacks, but is also created in a manner that follow established guidelines of reliable intrusion detection datasets in terms of realism, evaluation capabilities, total capture, completeness, and malicious activity [34]. The dataset contains multinomial class labels for each of the attack types carried out in the flow records. Table 1 shows the dataset used in this work.

Table 1. Type of attacks and amounts of subdata.

Sub-Data	Type of attacks	Total samples
Sub-Data 1	Benign	1,048,574
	DoS Hulk	
	DoS-SlowHTTPTest	
Sub-Data 2	Benign	1,044,751
	Brute Force-FTP	
	Brute Force-SSH	
Sub-Data 3	Benign	1,040,548
	DoS-GoldenEye	
	DoS-Slowloris	
Sub-Data 4	Benign	7,889,295
	DDoS-LOIC-HTTP	
Sub-Data 5	Benign	1,048,575
	DDoS-HOIC	
	DDoS-LOIC-UDP	
Sub-Data 6	Benign	1,042,965
	Brute Force -Web	
	Brute Force -XSS	
	SQL Injection	
Sub-Data 7	Benign	1,042,867
	Brute Force -Web	
	Brute Force -XSS	
	SQL Injection	
Sub-Data 8	Benign	606,902
	Infiltration	
Sub-Data 9	Benign	328,181
	Infiltration	
Sub-Data 9	Benign	1,044,525
	Bot	

Before the data can be used, it requires further pre-processing. The dataset pre-processing steps used are as follows.

3.2 Feature Selection

Feature selection is the process of determining the features to be used for learning by removing those that are not relevant or are redundant [13]. The main goal is to avoid overfitting the data in order to make further analysis possible. Feature selection techniques do not alter the original representation of the data. The subset of features selected followed the Occam's razor principle and also give the best performance according to some objective function. The pre-processed network flow data has 83 columns (e.g., duration, number of packets, number of bytes, and length of packets) that can be used as features, plus one label column and one flow ID column.

Since seven different kinds of attacks are contained in this dataset (i.e., brute-force against the SSH and Web, Heartbleed, botnet, denial of service (DoS), distributed denial of service (DDoS), cross-site scripting (XSS) and SQL injection attacks against websites, and infiltration) [6], a feature subset need to be selected for this study. A filter algorithm was used for subset selection. Filters work without taking the classifier into consideration making them very computationally efficient. The Markov Blanket Filtering (multivariate) method was used which finds features that are independent of the class label so that removing them will not affect the accuracy. According to [12], "a good feature subset is one that contains features highly correlated with the class yet uncorrelated with each other." Since the focus of the study was on detecting SSH-Brute force attacks, this filtering method was considered appropriate in identifying a subset data for the study.

3.3 Data Transformation

In order to develop a CNN-based SSH-Brute-force attack detection model, converting the selected features from the dataset into images was required. For computers, the essence of images is the array of pixel values. In this study, each of the pre-processed record of the dataset was transformed to generate a two-dimensional image matrix group that meets the requirements. Each labelled data was converted into 13x6 size of images since each data contains 78 features except the 'Label' feature. The 'Label' was used for image classification. Given the pixel values of the image range from 0 to 255, the attribute of each record is given as:

$$P_i = \mathrm{r}_i \times 255$$

Where, P_i is the element of the array and the images produced are inputted into the CNN network and the network weights are adjusted until the network learns the most relevant discriminative features for the classification task.

3.4 Data Normalization

The dataset was normalized in order to make convergence faster while training the network. Data normalization is carried out by subtracting the mean from each pixel, and then dividing the result by the standard deviation [4]. Thus, each input parameter would maintain a similar data distribution. For this purpose, the standard min-max scaling was used, a normalization method for scaling data to [0, 1] as follows:

$$X_{norm} = \frac{\mathrm{X} - X_{min}}{X_{max} - X_{min}}$$

Where X_{max} and X_{min} are the maximum and minimum value of feature X respectively. For categorical data contained in the dataset, entity embedding technique was used. Generally, the categorical data which have high cardinality need entity embedding to map them to low-dimensional real vectors in such a way that similar values remain close to each other [11]. This is the case for our datasets since there are many possible values for source IP addresses, destination IP addresses, source port numbers, and destination port numbers. The number of embedding dimensions are determined according to the formula [10];

$$dimension = \left\lfloor \sqrt[4]{possibe\ values} \right\rfloor$$

Where $possibe\ values$ represents the number of possible values a categorical feature can have. A categorical feature is first mapped to an integer between 0 and n-1, where n is the number of unique values that can be taken by the feature, and then encoded as a dense vector of floating point numbers according to the dimensions as calculated above. The parameters (weights) for the dense vector representation are initialized using a random uniform distribution in the range [−0.05, 0.05]. This representation is more computationally efficient, as well as holds inherent relationship information between the categorical feature, the other features in the dataset, and the label.

3.5 Dimensionality Reduction

Several studies have demonstrated that serious redundancy among the characteristic dimensions of network data as well as high correlation exists among the data of each dimension. Further, redundancy and correlation between feature dimensions not only reduce the response time of the intrusion detection system but also affect the learning efficiency of a model training process [40]. Principal component analysis (PCA) is the most normally used method for linear dimension reduction in machine learning, particularly in data analysis and pre-processing. PCA map high-dimensional data to a low-dimensional space representation by linear projection. This study applied PCA in Scikit–Learn library [27] to analyze the variance ratio of each principal component after PCA transformation.

3.6 Labelling

Class identification of the label record was numerically processed, with 0 for Benign, and 1 for SSH-Brute force, thereby enabling one-shot processing of the label in subsequent training and testing.

3.7 Experiment Setup

Deep learning modelling relies heavily on Graphics Processing Unit (GPU) with Compute Unified Device Architecture (CUDA) core enabled. The experiments were performed a server machine running on an Intel® Core™ i7-7500U @2.90 GHz processor, NVIDIA GeForce 940MX, Ms Window 10, 8GB memory machine. The experiments leveraged the popular open source TensorFlow machine learning framework [36] running on Keras backend. The Keras library provides a convenient wrapper for deep learning models to be used as classification or regression estimators. To evaluate the performance of the deep learning classifier, the following two different test cases were considered:

i). Classifying the records as either benign or SSH-Brute force attack with all features.

ii). Classifying the records as either benign or SSH-Brute force categorizing with minimal features.

3.8 Model Design & Training

Convolutional Neural Network (CNN) was used as the deep learning algorithm for model training. CNN is a neural network architecture that uses extensive weight-sharing to reduce the degrees of freedom of models that operate on spatially-correlated features [18]. CNNs are considered a subtype of discriminative deep architecture having shown satisfactory performance in processing two-dimensional data with grid-like topology [23], such as images and videos. CNNs have powerful learning ability mainly due to the use of multiple feature extraction stages (hidden layers) that can automatically learn representations from the data [15].

The basic topology of a CNN is composed of a stack of layers (learning stages) that consists of the convolutional layer, the pooling layer, and the fully connected layers. Figure 2 shows a CNN architecture. Many CNN models have a standard structure consisting of alternating convolutional layers and pooling layers. The last layers are a small number of fully-connected layers with a softmax or sigmoid classifier.

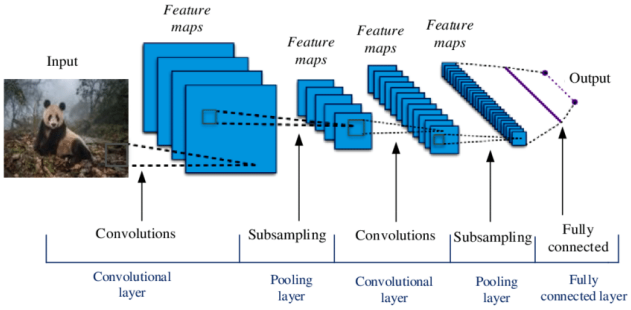

Figure 2: A Typical Convolutional Neural Network (Source:https://en.wikipedia.org/wiki/Convolutional_neural_network)

Each layer performs multiple transformations on the input data using a bank of convolutional kernels (filters). The convolution operation extracts locally correlated features by dividing the image into small slices (similar to the retina of the human eye), making it capable of learning suitable features [15]. Assuming that the input characteristic of the CNN is feature map of the layer z is $M_z(M_0 = \mathrm{X})$, then, the convolution process can be expressed as:

$$M_z = f(M_{z-1} \mathrm{x} P_z + t_{z)}$$

Where P_z is the convolution kernel weight vector of the z layer; the operation symbol "x" represents the convolution operation; t_z is the offset vector of the z layer; and f(x) is the activation function. The convolutional layer extracts different feature information of the data matrix M_{z-1} by specifying different window values, and extracts different features M_z in the data through different convolution kernels. In addition, during the convolution operation, the same convolution kernel follows the principle of "parameter sharing" (i.e., sharing the same weight and offset) which significantly reduces the number of parameters of the entire neural network [40].

The output of the convolutional kernels is then assigned to the pooling layer which not only helps in learning abstraction but also embeds non-linearity in the feature space. The pooling layer usually samples the feature maps in accordance with different sampling rules. Assume that M_z is the input to the pooling layer and M_{z+1} is the output of the pooling layer. Then, the pooling layer can be expressed as:

$$M_{z+1} = \text{Subsampling}\,(M_z)$$

Subsampling helps in summarizing the results and also makes the input invariant to geometrical distortions. Usually, the sampling criterion selects the maximum or the mean value of the window region. In other words, the pooling layer principally reduces the dimension of the feature, thus reducing the influence of redundant features on the model. Table 2 shows the proposed model architecture that was used.

Table 2: Model architecture of the proposed convolutional neural network

Layer	Kernel Size	Stride	Output Size (Width X Length X Depth X Filters)
Input	-	-	50 X 50 X 28
Convolution Layer 1	3 X 3 X 3	1	50 X 50 X 28 X 32
Pooling Layer 1	2 X 2 X2	2	25 X 25 X14 X 32
Convolution Layer	3 X 3 X 3	1	25 X 25 X 14 X 64
Pooling Layer	2 X 2 X2	2	13 X 13 X 7 X 64
Fully Connected Layer	-	-	12,544 X 516

For the realization of the experiments, the study used the proportions 70% and 30% for the training and test datasets, respectively. Since the dataset is highly imbalanced, there was need to ensure there is similar proportion of SSH-Brute force attack records in each training and test sample as there are in the dataset as a whole. This was achieved by stratifying the dataset to ensure the distribution of benign and the malicious traffic is equivalent in both training and test data sets. In addition, a separate hold out validation set was used during the training iterations. The CNN was trained using the back-propagation mechanism [21]. The hyper-parameters used to train the model are presented in Table 3. These parameters were determined empirically according to a set of experiments carried out on the whole dataset that give the best results of classification.

Table 3: CNN training hyper-parameters

Parameter	Value
Activation	ReLu
Loss function	Cross-Entropy
Optimization algorithm	Adam
Epochs	120
Batch size	12
Learning rate	0.05
Weight decay	0.0005
Momentum	0.9
Dropout	0.2
Iterations	30

The CNN consisting of four hidden layers of 64 units per layer was designed. Channeled into these hidden layers is an initial input layer consisting of the embedded categorical variables concatenated with the statistical input features. Each layer estimates non-linear features that are passed to the next layer and the last layer in the deep learning network performed the classification. The activation function on each hidden layer was the ReLU activation function, $R(t) = max(0, t)$ and for batch normalization and regularization, a dropout rate of 0.2 was used on each of the hidden layers to obviate overfitting and speed up the model training [20]. The optimizer used is Adam, with a default learning rate of 0.05. The output layer used a sigmoid activation function and contained 1 neurons for binary classification. The loss function used was binary cross-entropy given by,

$$loss(pd, ed) = -\frac{1}{N}\sum_{i=1}^{N}[ed_i \log pd_i + (1 - ed_i)\ \log(1 - pd_i)]$$

Where pd is a vector of predicted probability for all samples in the testing dataset, and ed is a vector of expected class label, values are either 0 or 1.

3.9 Model Testing and Validation

The proposed model was used to perform a classification task. Classification is the process of assigning a label to an object based on its features translated by its descriptors. The goal is to accurately predict the target class for each case in the data. The experiment was designed to answer two questions: (i) Can a deep learning model be developed to correctly detect SSH-brute force attacks? (ii) How does the resultant deep learning model compare with other machine learning methods used in SSH-Brute force attack detection?

In order to answer the aforementioned question 1, with the features selected, a CNN model was built using the training dataset. The model was then tested and validated by performing a binary classification, classifying network traffic flows as either benign or malicious (SSH-Brute force attacks) from the test dataset. To answer the second question, the study compared the effectiveness of our deep learning model with results obtained from experiments with 5 classical machine learning algorithms, namely Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbour, and Support Vector Machine. Four standard metrics namely, Accuracy, Precision, Recall, and the F-measure were used for comparison.

Accuracy: It's the ratio of the correctly recognized records to the entire test dataset. In this case, the SSH-Brute force attacks. If the model accuracy is higher, the resultant model is better. Accuracy serves as a good measure for the test dataset that contains balanced classes and is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where

True Positive (TP) - the number of SSH-Brute force records correctly classified.

True Negative (TN) - the number of Benign records correctly classified.

False Positive (FP) - the number of SSH-Brute force records wrongly classified as Benign.

False Negative (FN) - the number of Benign records wrongly classified to the SSH-Brute force records

Precision: It estimates the ratio of the correctly identified SSH-Brute force records to the number of all identified SSH-Brute force records. If the Precision is higher, the resultant model is better. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is also referred as True Positive Rate (TPR). It estimates the ratio of the correctly classified SSH-Brute force records to the total number of attack records. If the TPR is higher, the resultant model is better. TPR is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

F1 - Measure: F1-Measure, also called F1-Score, is the harmonic mean of Precision and Recall. If the F1-Score is higher, the resultant model is better. F1-Measure is defined as follows:

$$F1 - Measure = 2\ x\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

4. RESULTS AND DISCUSSION

The proposed CNN-based model was used to classify SSH-Brute force attacks based on the images contained in the testing dataset. The fully connected layer was used where each neuron provide a full connection to all learned feature maps issued from the previous layer in the CNN. These connected layers are based on the sigmoid activation function in order to compute the classes' scores. Using TensorFlow visualization on both training and testing dataset, we can view the accuracy of our approach as shown in Figure 3.

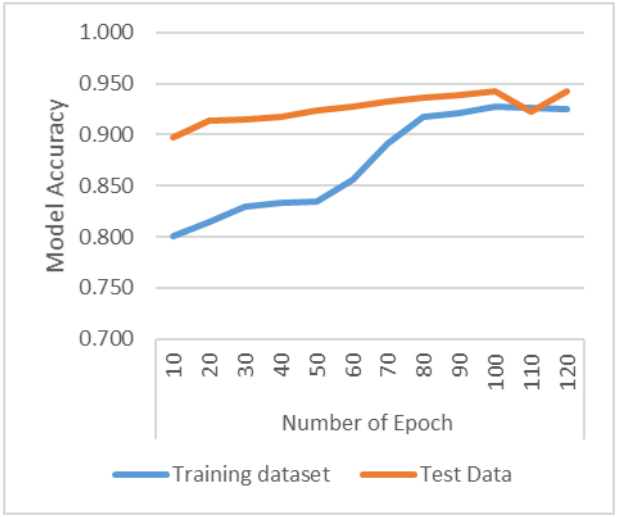

Figure 3: Model accuracy during the training and testing of the network

As shown, the accuracy of the trained data increases as number of steps (epochs) is increasing, until it reaches approximately 94 % of accuracy, which means that there is a change of 94% of detecting any SSH-Brute force attack destined towards the network. The final experimental results of the binary classification using all the features are reported in Table 4.

Table 4: Experimental results using all features

Class	Metric			
	Accuracy	Precision	Recall	F1-Measure
Benign	0.874	0.931	0.967	0.918
SSH-Brute Force	0.943	0.925	0.978	0.918

The results indicate that the model proposed in this study was able to classify SSH-Brute force attacks with 94.3% accuracy, a precision rate of 92.5%, recall rate of 97.8% and F1-score of 91.8%. Table 5 shows the classification using minimal features.

Table 5: Experimental results using minimal features

Class	Metric			
	Accuracy	Precision	Recall	F1-Measure
Benign	0.829	0.883	0.901	0.921
SSH-Brute Force	0.852	0.891	0.922	0.894

The results indicate that when minimal features were used for classification task the model performance was lower in all the metrics compared with the classification task using all the features. Using the minimal features to classify SSH-Brute force attacks, the model achieved 85.2% accuracy, a precision rate of 89.2%, recall rate of 92.2% and F1-score of 89.4%.

Finally, we compared the performance of our proposed CNN-based model with the results of the 5 classical machine learning algorithms, namely Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbour, and Support Vector Machine which used the same dataset.

The comparison of the classification results obtained are reported in figure 4 which shows that the deep learning based model has a better classification accuracy, recall and precision than the other machine learning algorithms. In addition, the F1 score of the model was slightly better, compared with the k-Nearest Neighbour, Logistic Regression and Support Vector Machine models.

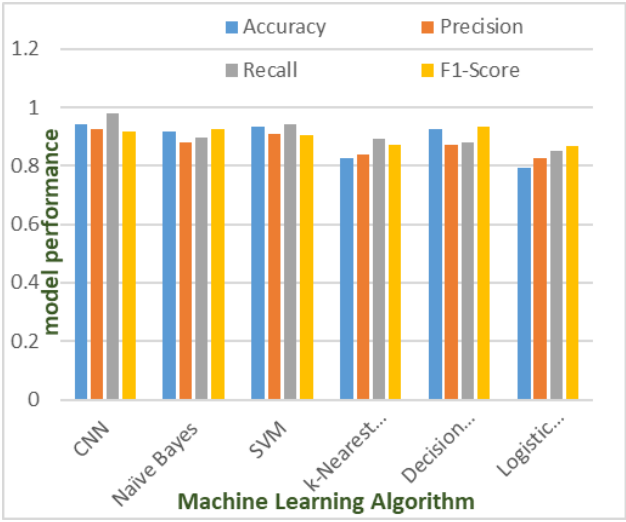

Figure 4: Performance comparison between the classical machine learning models and the CNN-based model.

The above experimental results demonstrates that the convolutional neural network model is superior to the traditional machine learning methods in terms of the ability to detect SSH-Brute force attacks. Utilizing the CICIDS dataset, and the features selected, we found success in classifying unknown network flows into benign and SSH-Brute force attacks. From just these features, we can see that the CNN-based model could identify the traits which characterize an SSH-Brute force attack.

SSH brute force attacks are common in network where an attacker attempts to guess the username and password of a user on the Secure Shell protocol. This type of network attack is simple to perform, with the results from a successfully compromised system triggering a number of destructive outcomes. Previous studies have also demonstrated that deep learning algorithms, particularly, convolutional neural networks, are effective in detecting and preventing these kinds of attacks as an alternative to the firewall techniques used today [6],[16],[22].

5. CONCLUSION AND FUTURE WORK

In this study, we proposed an approach for SSH-Brute force network attacks detection based on a supervised deep learning algorithm. A CNN-based model was designed and tested with the CIC-IDS 2018 dataset that was pre-processed for our experiment. The raw data was converted into images and then used for model training and testing. To evaluate the performance of the resultant model, two different test cases were considered; classifying the network connection record as either benign or SSH-Brute force attack with either all the features or with minimal features. The experimental results showed that our model detects benign and SSH-Brute force attack with higher accuracy and precision when all the features in are used.

Our model was further compared with experimental results obtained from 5 classical machine learning algorithms, namely Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbour, and Support Vector Machine. The results demonstrated that our deep learning model performed better in terms of classification accuracy, recall and precision than

the other machine learning algorithms. In addition, the F1-score of the CNN based model was slightly better, compared with the k-Nearest Neighbour, Logistic Regression and Support Vector Machine models.

In the future, further experiments can be performed with other deep learning algorithms such as Deep Belief Network (DBN), Generative Adversarial Network (GAN) and the results compared with our model. In addition, our model can be tested on a different benchmark dataset such as the ISCX IDS 2012 dataset.

6. REFERENCES

[1]. AL-Zwuiany, M., & Dongjun, H. (2015). DBFST: Detecting Distributed Brute Force Attack on a Single Target. *International Journal of Scientific & Engineering Research, 6*(3), 738-744.

[2]. CICFlowMeter. (2019). *CICFLOWMETER*. Retrieved from netflowmeter.ca: https://www.unb.ca/cic/research/applications.html#CICFlowMeter

[3]. Constantin, L. (2013, November 20). *GitHub bans weak passwords after brute-force attack results in compromised accounts*. Retrieved from PCWorld.com: https://www.pcworld.com/article/2065340/github-bans-weak-passwords-after-bruteforce-attack-results-in-compromised-accounts.html

[4]. Dhakal, A., & Shakya, S. (2018, July). Image-Based Plant Disease Detection with Deep Learning . *International Journal of Computer Trends and Technology (IJCTT), 61*(1), 26-29.

[5]. Faust, J. (2018). Distributed Analysis of SSH Brute Force and Dictionary Based Attacks. *Culminating Projects in Information Assurance, 56*. Retrieved March 19, 2020, from https://repository.stcloudstate.edu/msia_etds/56

[6]. Fernandez, G., & Xu, S. (2019, October 5). A Case Study on Using Deep Learning for Network Intrusion Detection. *arXiv Preprints*, arXiv:1910.02203v1.

[7]. Gautam, T., & Jain, A. (2015, November 10-11). Analysis of Brute Force Attack using TG – Dataset. *SAI Intelligent Systems Conference 2015*.

[8]. Giménez, C., Villegas, A., & Marañón, G. (2012). *HTTP Dataset CSIC 2010.* CSIC (Spanish Research National Council). Retrieved from http://www.isi.csic.es/dataset/

[9]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press. Retrieved from http://www.deeplearningbook.org

[10]. Google. (2019). *Machine learning crash course: Embeddings*. Retrieved from developers.google.com: https://developers.google.com/machine-learning/crash-course/

[11]. Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint*, arXiv:1604.06737.

[12]. Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the 17th International Conference on Machine Learning (ICML '00)* (pp. 359-366). San Francisco, Calif, USA: Morgan Kaufmann.

[13]. Hira, Z., & Gillies, D. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics, 2015*, 1-14. doi:10.1155/2015/198363

[14]. Javedy, M., & Paxson, V. (2013). Detecting Stealthy, Distributed SSH Brute-Forcing. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (pp. 85-96). Berlin, Germany: ACM. doi:10.1145/2508859.2516719.

[15]. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. (2019). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *arXiv.org*, arXiv:1901.06032.

[16]. Kim, J., Shin, Y., & Choi, E. (2019, December). An Intrusion Detection Model based on a Convolutional Neural Network. *Journal of Multimedia Information System, 6*(4), 165-172.

[17]. Kumar, M. (2013, April 12). *Massive Brute-force attack Targets Wordpress sites worldwide*. Retrieved March 16, 2020, from Thehackernews.com: https://thehackernews.com/2013/04/massive-brute-force-attack-targets.html

[18]. Lai, M. (2015). Deep Learning for Medical Image Segmentation, . *arXiv.org*, p. arXiv: 1505.02000.

[19]. Lazarevic, A., Banerjee, A., Chandola, V., Kumar, V., & Srivastava, J. (2008). Data Mining for Anomaly Detection. *Tutorial at the European Conference on Principles and Practices of Knowledge Discovery in Databases*.

[20]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

[21]. Leung, H., & Haykin, S. (1991, September). The complex backpropagation algorithm. *IEEE Trans. Signal Process, 39*(9), 2101-2104.

[22]. Liu, J., Song, X., Zhou, Y., Peng, X., Zhang, Y., Liu, P., & Wu, D. (2019). Deep Anomaly Detection in Packet Payload. *arXiv Preprints*, arXiv:1912.02549v1.

[23]. Liu, W., Wang, L., Liu, X., Zeng, N., & Liu, Y. (2017). A Survey of Deep Neural Network Architectures and Their Applications. *Neurocomputing, 234*, 11-26. doi:10.1016/j.neucom.2016.12.038

[24]. LookingGlassCyber. (2017, October 5). *Protecting Your Network Against Brute Force Password Attacks*. Retrieved from lookingglasscyber.com: https://www.lookingglasscyber.com/blog/threat-intelligence-insights/protecting-network-brute-force-password-attacks/

[25]. Mohammed, M., Degadzor, A., Effrim, B., & Appiah, K. (2017). BRUTE FORCE ATTACK DETECTION AND PREVENTION ON A NETWORK USING WIRESHARK ANALYSIS. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY(IJESRT), 6*(6), 26-37. doi:10.5281/zenodo.802797

[26]. Najafabadi, M., Khoshgoftaar, T., Calvert, C., & Kemp, C. (2015). Detection of SSH Brute Force Attacks Using Aggregated Netflow Data. *IEEE 14th International Conference on Machine Learning and Applications*, 283-288. doi:10.1109/ICMLA.2015.20

[27]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011, October). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825-2830.

[28]. Saito, S., Maruhashi, K., Takenaka, M., & Torri, S. (2016, March). TOPASE: Detection and Prevention of

Brute Force Attacks with Disciplined IPs from IDS Logs. *Journal of Information Processing, 24*(2), 217-226.

[29]. Sangkatsanee, P., Wattanapongsakorn, N., & Charnsripinyo, C. (2011). Practical real-time intrusion detection using machine learning approaches," , vol. 34, no. 18, pp. *Computer Communications*, 2227–2235. doi:10.1016/j.comcom.2011.07.001

[30]. Schwartz, M. (2017, July 24). *Mirai Malware Hacker Pleads Guilty in German Court*. Retrieved from bankinfosecurity.com: https://www.bankinfosecurity.com/mirai-malware-hacker-pleads-guilty-in-german-court-a-10140

[31]. Seals, T. (2016, February 8). Massive Brute-Force Attack on Alibaba Affects Millions. *Infosecurity Magazine*. Retrieved March 16, 2020, from https://www.infosecurity-magazine.com/news/massive-bruteforce-attack-on/

[32]. Seals, T. (2017, July 25). Widespread, Brute-Force, Cloud-to-Cloud Attacks Hit Office 365 Users. *Infosecurity Magazine*. Retrieved March 16, 2020, from https://www.infosecurity-magazine.com/news/widespread-bruteforce-office-365/

[33]. Sharafaldin, I., Lashkari, H., & Ghorbani, A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *4th International Conference on Information Systems Security and Privacy*, (pp. 108-116).

[34]. Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. (2012, May). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computer Security, 31*, 357-374.

[35]. Taher, K., Jisan, B., & Rahman, M. (2019). Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection. *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)* (pp. 643-646). IEEE.

[36]. TensorFlow. (2019). *TensorFlow: An end-to-end open source machine learning platform.* Retrieved from TensorFlow.org: https://www.tensorflow.org/about/

[37]. Tsai, C., Hsu, Y., Lin, C., & Lin, W. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications, 36*(10), 11 994–12 000. doi:10.1016/j.eswa.2009.05.029

[38]. Verizon. (2018). *2018 Data Breach Investigations Report.* Verizon.

[39]. Vinayakumar, R., Alazab, M., Soman, K., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access, 7*, 41525-41550. doi:10.1109/ACCESS.2019.2895334

[40]. Xiao, Y., Xing, C., Zhang, T., & Zhao, Z. (2019). An Intrusion Detection Model Based on Feature Reduction and Convolutional Neural Networks. *IEEE Access, 7*, 42210-42219.

[41]. Yuvaraj, M., Bharathidasan, A., & Kumar, N. (2014). Implementation of Password Guessing Resistant Protocol (PGRP) to Prevent Online Attacks. *International Journal of Computer Science and Mobile Computing (IJCSMC), 3*(2), 815-826.

Stephen Kahara Wanjau received his B.Sc. Degree in Information Sciences from Moi University, Kenya, in 2006 and MSc. Degree in Computer Systems from Jomo Kenyatta University of Agriculture and Technology, Kenya, in 2018. Currently, he is pursuing a PhD in Computer Science at Murang'a University of Technology. He is currently serving as the Director of ICT at Murang'a University of Technology, Kenya. His research interests include machine learning, network security, big data analytics, knowledge management, and cloud computing.

Geoffrey Mariga Wambugu received his B.Sc. degree in Mathematics and Computer Science from Jomo Kenyatta University of Agriculture and Technology (JKUAT), Juja, Kenya, in 2000, the M.Sc. degree in Information Systems from The University of Nairobi, Nairobi, Kenya, in 2012, and the Ph.D. degree in Information Technology JKUAT, in 2019.

He have served for over 10 years' as head of department in higher education institutions in Kenya and also been involved in the design, development, review and implementation of Computing Curricula in different universities in Kenya. Currently he is a Lecturer and Information Technology head in Murang'a University of Technology. His research interests include Probabilistic Machine Learning, Text Analytics, Natural Language Processing, Data mining, Big Data Analytics. At present, He is engaged in university teaching and research supervision.

Gabriel Ndung'u Kamau received his Bed (Art) degree in Mathematics and Business in 1999 from Kenyatta University and Master of Business Administration (Management Information Systems) in 2008 and PhD in Strategic Information System in 2017 from University of Nairobi. Gabriel is also a Certified Network Security Specialist (2020) and Big Data Analyst (2019). Gabriel was a teaching assistant lecturer with Department of Computer and Information Technology, Kenya Methodist University from 2009 to April 2013. In June, 2013, Gabriel Joined Murang'a University of Technology as a lecturer in the department of Information Technology. Gabriel has a vast knowledge and teaching experience in the area of management information systems, information security and applied cryptography, computer forensics, enterprise risk management of information systems, and IT Governance.

His research interest includes ICT4D, cybersecurity and forensics, data analytics, computing and information technology philosophy perspectives.