

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES



ISSN: 2616-9592



Volume: 3

Issue: 2

Jul - Dec

2020



KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

Volume 3, Issue 2, 2020

ISSN: 2616-9592

Frequency Bi-Annual

Editorial Board

Patron

Air Vice Marshal (Retd) Tubrez Asif, HI(M) - President, KIET

Editor-in-Chief

Prof. Dr. Muzaffar Mahmood

Associate Editor

Dr. Muhammad Affan Alim

Managing Editor

Prof. Dr. Muhammad Khalid Khan

Manager Production & Circulation

Mr. Muhammad Furqan Abbasi



College of Computing & Information Sciences
Karachi Institute of Economics & Technology

College of Computing & Information Sciences

Vision

To develop technology entrepreneurs & leaders for national & international market

Mission

To produce quality professionals by using diverse learning methodologies, aspiring faculty, innovative curriculum and cutting edge research, in the field of computing & information sciences.



AIMS AND SCOPE

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by **College of Computing & Information Sciences (CoCIS)** at **Karachi Institute of Economics and Technology (KIET)**, Karachi, Pakistan. **KJCIS** aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at global level.

It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at global level. The journal also fills the gap between academician and industrial research community.

KJCIS focused areas for publication includes; but not limited to:

- Data mining
- Big data
- Machine learning
- Artificial intelligence
- Mobile applications
- Computer networks
- Cryptography and information security
- Mobile and wireless communication
- Adhoc and body area networks
- Software engineering
- Speech and pattern recognition
- Evolutionary computation
- Semantic web and its application
- Data base technologies and its applications
- Internet of things (IoT)
- Computer vision
- Distributed computing
- Grid and cloud computing

OPEN ACCESS POLICY

For the benefit of authors and research community, this journal adopts open access policy, which means that the authors can self-archive their published articles on their own website or their institutional repositories. The readers can download or reuse any article free of charge for research, further study or any other non profitable academic activity.

PEER REVIEW POLICY

Peer review is the process to uphold the quality and validity of the published articles. KJCIS uses double-blind peer review policy to ensure only high-quality publications are selected for the journal. Papers are referred to at least two experts as suggested by the editorial board. All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. We expect our Board of Reviewing Editors and reviewers to treat manuscripts as confidential material. The identities of authors and reviewers remain confidential throughout the process.

COPYRIGHT

All rights reserved. No part of this publication may be produced, translated or stored in a retrieval system or transmitted in any form or by any means; electronic, mechanical, photocopying and/ or otherwise the prior permission of publication authorities.

DISCLAIMER

The opinions expressed in **KIET Journal of Computing and Information Sciences (KJCIS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board and the editorial board. Papers published in KJCIS are processed through double blind peer-review by subject specialists and language experts. Neither the **CoCIS** nor the editors of **KJCIS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead; errors should be reported directly to the corresponding authors of the articles.

Academic Editorial Board

Dr. Ronald Jabangwe University of Southern Denmark, Denmark	Dr. Sardar Anisul Haque Alcorn State University, USA
Dr. M. Ajmal Khan Ohio Northern University, USA	Dr. Yasser Ismail Southern University Louisiana, USA
Dr. Suliman A. Alsuhibany Qassim University, Saudi Arabia	Dr. Manzoor Ahmed Hashmani University of Technology Petronas, Malaysia
Dr. Wael M El-Medany University of Bahrain, Bahrain	Dr. Atif Tahir FAST NUCES, Pakistan
Dr. Asim Imdad Wagan Mohammad Ali Jinnah University, Pakistan	Dr. Maaz Bin Ahmed Karachi Institute of Economics & Tech, Pakistan
Dr. Salman A. Khan Karachi Institute of Economics & Tech, Pakistan	Dr. Taha Jilani Karachi Institute of Economics & Tech, Pakistan

Advisory Board

Dr. Andries Engel brecht University of Pretoria, South Africa	Dr. Mohamed Amin Embi University Kebangsaan, Malaysia
Dr. Rashid Mehmood King Abdul Aziz University, Saudi Arabia	Dr. Anh Nguyen-Duc Norwegian University of Technology, Norway
Dr. Ibrahima Faye University of Technology Petronas, Malaysia	Dr. Tahir Riaz Data Architect, SleeknoteApS, Denmark
Dr. Faraz Rasheed Microsoft, USA	Dr. Mostafa Abd-El-Barr Kuwait University, Kuwait
Dr. Abdul Naser Mohamed Rashid Qassim University, Saudi Arabia	Dr. Mohd Fadzil Bin Hassan University of Technology Petronas, Malaysia
Dr. Syed Irfan Hyder Institute of Business Management, Pakistan	Dr. Bawani S. Chowdry Mehran University, Jamshoro, Pakistan
Dr. Jawad Shami FAST - NUCES, Pakistan	Dr. Nasir Tauheed Institute of Business Administration, Pakistan

Table of Content

1 1- 12	Patient Benefactor Linker <i>Hira Zahid, Sidra Abid Syed, Marissa Jerome, Rida Batool, Sarmad Shams</i>	
2 13-28	Sentiment Analysis through Big Data in online Retail Industry: A Conceptual Quantitative Study on linkage of Big-Data and Assortment Proactive of Online Retailers <i>Muhammad Faisal Sultan, Mewish Jabeen, Muhammad Adeel Mannan</i>	
3 29-44	Robust Food Supply Chain Traceability System based on HACCP using Federated Blockchain <i>Muhammad Danish, Muhammad Shahwaiz Hasan</i>	
4 45-52	Improved User Authentication Process for Third-Party Identity Management in Distributed Environment <i>Kashif Nisar, Shamsuddeen Bala, AbubakarAminu Mu'azu, Ibrahim A. Lawal</i>	
5 53-62	Predicting the Visibility of the First Crescent <i>Tafseer Ahmed</i>	

Patient Benefactor Linker

Hira Zahid¹

Sidra Abid Syed²

Marissa Jerome³

Rida Batool⁴

Sarmad Shams⁵

Abstract

This paper discusses a new evolution in the healthcare sector through a device by investigating the principle application of Artificial Neural Networks (ANN) for the selection of an optimal benefactor-donor match in organ transplantation. The device aims to correlate ABO blood type, age and bone density of healthy subjects. Firstly, linker phase integrates a light intensity(lux) meter and an RGB Color Sensor module to perform an experimental observation of agglutination of RBC's which is measured through a halogen illumination source that measures the light intensity which is displayed on a screen through the microprocessor interface. Secondly, we aim to study the possibility of calcium quantification via near-infrared spectroscopy to estimate bone density which involves the use of an emitting source and a photodiode as a detector/receiver. At last the device involves designing an Artificial Neural Network (ANN) model through the Neural Network Toolbox of MATLAB software to get the optimal network architecture suitable for the analysis. This architecture is achieved by simulating different Artificial Neural Network (ANN) configurations. We used a non-linear ANN which can predict benefactor and patient organ matches, while measuring ABO blood typing and calcium density of the donors in real time and for recognizing mapping functions for which there is no requirement for a particular basis of functions. A database was created through an intensive survey of benefactor profiles. The results generated by ANN are promising for identifying optimal benefactor and patient matches. This approach has potential benefits as an increase in the number of input and parameters will provide better matches and risk associated with human error are reduced. The network can further be modelled to predict survival rates.

Keyword: Artificial Neural Network, RGB color sensor module, infrared spectroscopy, MATLAB Neural Network Toolbox, Optimal benefactor-donor match

1 Introduction

Organ transplantation serves as a life-saving remedy for all the patients that are subject to organ deterioration. [1] The important factors to accomplish this are the availability of organs and the proper matching of donor and recipient organs. Higher life expectancy and improved survival rates have somewhat urged patients to opt for this treatment but lack of organs limits their hope for better lives. Today, the fundamental criteria standard for matching a potential recipient with a suitable donor are ABO blood typing, age, gender and urgency level. [4] The immune system of a human acts up when a foreign invasion takes place, it is obvious that there

¹Biomedical Engineering Department,, Ziauddin University | hira.zahid@zu.edu.pk

²Biomedical Engineering Department, Ziauddin University | sidra.gha@zu.edu.pk

³Biomedical Engineering Department, Ziauddin University | marissa_jerome@outlook.com

⁴Biomedical Engineering Department, Ziauddin University | r.batool94@gmail.com

⁵Biomedical Engineering Department, Sir Syed University of Engg. and Technology | sarmadshams@ssuet.edu.pk

will be immunological incompatibility and this serves as a risk. A perfectly matched organ is considered to lower risks of immunological rejection and infection improving survival rates overall.[5] Since organ transplantation has come into practice it is mandatory for ABO blood typing to be compatible for both benefactor and patient. High risk of hyper acute rejection is the main consequence of ABO mismatch. [6] Organ transplantation and ABO incompatibility are never taken intentionally because of high dying rate of some accidental cases that have been reported [7]. In order to avoid such fatal accidents, it is important to match ABO blood typing in adults, as neonates have delayed development of immunity to antigens [8]. Benefactor age is another important criterion in organ matching. It is said that organs can be donated at any age, donors have given up their organs at the age of ninety years. It is up to doctors to check the viability of these organs. Relative studies and researches show that old age organs can affect the transplantation process. A significantly greater impact occurs when old age donor organs are transplanted, age matters most in liver transplantations. [9] Donor with advance ages has disadvantageous influence on patient continuity for liver patients, this is seen when the donor's age exceed from forty years or over^[9].










ABO BLOOD GROUPS					
Antigen (on RBC)	Antigen A (A ₁ , A ₂ , A _x , etc)  Density ≈ 10 ⁶ /cell	Antigen B  Density ≈ 7.5 x 10 ⁵ /cell	Antigen A + B  Density ≈ 8.5 x 10 ⁵ /cell	Neither A or B  Density ≈ 10 ⁶ H-antigens/cell	Neither A or B or H 
Antibody (in Serum or Plasma)	Anti-B Antibody 	Anti-A Antibody 	Neither Antibody	Anti-A, Anti-B and Anti-A,B 	Anti-A, Anti-B, Anti-A,B and Anti-H 
Blood Type	Type A A-subsets can produce anti-A ₁ (A ₂ =1%; A ₂ B=25%) Anti-B and anti-A ₁ can be clinically significant IgM, IgG, IgA Hemolysis due to complement activation Antibodies found in IVIG IVIG = Intravenous Immunoglobulin	Type B Anti-A is more potent with higher titers than anti-B Can be clinically significant IgM, IgG, IgA Hemolysis due to complement activation Antibodies found in IVIG	Type AB No isoagglutinins Ideal for producing IVIG	Type O Anti-A and Anti-B similar to Type A and Type B blood Anti-A,B mostly IgG Hemolysis due to complement activation Antibodies found in IVIG Anti-A,B recognizes an antigen that is similar but different from A or B, may be difficult to remove	Type Bombay Anti-A, Anti-B and Anti-A,B similar to Type O blood Anti-H highly clinically significant IgM, IgG Rare, likely not ever found in IVIG

Figure 1: ABO Blood groups and their respective antibodies and antigens (from Ref [27])

This study looks into an unusual but important factor that can serve as a matching criterion for organs. The role of calcium as the fifth most occurring element in a healthy person. Prospective studies show that calcium levels, bone density and bone mass are usually lost after successful transplantations. Fractures and bone loss produces significant morbidity, especially during the period of early post-transplant.[10] It is suggested that both donors and patients have optimum calcium levels so that the transplantation process does not affect them so much as it would affect patients with lower calcium and bone density levels. All candidates of organ transplantation should go through a bone health evaluation before proceeding as it would deem beneficial for them. [11] The aim of this research is to devise an artificial neural network that will predict the match percentage between benefactor and patient keeping in mind the three criteria, ABO blood typing, age and lastly the calcium density. Prior work has been done to predict the perfect match for heart transplantation, using data from a registry and then processing it through the algorithm [2].

Figure 1 illustrates the five known blood types or groups. Branch [27] pointed out in his research that the most prevalent type of blood group is the A-type, similarly, for our research, the subjects which were tested to prove our hypothesis were majority A-blood type participants. The percentage of Type O and Type A is approximately the same, as indicated by Branch, however, among the individuals tested in our study, Type O was the second populous group in the sample. Type “Bombay” did not exist in the study sample. Since Bombay lacks anti-A and anti-B if the sample had at least one individual with this type the results may have significantly varied. On the other hand, “artificial neural networks (ANNs)” are the algorithms which computer-based and modelled at the behaviour and composition of neurons in the brain and can be applied for identification and classification of complex patterns [12].

The parameters of the ANN are adjusted to achieve the pattern recognition through a process known as “error minimization” which corresponds to skill development through experience. Calibration of ANN may be done through input data of any type, such as levels of gene-expression generated by microarrays of cDNA. The output of trained ANN may be grouped into different classes. There are various modern world applications of ANNs such as the application in clinical problems ones such examples is its use to diagnose myocardial infarcts[13] and arrhythmias[14] through the interpretation of radiographs and electrocardiograms.[16] and magnetic resonance images[15]. We applied ANNs to obtain a perfect match between benefactor and donor organs depending upon the three criteria. This will be done by collecting patient data and training the network over the three parameters and then the benefactor data will be collected by the linker in the form of ABO blood group and calcium density.

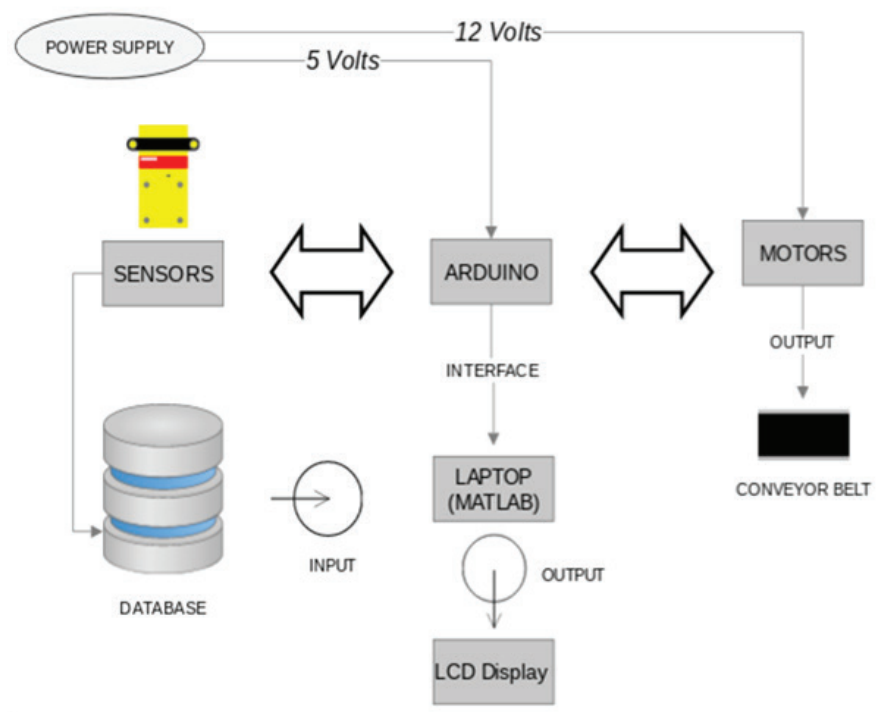


Figure 2: Hardware Flow (Source: self-made)

After the network has been trained, the inputs will be added from the Red blood cell agglutination sensor and the calcium density sensor as shown in Figure 2. The ANN will process the data and as a result, will display the percentage match of organs on the liquid crystal display. Figure 3 displays how the network will flow, the decision box displays the processing of data and working of an algorithm which then results in a digital output of either 1 or 0.

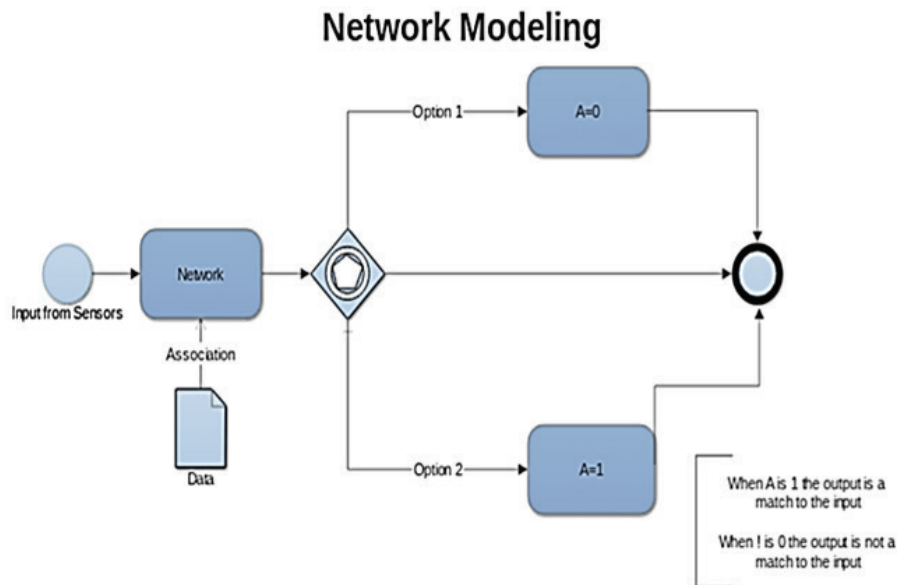


Figure 3: Example of the working of the Network (Source: self-made)

2 Materials And Methods

A ABO Blood Typing Using a Luminosity Sensor

Blood type is typically explored via data recordings of light intensity (lux) calculations that occur within the range from 0.1 - 40000 Lux. However, observation and interpretation of the data exclusively cannot be utilized to establish a cross-match between recipient and Donor blood type. A relationship can be generally identified using a model-based approach. The approach undergoes learning and prediction steps.

Light Source Selection

Previous studies investigated the effects of wavelength (500 to 900 nm) on aspects of RBC aggregation for regular blood suspensions which recorded both reduced and accelerated aggregation statically and dynamically [18].

Results from experiments show that green laser beam (527 nm) performed better for determination of RBC agglutinations in blood typing tests [17]. The routine immunohaematological tests that are performed by manual techniques that do not use a light source to confirm the results. In the conventional tube technique examination, Anti-A and Anti-B both are

required to deduce if red blood cells acquire or do not have A and or B blood group antigens. “Negative test result” is a lack of agglutination, which illustrates the absence of corresponding investigated antigen. “Positive test result” is the agglutination of red blood cells with given reagent, which indicates the existence of the desired ABO antibody.

B Predicting Match Using a Model-free Approach

Model-Free Approach

The term model-free refers to construction of a machine learning a computer-based tool, rather than a physical model to approximate a numerical function. This numerical function is then used to produce a map between the given inputs. For our research, we used Artificial neural networks (ANNs) for recognizing mapping functions for which there is no requirement for a particular basis of functions. Subject-specific neural models have been established from experimental data that primarily include light intensities and electric potential differences.

In Model-free approach, equations are not used to form relationships thus the process is free from equations modelling it. Therefore, we used the precise approach of a “black-box” where intermediate functional relationships between the observed experimental parameters are done through “macroscopic transfer function” instead of separate modelling. In the context of Cross-matching the mismatch of Donor-Recipient factors such as “gender, allograft, ischemia time, medical conditions prior to transplant, and human leukocyte antigen (HLA)” have all been classified as risk determinants for organ rejection but may not be benefited in the matching of organ. [19, 20, 21, 22]

Age, ABO blood group compatibility and Bone density measurements (in case of Bone graft implants and also to estimate the overall health of Donor) are the parameters chosen to be the constitutive units to the neural network. Conversely, a method to ascertain the conversion of each input variable into output variables via constitutive unit. The term used to describe this relation is called “functional relation”.

Due to the inborn complication of “coupled nonlinear biological systems”, computational model production is needed for quantitative awareness of their structure and function in medical studies [23]. However, there exists a limitation to this approach a single macroscopic nonlinear transfer function may not prove to be adequate enough to apprehend the complex various “Intermediate nonlinear components” in the observed variables. For data fitting difficulties we desire artificial neural network to sketch between a data set of numeric inputs and of numeric targets. In this context, the interpretation of the regression map is as follows, the more complicated the regression map, the more it will over fit the data resulting in deficit of generalization (i.e. classic overfitting problem).

3 Bone Densitometry Using Near-Infrared Spectroscopy

The traditional method to estimate bone mineral density is by using the “Dual-energy x-ray absorptiometry (DXA)”.

Table I: Comparison of radiation doses. Dose limits for occupational exposures are expressed in equivalent doses (from Ref. [24])

Type	Model	Patient Dose(microSv)
i) Body CT scan		5,000-15,000
ii) Head CT scan		2,000-4,000
iii) Lumbar Spine X-ray		600-1,700
iv) Lateral Spine X-ray		820
v) Dental Bitewing		60
vi) Chest X-ray		
vii) DEXA Total body	Lunar Prodigy	0.37
viii) DEXA Total Body	Lunar DPX-L	0.20

Table 1. Shows that the radiation dose for a DEXA scan is very small as compared to other modalities however there is a chance of cancer from excessive exposure to radiation. Radiations may cause cancer due to mutation of cell. The aim is to design a device to measure bone density using the near-infrared emitter and Photo-detector.

A *Optical Wavelength*

According to the light absorbance properties at near-infrared of skin and bone, it is known that light absorbance of skin and bone largely differs with changing wavelength [25]. In order to select the optimal light source for the hypotheses, two requirements were necessary. To ensure that the collected spectral data is limited to the area of tissue of interest firstly such wavelength of light would be considered which could pass the skin and penetrate into the bone so as to estimate the wavelength of light that was transmitted through the bone. The high value of absorbance of light by the bone can clearly identify the difference of bone density [25]. Secondly, the wavelength of light must pass the skin or absorb light so that experimental results would only provide bone absorption. It was observed that the penetration depth of NIR fluctuated from approximately "1 mm to 2 mm in the 4000-5100 cm⁻¹ range to approximately 3 mm in the 5100-7000 cm⁻¹ range, and to approximately 5 mm in the 7000-9000 cm⁻¹ frequency range" [26]. These findings suggest that the chosen optical wavelength which is 1720 nm at a distance of 25cm could be used for the experiment.

4 Experiment

In order to understand the thickness or bone density, different wavelength were used to illuminate the target area. NIR spectra helped detect the thickness using NIR spectrum. The light source was placed on the thumb or any finger of the subject and the transmitted light was detected by the photodetector. The correlation coefficient R determined confirms the high potential of an optical wavelength at 1720nm. For the blood sample clots, it implies that the blood has reacted with the antibody present in the reagent/Antisera. For instance, if the blood forms agglutination with Antisera A, it has antigen A. The slides may be viewed under a microscope to confirm agglutination. However, we have used a lux sensor in our project to perform the task.

A Training Algorithm

For faster training and processing of input data, a high-performance training algorithm was required. Two main categories of faster algorithm exist. The first category uses a heuristic approach and the other category uses standard numerical optimization techniques. We have used the numerical optimization technique. The best correlation coefficient was observed using Levenberg-Marquardt Training Algorithm. This method is applied whenever the current solution is far from the accurate solution. In case the current solution is close to the accurate solution LM applies the Gauss-Newton method. LM is capable to alternate between a slow decline approach when moving far from the minimum and a swift convergence when being at the minimum vicinity.

1 Network Layers

ANNs are structured by layers of neurons. For the study, two layers which are mandatory are the output and input layer. Hidden layer is in between the input and output layer. A hidden layer can be more than one in number. Hence there exist an arbitrary number of hidden layers each of them being an approximate size. The number of hidden layers for our project is 20. A suitable correlation was obtained using 20 hidden layers and one input and one output layer.

2 Learning Rate

During training in order to measure a more correct output the network figures out the direction in which link value and each bias value can be changed at each step. Learning depends upon the learning rate. The higher the learning rates the faster the learning. Learning rate can be set to a maximum value of 1.0.

3 Epochs

The number of epochs the network uses to perform learning was a total of 1000 epochs.

B Participants

A data sample of 2000 patients included in this study was extracted from a hospital with the consent of the patients and healthy subjects were chosen randomly, their medical profiles were extracted from records for research purposes. Informed consent was obtained by the hospital from patients prior to the start of the study. The full dataset was used to train the network. Dataset consisted of three important parameters which were Age, Blood group and calcium density. The main selection criteria for subjects was their age. Individuals were grouped and a data set ranging from 25 to 60 years of age was constructed.

5 Results

The efficacy of the network is promising. In order to analyze the efficacy of the model, we have utilized a classifier to adjust the imbalanced data. This data is split into training and testing data sets. The training data set is for the purpose of setting and computing hyper-parameters. The test data set is for appraising the classifiers. The performance is adjusted by estimating the means and standard deviations. Efforts have been made to reduce the computation time. The time taken for the

simulation of the network is crucial because we aim to provide instant results but due to a large dataset, a trade-off between computational efficiency and performance is observed.

The results from the NNtool application in MATLAB used for the study provided below:

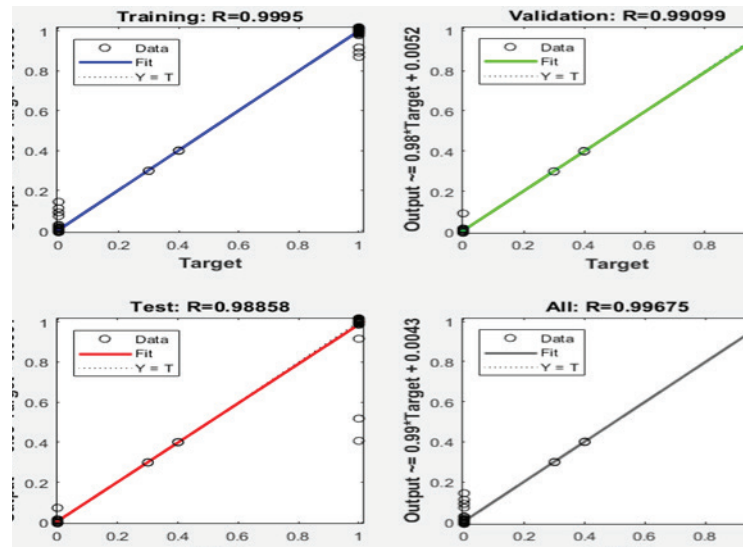


Figure 4: Regression Plot of the neural network. The correlation coefficient $R=0.988$ for test input. Hidden layer = 10 and the network is trained using the Levenberg-Marquardt algorithm

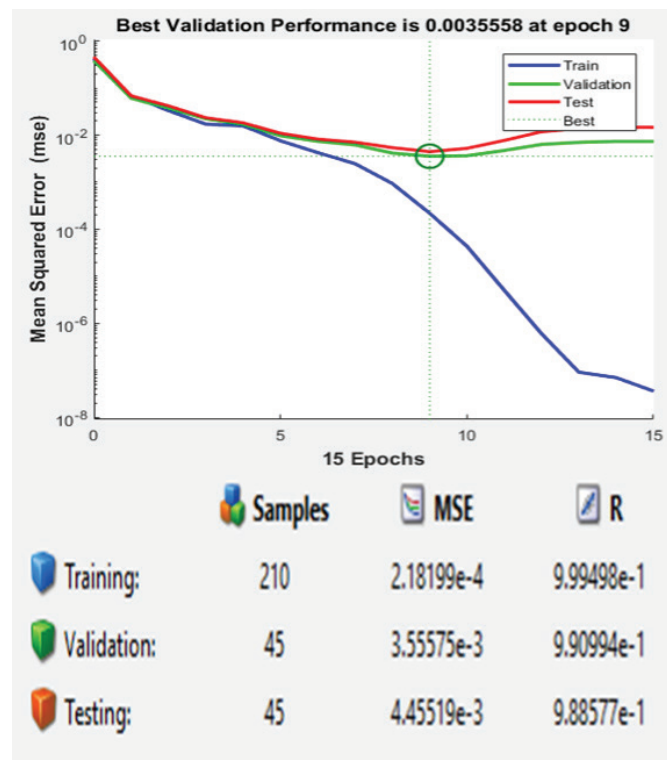


Figure 5: (Top) Plot performances at epoch 9. A total of 15 epochs were used to train the network. (Bottom) Mean square error for 10 hidden layers.

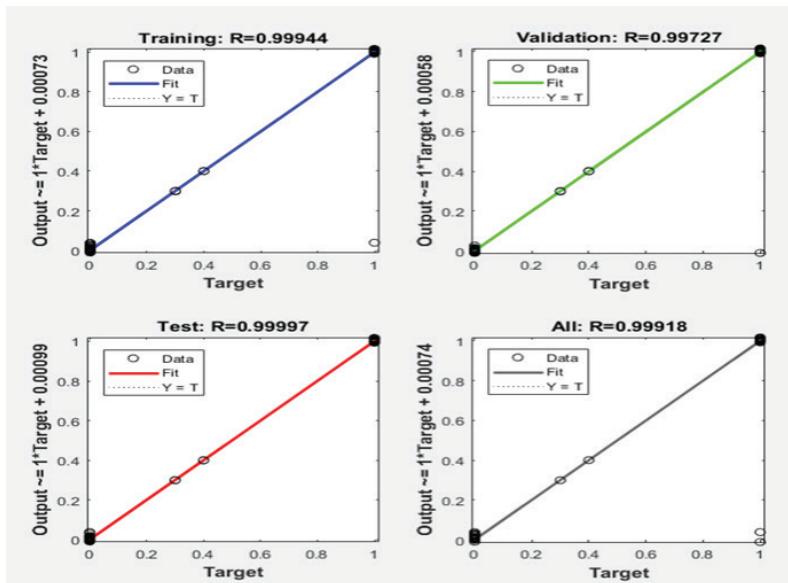


Figure 6: Regression plot for 20 hidden layers. Input= 2000

Results			
	Samples	MSE	R
Training:	1400	2.29792e-4	9.99437e-1
Validation:	300	1.14221e-3	9.97268e-1
Testing:	300	1.16214e-5	9.99972e-1

Buttons: Plot Fit, Plot Error Histogram, Plot Regression

Figure 7: Mean square error for 20 hidden layers. Input=2000

6 Discussion

In this paper, we present a non-linear ANN which can predict benefactor and patient organ matches, while measuring ABO blood typing and calcium density of the donors in real time. This is important as many patients are already waiting for transplants and when they do pass the waiting list there is an immense risk of a mismatch. The network has shown to be somewhat successful in achieving the hypotheses laid out in the beginning. The predicted results have, however, not been achieved yet. Artificial Intelligence is the future and eventually, all problems in medicine will be overcome. The network will hopefully work best in the future with a bit tweaking in the proposed network architecture. At the moment it lacks a bit of accuracy and validity but it proves to be reliable. The RBC agglutination sensor has given us promising results in identifying the donor blood group which was questionable at the beginning. A relatively cheaper method of determining calcium density in the body has been worked out and it has proven to give great results, with a 70-80% match to the original DXA values.

7 Conclusion

This study introduced to explore another technique for the cross-linking of beneficiary benefactor for patients requiring a transplant by utilized artificial neural systems. The device is designed as a significant tool which will not only help the clinicians and doctors to predict the best match but will also reduce patient death due to the rejection of organs. To achieve a profound understanding of the risk factors that cause organ rejection and influence long-term survival rates, the development of the subject-specific data-driven network is a key facet. This approach shows potential to accurately predict the best matches and provides concrete opportunities for easier and more reliable healthcare systems. The proposed device is a useful tool in clinical practice and can also be used to automate other medical procedures. The development of such models in the biomedical field can eradicate human-associated error to a greater degree. The goal of our project was to illustrate 4 to 5 possible matches for a patient and the goal has been achieved with a correlation coefficient of $R=0.99$. This device exchanges the data obtained from several patients profile to train the network then finally the test data is obtained using the RGB and bone density measuring sensor. The matching is purely done by the ANN and hence no doctor/physician is to be blamed for the outcome.

References

- [1] Simon, DM. Levin, S., 2001. Infectious Complications of Solid Organ Transplantations. *Infectious Disease Clinics of North America*, 15(2), 521-549
- [2] Nilsson, J. Ohlsson, M. Höglund, P. Ekmehag, B. Koul, B. Andersson, B., 2015. The International Heart Transplant Survival Algorithm (IHTSA): A New Model to Improve Organ Sharing and Survival. *PLoS ONE-Public Library of Science*, [online] Available at <<https://doi.org/10.1371/journal.pone.0118644>> [Accessed 10 October 2017]
- [3] Opelz, G. Wujciak, T. 1994. The influence of HLA compatibility on graft survival after heart transplantation. *The Collaborative Transplant Study. The New England Journal of Medicine*, 330(12), 816-819.
- [4] Costanzo, MR. Costanzo, MR. Dipchand, A. Starling, R. Anderson, A. Chan, M et al. 2010. The International Society of Heart and Lung Transplantation Guidelines for the care of heart transplant recipients. *J Heart Lung Transplant*, 29(8), 914-956
- [5] Russo, MJ. Iribarne, A. Hong, KN. Ramlawi, B. Chen, JM. Takayama, H et al., 2010. Factors associated with primary graft failure after heart transplantation. *Transplantation*, 90(4), 444-450
- [6] West, LJ. Karamlou, T. Dipchand, A. Pollock-Barziv, SM. Coles, JG. McCrindle, BW. 2006. Impact on outcomes after listing and transplantation, of a strategy to accept ABO blood group-incompatible donor hearts for neonates and infants. *The Journal of Thoracic and Cardiovascular Surgery*, 131(2), 455-461
- [7] Cooper, D.K, 1990. Clinical survey of heart transplantation between ABO blood group-incompatible recipients and donors. *The Journal of Heart Transplantation*, 9(4), 376-381.
- [8] West, LJ. Pollock-Barziv, SM. Dipchand, AI. Lee, KJ. Cardella, CJ. Benson, LN et al. 2001.

- ABO-incompatible heart transplantation in infants. *The New England Journal of Medicine*, 344(11), 793-800.
- [9] Mutimer, DJ. Gunson, B. Chen, J. Berenguer, J. Neuhaus, P. Castaing, D. Garcia-Valdecasas, J. C. Salizzoni, M. M. G. E. Mirza, D., 2006. Impact of Donor Age and Year of Transplantation on Graft and Patient Survival Following Liver Transplantation for Hepatitis C Virus. *Transplantation*, 81(1), 7-14.
- [10] Torres, A. Lorenzo, V. Salido, E. 2002. Calcium Metabolism and Skeletal Problems after Transplantation. *Journal of the American Society of Nephrology*, 13(2), 551-558.
- [11] Foundation, NK. 2003. K/DOQI clinical practice guidelines for bone metabolism and disease in chronic kidney disease. *American Journal of Kidney Diseases*, 42(4 Suppl 3): S1-201.
- [12] Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Oxford: Clarendon
- [13] Heden B, Ohlin H, Rittner R, Edenbrandt L. 1997. Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*. 96(6), 1798-1802
- [14] Silipo R, Gori M, Taddei A, Varanini M, Marchesi C. 1995. Classification of arrhythmic events in ambulatory electrocardiogram, using artificial neural networks. *Computers and Biomedical Research*. 28,305-318.
- [15] Ashizawa K, ET al. 1999. Artificial neural networks in chest radiography: application to the differential diagnosis of interstitial lung disease. *Academic Radiology*, 6, 2-9.
- [16] Abdolmaleki P, ET al. 1997. Neural network analysis of breast cancer from MRI findings. *Radiation Medicine*. 15, 283-293.
- [17] Chang, YJ. Chang, W. Lin, YT. 2014. Detection of RBC agglutination in blood typing test using integrated Light-Eye-Technology (iLeyeT). *International Symposium on Bioelectronics and Bioinformatics, IEEE*. [e-journal] <http://dx.doi.org/10.1109/ISBB.2014.6820952>
- [18] Niemann, C. Divol, L. Froula, D. Glanzer, S. Gregori, G. Kirkwood, R. Mackinnon, A. Meezan, N. Moody, J. Source, C. Bahr, R. Seka, W. 2005. *IEEE International Conference on Plasma Science*. [e-journal] <http://dx.doi.org/10.1109/PLASMA.2005.359167>
- [19] Weiss ET al. 2009. Impact of recipient body mass index on organ allocation and mortality in orthotopic heart transplantation. *Journal of Heart and Lung Transplantation*. 28(11), 1150-1157
- [20] Russo, M. J. et al. 2007. The effect of ischemic time on survival after heart transplantation varies by donor age: an analysis of the United Network for Organ Sharing database. *The Journal of Thoracic and Cardiovascular Surgery*. 133, 554-559.
- [21] Smith, J. D., Rose, M. L., Pomerance, A., Burke, M. Yacoub, M.H. 1995. Reduction of cellular rejection and increase in longer-term survival after heart transplantation after HLA-DR matching. *Lancet* 346, 1318-1322

- [22] Kilic, A. et al. 2012. What predicts long-term survival after heart transplantation? An analysis of 9,400 ten-year survivors. *The Annals of Thoracic Surgery*.93, 699–704
- [23] Winslow, R. L., Trayanova, N., Geman, D. Miller, M. I.2012. Computational medicine: translating models to clinical care. *Science Translational Medicine*, 4(158) [e-journal] <http://dx.doi.org/10.1126/scitranslmed.3003528>.
- [24] Albanese, CV. Diessel, E. Genant, HK.2003. Clinical applications of body composition measurements using DXA. *Journal of Clinical Densitometry*.6 (2), 75-85.
- [25] Chaichanakol, S. Tanaka, SM. Khantachawana, A. 2016. Quantitative Detection of Calcium using Near Infrared Spectroscopy for apply in Bone Densitometry. *International Journal of Mechanical and Production Engineering*.4 (4).
- [26] Padalkara, M.V. Pleshkoa, N. 2015. Wavelength-Dependent Penetration Depth of Near Infrared Radiation into Cartilage.*Analyst*.140 (7), 2093-2100.
- [27] Branch, D.R. “Anti-A and anti-B: what are they and where do they come from?” *Transfusion*, vol. 55, pp.S74-S79, July 2015.

Sentiment Analysis through Big Data in online Retail Industry: A Conceptual Quantitative Study on linkage of Big-Data and Assortment Proactive of Online Retailers

Muhammad Faisal Sultan¹

Mehwish Jabeen²

Muhammad Adeel Mannan³

Abstract

Big-Data is the recent trend in data sciences prevailing all over the globe. The tool aids significantly in optimization of knowledge and has predominant use in optimization of knowledge and productivity. However, there is lack of understanding of concept and its application in Pakistan as indicated by Gallup Pakistan (2018) and stream of data is going to be doubled in two years' time Tankard (2012). Therefore, there is a definite need of research which optimizes understanding associated with technology and its application from the context of Pakistan. Hence considering the application of big-data in retail sector this study aims to explore the impact of sentiment analysis through relating impact of big-data with effective assortment s of online stores. Although data has been collected from IT experts associated with online retail sector via quota sampling and SMART-PLS has been incorporated for the purpose of analysis. Results of the study highlights that big-data is perceived as the major tool for the betterment of assortment in online retail stores although data scientist and their applicability might diminish the impact of the use of big-data.

Keyword: Big-Data, Sentiment Analysis, SMART-PLS, Assortment and Sentiment Analysis.

1 Introduction

Big-Data is not a new terminology which is developed in 2011 after the continuous and thorough efforts made in the field of data management (Surbakti, Wang, Indulska & Sadiq, 2020) In recent times organizations are gathering heaps of data in order to gain benefit in upcoming time. Therefore, there is a necessity of managing this enormous amount of data effectively so to extract optimal information at the time of need (Fan & Bifet, 2013). Although in the time of need one of the latest IT trend trends i.e. Big-Data is achieving massive attention from researchers & practitioners (Bollen, Mao & Zeng, 2011). There are numerous advantages of using big-data which will provide access to new markets and aids in innovation of business model. Moreover, the technology will also aid in assessing customers' needs which also results in betterment of customer service (Rajeb, Rajeb & Keogh, 2020).

2 Statement of Problem

Rajeb (2020) postulated that there are some legitimate indications regarding increase of big-data application in terms of marketing. However, study uses the reference of Pantano Giglio and Keogh (2019) that regard less of increase in literature on big-data on various marketing

¹Khadim Ali Shah Bukhari Institute of Technology (KASBIT) | mfaisal@kasbit.edu.pk

²Khadim Ali Shah Bukhari Institute of Technology (KASBIT) | mehwish.jabeen@kasbit.edu.pk

³Hamdard University | adeel.mannan@hamdard.edu.pk

functions although literature provide minimal evidence for association of big-data & marketing functions. Although it's a time when each and every industry is striving to know how big-data might be associated with the solving problems. In fact, some of the industries have already implemented this technology (Le & Liaw, 2017). However, there are significant lacking in the use of big-data technology in Pakistan and only NADRA has some advanced mechanism of big-data (Ashraf, 2013).

This actually happens due to lack of knowledge and understanding regarding the technology (Gallup Pakistan, 2018) and difference of culture as compared to the western world (Latif, Tunio, Pathan, Jianqiu, Ximei & Sadozai, 2018). Hence there is a significant need of research which may analyze the effect of big-data technology with reference to the business operating in. Pakistan. Especially in the context of marketing as big-data has the ability to transform marketing related functions in near future (Rajeb et al., 2020).

3 Theoretical Framework and Delimitations

Application of big data are linked with top ten industries including wholesale and retail (Rajeb et al., 2020). Among these retailing and wholesaling are treated as part of daily life (Le & Liaw, 2017). On other hand use of information technology might significantly aids in the profitability of retail sector (Ali Subzwari and Tariq, 2016) though there was lack of evidence regarding the application of big-data in marketing functions but have the ability to transform market and interrelated functions in near future (Rajeb et al., 2020).

Hence study will figure out the impact of big-data on marketing functions of retail sector which is associated with exponential growth (Fazl-e-Haider, 2018), and major elements are pricing & assortment (Aktas & Meng, 2017). However, to deal with big-data analytics there is a requirement of technical mix of creativity and analytical skills especially in the context of marketing functions (Glass & Callahan, 2014). Though this is not possible without conducting quantitative analysis to forecast these measures in more effective manner (Aktas & Meng, 2017). Therefore, this study will use availability of skilled data scientist as the moderator with assortment of retail sector of Pakistan.

The research has only one IV (big-data) and one DV (assortment strategies) due to its conduction in Pakistan where there is lacking of understanding of the concept of big-data (Gallup Pakistan, 2018). However, most of the retailers are dealing in fast moving consumer goods (IBM, 2018). and online retailing is still infancy and in initial stages (Ali et a., 2016). Therefore, in order to provide significant impact to economy it is valid to measure it efficiency in terms of online retail stores.

4 Research Questions:

RQ1: Whether Big-Data is applicable to developing countries like Pakistan?

RQ2: Whether knowledge of IT experts really signifies the use of Big-Data?

RQ3: Whether Big-Data is applicable to retail sector of Pakistan?

5 Significance

Rajeb et al. (2020) indicated that use of big-data analytics is viable in top ten industries all over the globe. Although there is lack of understanding regarding the concept and application in the context of Pakistan (Gallop Pakistan, 2018). However other countries are dominant in taking advantage of big-data application which are even observable in leading industries like banking, securities, media, whole selling and retaining etc (Aktas & Meng, 2017).

However whole selling and retailing are treated as the part of daily life Li and Liaw (2017) and therefore conducting quantitative study on implication of big-data on retailing industry must be termed pervasive in nature. The claim is valid as study will fulfill need of big-data linkage with marketing functions as highlighted by Rajeb et al (2020).

Moreover, study will also help in removing the lacking of studies linked with big-data from emerging markets (New Desk, 2020), hence the study is beneficial for academia as well a pragmatic world and must be termed as pervasive.

6 Literature Review

One of the latest studies of 2020 indicated that still a significant lacking of studies which highlights the most appropriate use of big data technology (Surbakti et al., 2020). The terminology is actually associated with large data sets i.e. of terabyte and exabyte (Rajeb, Rajeb & Keogh, 2020). However, there is also a significant probability for the increase of organizational data (Zanini & Dhawan, 2015) thus several studies which tries to explore challenges pertaining to big-data and most of these indicated further investigation of benefits which organizations may urge through using big-data (Surbakti et al., 2020). There was lack of evidence for the relationship of big-data and marketing functions but big-data provide companies way to create edge and value (Zeng & Glaister, 2018)

However recent work of Rajeb et al (2020) highlighted that big-data is playing major role in improvement of functions like Marketing, supply chain as well as in process of decision making. This form of data is also applicable social media to observe thoughts, views and reviews of clients and this field is termed as sentiment analysis with purpose to device effective marketing strategies (Zanini & Dhawan, 2015). Though as mentioned earlier there is lack of evidence for the relationship of big-data and marketing functions (Keogh, 2019) but big-data has significant importance in retail business (Bradlow, Gangwar, Kopalle & Voleti, 2017).

Therefore, the study is optimal to relate study with sentiment analysis as a form of big-data analytics so to collect, organize and analyze large data sets in order to uncover new patterns and large data sets (Vijayarani & Sharmila, 2016)

7 Big-Data & Assortment

Product attributes as well as level of product attributes are the major components of product information at retail stores (Bradlow et al., 2017). Products attributes are potent predictor of customer's satisfaction as through effective analysis firm may decrease the percentage of stock

out products (Matsa, 2011) & may also increase total number of items (Briesch, Chintagunta & Fox, 2009). On the other side retailer are in need of right data to gain information about customer insights and value might be gauged through testing the data in terms of transaction log, loyalty information, pricing strategies and campaign results (Howe, 2014). Analysis of purchase records by Gielens (2014) was based on six retail stores for thousands of customers and resulted in effective negotiation with suppliers.

Similarly, Kumar and Kapoor (2014) indicated buying behavior has been measures on the bases of three characteristics i.e. quantity of purchase, frequency of purchase & preferred location of retail outlet. Thus, these types of information can be used by retailers in order to optimize practices of inventory management and may also deal effectively with change in demands (Ridge, Johnston & O'Donovan, 2016). Big-Data is also capable of assisting retailers in recognition of anomalies through observing practices and patterns looks unusual (Kaur & Jagdev, 2017).

Although several retailers are failed to incorporate effective data collection techniques due to privacy issues but these types of issues must not be treated as reason for non-incorporation of better data driven decisions. On the other hand, if retailers became able to cope up with the opportunity of associating data with business analytics might be able to track entire purchase journey (Shankar, 2019).

Though this might only be possible if data scientists have sufficient skill inventory to deal with issues of knowledge extraction (Dolezel & McLeod, 2019). Hence, there is severe lacking of inhouse data management specialists due to uncertainty of return on investment from big-data analytics (Iqbal, Kazmi, Manzoor, Soomrani, Butt & Shaikh, 2018). In large firms the functions is carried out through using experts of different IT fields. However, small and medium sized enterprises (SME) faces lot of difficulties in managing the same as they need cross sectional experts. Contrary to this unavailability of experts of the field & high cost of staffing are also resulting in the shortage of in-house data management specialists (Iqbal et al., 2018).

8 Research Methodology

Research methodology is a generic logic used in devising research although methods are the specific set of strategies and procedure used in the process of analysis. Research methodology is also supplemented with epistemological or ontological assumptions (Long, 2014). However, if we use facts as a truth then epistemology would help us in posing factual questions i.e. how do we know the truth? and what counts as knowledge? etc.

Although ontology is a branch of philosophy through which we evaluate things which making sense are real or not and through this researcher tries to explore (Kivunja & Kuyini, 2017). Hence in association with parameters of Kivunja and Kuyini (2017) the philosophy associated with this study is epistemology. Reason being the study aims to identify linkage of big-data in terms of retail industry of Pakistan in order to seek knowledge rather than to check the nature of reality. For discussing further on research methodology, one has to discuss about two of its parts i.e. Research Design & Sampling Design as indicated by Sileyew (2019).

A Research Design

Research design the part of research methodology which is used to provide answer to those questions in which researchers are interested (Oso & Onen, 2009). On the other hand, Mkansi and Acheampong (2012) indicated that most of the research terminologies which are used in academic research are consistent with the work done by Saunders Lewis and Thornhill (2009). Therefore, most of the terminologies used in research design are consistent with Saunders et al., (2009). Therefore, the philosophical stance is required for the determination of most adequate method for collection and analysis (Zukauskas, Vveinhardt & Andriukaiteiene, 2018) which is applicable to qualitative as well as quantitative research design (Saunders, Lewis & Thornhill 2015). Though the method of data collection is mono-method i.e. quantities (Saunders et al., 2015) and unit of analysis is individual while time horizon was cross-sectional (Sekaran & Bougie, 2016).

B Sampling Design

The reason due to which one may realize the base to include specific units' items in research are termed as sampling design (Mugenda, 20003).

However, in order to decrease overall cost associated with data collection items which are included in sample must able to assure research objectives (Leedy & Ormrod, 2005). Therefore, this study uses IT experts from online retail sector as the elements of sampling regardless of the infancy of online retailing (Ali et al., 2016). However, most of the retailers in Pakistan do not preferred online method (IBM, 2018) hence there are few IT experts available who are associated with online retail business. Therefore, the sample size for this study is 50 respondents which are justifiable as the research is linked with theory building approach due to its linkage with big data application on online retail sector. However, the data has been collected through using quota sampling so to deal with slow response rate and also to excessive sampling (Yang & Banamah, 2014).

C Questionnaire

Data has been done through of closed ended questionnaire which is adapted from Le and Liaw (2017) and Seetharaman Niranjana Tandon and Saravanan (2016). Moreover, measures indicated by Aktas and Meng (2017) and Valchanov (2017) has also been added to the questionnaire. Last but not the least availability of data scientist is a form of construct which has been developed on the bases of characteristics of data scientist and their use indicated by De Mauro Greco Grimaldi and Nobili (2016). Thus, through these considering systematic pattern indicated by prior studies the questionnaire was formulated.

D Statistical Testing and Analysis

Now a days SMART-PLS is treated as the better option for statistical testing especially in the studies of management sciences (Benitez, Henseler, Castillo & Schuberth, 2020).

Software use two types of models i.e. reflective and formative for the analysis of data (Benitez, Henseler, Castillo & Schuberth, 2020). The model of this study is a form of reflective-measurement model and therefore analysis will follow the indications made by Afthanorhan (2014) and Benitez et al. (2020). Hence the initial tables included in the section will demonstrate about descriptive statistical measures and the later one will indicate inferential statistical measures.

E Outer Loadings

Table 1: Outer Loading

	Assortment	Big-Data	Moderating Effect 1	Skilled Data Scientists
A1	0.743			
A2	0.893			
A3	0.831			
A4	0.891			
A5	0.704			
BD1		0.866		
BD2		0.877		
BD3		0.934		
BD4		0.939		
BD5		0.918		
Big-Data * Skilled Data Scientists			1.166	
SD1				0.686
SD2				0.879
SD3				0.782
SD4				0.876
SD5				0.721

The purpose of table 1 is to indicate outer loading for each element in order to highlight reliability of the construct associated with big-data and assortment strategies of online retailers. The minimum value which may validate the selection of any element in the construct is 0.60 (Afthanorhan, 2014).

Though the optimal range of value for outer loading starts from 0.708 (Sarstedt, Ringle & Mena, 2012) and if the value of element is lesser than 0.708 and also causing in decrease of overall reliability of construct then it must be deleted (Hair Jr, Hult, Ringle & Sarstedt, 2016). Hence all the elements include in table 1 seems to be effective as the minimal value of any element included in table 1 is 0.686 which is higher than the conidian indicated by Afthanorhan (2014).

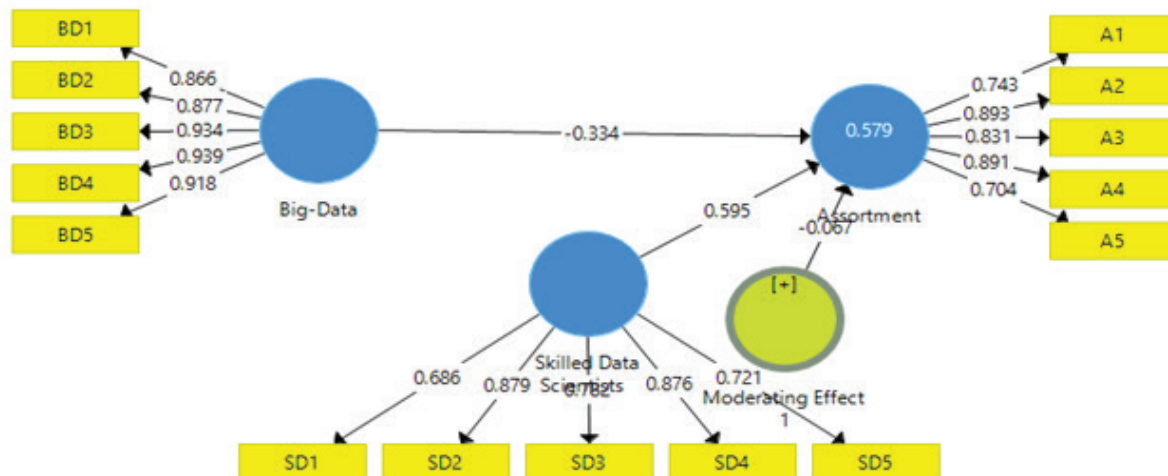


Figure 1: Highlighting p-values to show confirmatory factor analysis (CFA) for the model of big data on assortment strategies

Table 2: Predictive Accuracy (Quality Criteria)

R Square		
	R Square	R Square Adjusted
Assortment	0.579	0.574

Table 2 is termed as quality criteria or predictive accuracy and the use of the tool is to indicate the degree of explained variance caused by independent variable (Benitez et al., 2020).

However the method of evaluation of R is same as that of regression (Andreev, Heart, Moaz & Pliskin, 2009) the minimal value for the tool is 0.26 & of 0.5 & 0.75 are treated as moderate and extensive (Cheah, Memon, Chuah, Ting & Ramayah, 2018). Therefore in accordance with these measure the value of R is treated as moderate as the value for the tool is 0.574 which is lower than the standard of extensive fit.

F Construct Reliability and Validity

Table 3: Construct Reliability & Convergent Validity

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Assortment	0.871	0.873	0.908	0.666
Big-Data	0.946	0.950	0.959	0.823
Moderating Effect 1	1.000	1.000	1.000	1.000
Skilled Data Scientists	0.849	0.853	0.893	0.628

Table 3 is indicating construct reliability & convergent validity (Ab Hamid, Sami & Sidek, 2017 & Sijtsma, 2009 a&b).

Table 3 is used to indicate convergent validity that indicates how well parameters associated with one latent variable measures the same construct (Benitez, Henseler, Castillo & Schuberth, 2020). Though table also contains two reliability measures i.e. Cronbach's Alpha (α) and Dillon-Goldstein rho in order to highlight construct reliability through Cronbach's Alpha (α), Dillon-Goldstein's rho & AVE (Sijtsma, 2009a&b). Therefore, in the light of these measures the model is effective to ensure construct reliability and convergent validity. These statements are valid as the values of α and Dillon-Goldstein rho is more than 0.7 to ensure construct reliability while values of composite reliability and AVE are more than 0.5.

Table 4: Discriminant Validity through Heterotrait-Monotrait Ratio (HTMT)

	Assortment	Big-Data	Moderating Effect 1	Skilled Data Scientists
Assortment				
Big-Data	0.550			
Moderating Effect 1	0.078	0.133		
Skilled Data Scientists	0.779	0.288	0.140	

Table 4 is used to indicate discriminant validity to show heterogeneousness among different variable of same research model (Cheung & Lee, 2010). Reason for using HTMT as the way to show discriminant validity is the ratio is treated as most effective for highlighting that (Benitez et al., 2020). However, the peak value which is tolerate able in the case of HTMT is 0.85 (Hair Jr, Sarstedt, Ringle & Gudergan, 2017) and any value greater than this is ineffective for highlighting the discriminant validity.

G Mean, STDEV, T-Values, P-Values

Table 5: Path Coefficient

	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (O/STDEV)	P Values
Big-Data -> Assortment	-0.334	-0.334	0.053	6.290	0.000
Moderating Effect 1 -> Assortment	-0.067	-0.066	0.033	2.062	0.040
Skilled Data Scientists -> Assortment	0.595	0.594	0.054	11.114	0.000

Table 5 is the table which is used to highlight inferential measures of big-data on the assortment strategies of e-retailers. Inferential stats is one of the major part of the measurement models used in SMART-PLS (Hair, Risher, Sarstedt & Ringle 2019). Software does this through t-statistics (Durate & Amaro, 2018) and p-values (Kock & Hadaya, 2018) in order to indicate relationship of variables inferentially.

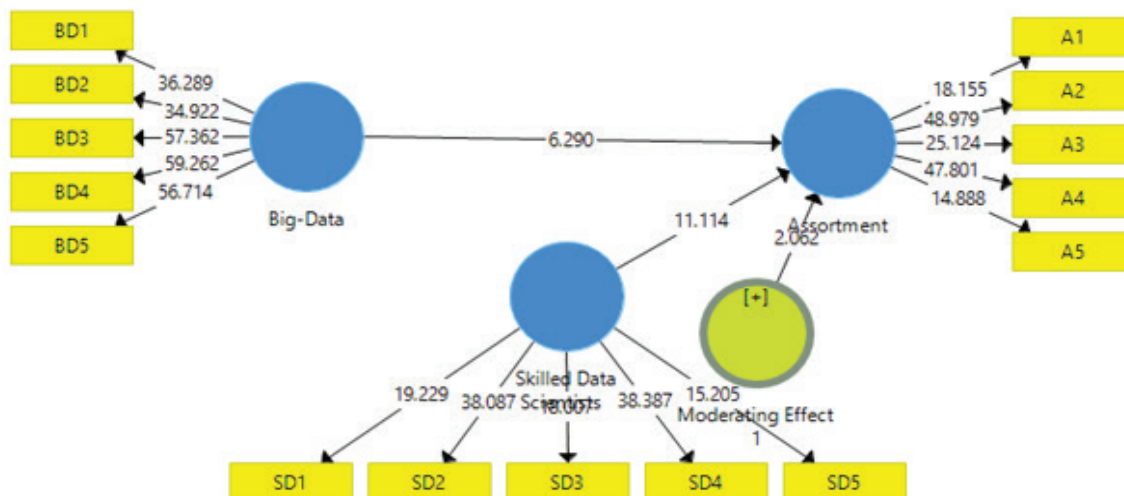


Figure 2: Path Coefficients and regressions weights for the construct of Big-Data on Assortment Strategies of Online Retailers

The benchmark value for t-statistics is 1.97 (Hair, Ringle & Sarstedt (2011) and for p-values the cutoff value is 0.05 above which there is no effect of variable on another (Kock & Hadaya, 2018). The minimum t-value required to indicate relationship between variables of the construct is 1.97 (Hair et al., 2011). Therefore in the light of these parameters it is legitimate to indicate that big-data is perceived as an effective tool for the optimization of assortment of e-retailers and skills of data scientist are also perceived as crucial in optimization of assortment practices.

Although the abilities of data scientist working with e-retailers are not adequate to affect assortment practices at e-retailers. Thus the moderation of skilled data scientist are diminishing the effect of big-dat technology on assortment practices of e-retailres which can be I observed through the increase of p-value in case of moderation of skilled data scientist

9 Discussion And Managerial Implications:

Study highlighted that the indications made by prior studies like Rajeb et al (2020) are appropriate to highlight significant impact of bi-data on betterment of functions like marketing, & supply chain etc. Moreover the study is based on e-retailers thus positive relation of big-data on assortment also indicates that big-data is also applicable to social media. Hence optimal to observe thoughts views and reviews of clients as indicated by Zanini & Dhawan (2015). Thus also justifies the indication of Bradlow wt al. (2017) that big-data is perceived as important tool for the retail business & findings of the study opposes the Keogh (2019). This is legitimate as assortment is important tool for retailers as well as in the field of marketing.

Hence legitimate to declare big-data is a potent tool for the analysis of large data sets (Vijayarani et al., 2016) which also required in retail business to assess right information so to predict customers. Furthermore the moderation of skilled data scientist actually diminishes the impact of big-data from assortment strategies of online retailers therefore also linked with Shankar et al. (2019) and Iqbal et al (2018).

Hence legitimate to declare the ineffective skill inventory of data scientist might be due to imbalance of payments and returns on big-data technology.

10 Area For Future Research

This research is based on the analysis of perceived uses of big-data and results achieved are only from the IT managers and experts associated with online retail sector. However most of the retailers preferred bricks and mortars as business model therefore studies would become legitimate if able to relate big-data with the operations and practices of physical retail businesses. Moreover future studies might try to explore the difference in big-data practices of those retailers which prefer bricks and clicks and flips and clicks in order to check the application of big-data in hybrid form and online form of retailing.

References

- [1] Ab Hamid, M. R., Sami, W., & Sidek, M. M. (2017, September). Discriminant validity assessment: Use of Fornell & Larcker criterion versus HTMT criterion. In *Journal of Physics: Conference Series* (Vol. 890, No. 1, p. 012163). IOP Publishing
- [2] Afthanorhan, W. M. A. B. W. (2014). Hierarchical component using reflective-formative measurement model in partial least square structural equation modeling (PLS-Sem). *International Journal of Mathematics*, 2(2), 33-49
- [3] Aktas, E., & Meng, Y. (2017). An Exploration of Big Data Practices in Retail Sector. *Logistics*, 1(2), 12
- [4] Ali, R., Subzwari, M., & Tariq, S. (2016). Impact of Information Technology on Retail Sector in Pakistan. *KASBIT Journal of Management & Social Science*, 9(1), 63-93.
- [5] Andreev, P., Heart, T., Maoz, H., & Pliskin, N. (2009). Validating formative partial least squares (PLS) models: methodological review and empirical illustration. *ICIS 2009 proceedings*, 193
- [6] Ashraf, S., (2013), "Can we Optimize Pakistan via Big Data?", Tech Juice, <https://www.techjuice.pk/can-optimize-pakistan-via-big-data/>
- [7] Baker, W., Kiewell, D., & Winkler, G. (2014). Using big data to make better pricing decisions. McKinsey Analysis. McKinsey & Company, <https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions#>
- [8] Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information & Management*, 57(2), 103168

- [9] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8.
- [10] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679
- [11] Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1), 79-95
- [12] Briesch, R. A., Chintagunta, P. K., & Fox, E. J. (2009). How does assortment affect grocery store choice? *Journal of Marketing research*, 46(2), 176-189
- [13] Brown, B. Bughin, J., Byers, A.H.; Chui, M.; Dobbs, R.; Manyika, J. and Roxburgh, C., (2011), "Big Data: The Next Frontier for Innovation, Competition, and Productivity", Technical Report for Mckinsey& Company, Washington, DC, USA
- [14] Cheah, J. H., Memon, M. A., Chuah, F., Ting, H., & Ramayah, T. (2018). Assessing reflective models in marketing research: A comparison between PLS and PLSC estimates. *International Journal of Business and Society*, 19(1), 139-163
- [15] Chong, A. Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via BigData: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, 55(17), 5142-5156, doi: 10.1080/00207543.2015.1066519
- [16] Clark, R., & Vincent, N. (2012). Capacity-contingent pricing and competition in the airline industry. *Journal of Air Transport Management*, 24, 7-11
- [17] Cohen, L., Manion, L., & Morrison, K. (2007). Observation. *Research methods in education*, 6, 396-412
- [18] Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and aging*, 21(2), 333.
- [19] Danah, B., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662-679.
- [20] De Mauro, A., Greco, M., Grimaldi, M., & Nobili, G. (2016). Beyond data scientists: a review of big data skills and job families. *Proceedings of IFKAD*, 1844-1857
- [21] Ducange, P., Pecori, R., & Mezzina, P. (2018). A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing*, 22(1), 325-342.
- [22] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- [23] Fazl-e-Haider, S. (2018 February 19). Booming Growth in Pakistan's Retail Sector. *Economist*, <http://www.pakistaneconomist.com/2018/02/19/booming-growth-pakistans-retail-sector/>

- [24] Friedrich, O. Stoler, P. Moritz, M. & Nash, J. N. (1983). Machine of the year: The computer moves. *Time Magazine*, 121(1). 14–28
- [23] Gallup Pakistan, (2018), “Big Data Analysis Reports”. Retrieved from <http://gallup.com.pk/polls/gallup-history-project/big-data-analysis/>
- [24] Gielens, K., Gijsbrechts, E., & Dekimpe, M. G. (2014). Gains and losses of exclusivity in grocery retailing. *International Journal of Research in Marketing*, 31(3), 239-252
- [25] Glass, R., & Callahan, S. (2014). *The Big Data-driven business: How to use big data to win customers, beat competitors, and boost profits*. John Wiley & Sons.
- [26] Grewal, D., & Levy, M. (2007). Retailing research: Past, present, and future. *Journal of retailing*, 83(4), 447-464.
- [27] Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, 2(163-194), 105.
- [28] Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications
- [29] Hair Jr, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2017). *Advanced issues in partial least squares structural equation modeling*. sage publications
- [30] Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long range planning*, 46(1-2), 1-12
- [31] Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2-24
- [32] Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the academy of marketing science*, 40(3), 414-433
- [33] Hair, J.F., Ringle, C.M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, 19(2), 139-152
- [34] Hajirahimova, M. S., & Aliyeva, A. S. (2017a). About Big Data Measurement Methodologies and Indicators. *International Journal of Modern Education and Computer Science*, 9(10), 1
- [35] Hajirahimova, M. S., & Aliyeva, A. S. (2017b). Big Data Initiatives to Developed Countries. *Problems of information society*, 1, 10-19, doi: 10.25045/jpis.v08.i1.02
- [36] Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing*. Emerald Group Publishing Limited.
- [37] Howe, K. (2014). Beyond big data: How next-generation shopper analytics and the internet of everything transform the retail business. *Cisco*, 1-10
- [38] IBM Institute of Business Value IBM, (2018), “Analytics: The real-world use of big data in retail”, How innovative retailers extract value from uncertain data, <https://www-935.ibm.com/services/us/gbs/thoughtleadership/big-data-retail/>

- [39] IDC, (2014), "Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives", Retrieved from <https://www.emc.com/leadership/digital-universe/2014iview/executivesummary.html>
- [40] Integreon Insight (2012). "Big just got bigger". Grail Research, 1-17 (Retrieved from http://www.integreon.com/pdf/Blog/Grail-Research-Big-Data-Just-Got-Bigger_232.pdf)
- [41] Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. More than just convenient: The scientific merits of homogeneous convenience samples. *Monographs of the Society for Research in Child Development*, 82(2), 13-30
- [42] Jun, S. P., & Park, D. H. (2017). Visualization of brand positioning based on consumer web search information: Using social network analysis. *Internet Research*, 27(2), 381-407.
- [43] Kaur, R., & Jagdev, G. (2017). Big Data in retail sector-an evolution that turned to a revolution. *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)*, 4(4), 43-52
- [44] Khan, A., (2017, July 31), "Trading data", Retrieved from <https://www.thenews.com.pk/magazine/money-matters/194614-Trading-data>
- [45] Khan, M. W., Khan, M. A., Alam, M., & Ali, W. (2018). Impact of Big Data over Telecom Industry. *Pakistan Journal of Engineering, Technology & Science*, 6(2), 116-126, <http://dx.doi.org/10.22555/pjets.v6i2.1958>
- [46] Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, 6(5), 26-41
- [47] Kshetri, N. (2016). *Big data's big potential in developing economies: impact on agriculture, health and environmental security*. CABI. Wallingford, doi10.1079/9781780648682.0000, ISBN 9781780648682
- [48] Latif, Z., Tunio, M. Z., Pathan, Z. H., Jianqiu, Z., Ximei, L., & Sadozai, S. K. (2018, March). A review of policies concerning development of big data industry in Pakistan: Subtitle: Development of big data industry in Pakistan. In *Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018 International Conference on (pp. 1-5). IEEE.
- [49] Le, T. M., & Liaw, S. Y. (2017). Effects of Pros and Cons of Applying Big Data Analytics to Consumers' Responses in an E-Commerce Context. *Sustainability*, 9(5), 798.
- [50] Leedy, P. D., & Ormrod, J. E. (2005). *Practical Research Planning and Design*. New Jersey: Pearson Merrill Prentice Hall.
- [51] Lycett, M. (2013). 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, 22(4), 381-386
- [52] Maheshwari, A. (2014). *Business Intelligence and Data Mining*. Business Expert Press.
- [53] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company, Washington, USA,

- [54] Matsa, D. A. (2011) Competition and product quality in the supermarket industry, *The Quarterly Journal of Economics*, 126, 1539–91. doi:10.1093/qje/qjr031
- [55] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-68.
- [56] Mkansi, M., & Acheampong, E. A. (2012). Research philosophy debates and classifications: students' dilemma. *Electronic journal of business research methods*, 10(2), 132-140
- [57] Mugenda, A. (2003). *Research methods Quantitative and qualitative approaches* by Mugenda. Nairobi, Kenya.
- [58] Mukherjee, S., & Shaw, R. (2016). Big Data–Concepts, Applications, Challenges and Future Scope. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(2), 66-74.
- [59] New Desk. (2020, May 13). Profit, <https://profit.pakistantoday.com.pk/2020/05/13/pakistan-retains-its-place-in-msci-emerging-markets-index/>
- [60] Oracle (2012). Big data for the enterprise. Oracle White Paper, 1–14, <http://www.oracle.com/us/products/database/big-data-forenterprise-519135.pdf>.
- [61] Oso, W. Y., & Onen, D. (2009). *A general guide to writing research proposal and report*. Jomo Kenyatta Foundation.
- [62] Pantano, E., Giglio, S., & Dennis, C. (2019). Making sense of consumers' tweets. *International Journal of Retail & Distribution Management*, 47(9), 915-927.
- [63] Pathirage, C. P., Amaratunga, R. D. G., & Haigh, R. P. (2008). The role of philosophical context in the development of research methodology and theory. *The Built and Human Environment Review*, 1(1), 1-10
- [64] Pecori, R. (2016). S-Kademlia: A trust and reputation method to mitigate a Sybil attack in Kademlia. *Computer Networks*, 94, 205-218.
- [65] Ravand, H., & Baghaei, P. (2016). Partial least squares structural equation modeling with R. *Practical Assessment, Research, and Evaluation*, 21(1), 1-16, <https://doi.org/10.7275/d2fa-qv48>
- [66] Rejeb, A., Rejeb, K., & Keogh, J. G. (2020). Potential of Big Data for Marketing: A Literature Review. *Management Research and Practice*, 12(3), 60-73
- [67] Ridge, M., Johnston, K. A., & O'Donovan, B. (2015). The use of big data analytics in the retail industries in South Africa. *African Journal of Business Management*, 9(19), 688-703
- [68] Salvador, A. B., & Ikeda, A. A. (2014). Big data usage in the marketing information system. *Journal of Data Analysis and Information Processing*
- [69] Santoro, G., Fiano, F., Bertoldi, B., & Ciampi, F. (2019). Big data for business management in the retail industry. *Management Decision*, 57(8), 980-1992
- [70] Saunders, M. N. K., Lewis, P., Thornhill, A., & Bristow, A. (2015). Understanding research philosophies and approaches: Research methods for business students

- [71] Saunders, M., Lewis, P. & Thornhill, A. (2007). Research methods. Business Students 4th edition Pearson Education Limited, England
- [72] Schultz, J., (2017, October 10), "How Much Data is Created on the Internet Each Day?" <https://blog.microfocus.com/how-much-data-is-created-onthe-internet-each-day>
- [73] Seetharaman, A., Niranjan, I., Tandon, V. and Saravanan, , A. S., (2016), "Impact of Big Data on the Retail Industry" Journal of Corporate Ownership & Control, 14(1), 506-518
- [74] Sekaran, U. and Bougie, R., (2016), "Research Methods For Business: A Skill Building Approach", John Wiley & Sons, 1-448, ISBN 1119165555, 9781119165552
- [75] Shankar, V. (2019). Big Data and Analytics in Retailing. NIM Marketing Intelligence Review, 11(1), 36-40
- [76] Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107
- [77] Sijtsma, K. (2009). Over misverstanden rond Cronbachs alfa en de wenselijkheid van alternatieven. Psycholoog, 44(11), 561
- [78] Sileyew, K. J. (2019). Research Design and Methodology. In Text Mining-Analysis, Programming and Application. Intech Open, <https://www.intechopen.com/books/cyberspace/research-design-and-methodology>
- [79] Spiess, J., T'Joens, Y., Dragnea, R., Spencer, P., & Philippart, L. (2014). Using big data to improve customer experience and business performance. Bell labs technical journal, 18(4), 3-17.
- [80] State Bank of Pakistan, (2014), "Supermarkets and Retail Shops", Research Report on 'Supermarkets and Retail Shops' Segment", http://www.sbp.org.pk/departments/ihfd/Sub_Segment%20Booklets/Supermarkets%20and%20Retail%20Shops.pdf
- [81] Stoicescu, C. (2016). Big Data, the perfect instrument to study today's consumer behavior. Database Syst. J, 6, 28-42.
- [82] Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. Information & Management, 57(1), 103146
- [83] Tankard, C. (2012). Big Data Security. Network Security Newsletter, Elsevier, ISSN 1353-4858
- [84] Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International journal of medical education, 2, 53-55, doi 10.5116/ijme.4dfb.8dfd
- [85] Thakuriah, P. V., Tilahun, N. Y., & Zellner, M. (2017). Big data and urban informatics: innovations and challenges to urban planning and knowledge discovery. In Seeing cities through big data (pp. 11-45). Springer, Cham.
- [86] Thau, B., (2016, March 2), "Retail pricing strategies getting a makeover from data analytics", IBM Big Data and Analytics Hub, <https://www.ibmbigdatahub.com/blog/retail-pricing-strategies-getting-makeover-data-analytics>
- [87] Valchanov, I. (2017). Is data science really a rising career? <https://www.quora.com/Is-data-science-really-a-rising-career/answer/IliyaValchanov>

- [88] Vargas-Sánchez, A., do Valle, P. O., da Costa Mendes, J., & Silva, J. A. (2015). Residents' attitude and level of destination development: An international comparison. *Tourism Management*, 48, 199-210.
- [89] Vijayarani, S., & Sharmila, S. (2016, August). Comparative analysis of association rule mining algorithms. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (Vol. 3, pp. 1-6). IEEE
- [90] Vikas, D. & Nadir, Z. (2014). Big data and social media analytics. A Cambridge Assessment publication, 18, 36-41, www.cambridgeassessment.org.uk/research-matters/
- [91] Voleti, S., Kopalle, P.K., & Ghosh, P. (2015). An inter-product competition model incorporating branding hierarchy and product similarities using store-level data. *Management Science*, 61(11), 2720-2738.
- [92] Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6), 97-121.
- [93] Werner, K., Lerzan, A., Yakov, B., Kristina, H., Sertan, K., Francisco, V. O., Marianna S., David, D. and Babis, T., (2017) "Customer engagement in a Big Data world", *Journal of Services Marketing*, 31(2), 161-17
- [94] Zanini, N., & Dhawan, V. (2015). Text Mining: An introduction to theory and some applications. *Research Matters*, 19, 38-45
- [95] Zeng, J., & Glaister, K. W. (2018). Value creation from big data: Looking inside the black box. *Strategic Organization*, 16(2), 105-140
- [96] Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media. Erevelles, S., Nobuyuki
- [97] Žukauskas, P., Vveinhardt, J., & Andriukaitienė, R. (2018). Philosophy and paradigm of scientific research. *Management Culture and Corporate Social Responsibility*, 121.

Robust Food Supply Chain Traceability System based on HACCP using Federated Blockchain

Muhammad Danish¹

Muhammad Shahwaiz Hasan²

Abstract

In this fast growing world, everything needs to change and upgrade with the technology to survive in this rapidly changing environment. The Food Supply Chain system is one of them. In traditional supply chain management functions, it is hard to manage the flow of traceability information of products timely and effectively among upstream to downstream stakeholders. This gap in the flow of information and untrusted traceability of product causes to generate food contamination hazards and also increasing the network of modern FSC systems is still challenging to provide quality, privacy, and integrity of traceability mechanism. HACCP and Blockchain-based FSC provide various functions to enhance the FSC process such as traceability, transparency, security among all stakeholders. In these functions, some most important functions of the Food Supply Chain system are traceability, amount of data sharing, and privacy of data of products among stakeholders.

In this paper, we are going to provide an Enhanced Food Supply Chain Traceability system based on Federated Blockchain with HACCP using smart contracts. This model will manage by pre-selected leaders' nodes. These leaders' nodes have the right to control the action of middle nodes and lower nodes. The amount of data sharing can also be controlled by the leaders' nodes. So, any particular transaction is restricted to perform by middle and lower nodes because the Federated Blockchain is partially decentralized. This model overcomes the issue of privacy concern which remains unsolved by blockchain-based FSC and makes a better and enhanced Federated Blockchain-based FSC traceability system.

Keyword: Food supply chain, Federated blockchain, smart contract, traceability, HACCP, security

1 Introduction

In the industry line, the efficient work and smooth flow of the operations is a key strategy of every industry. Companies require reliable and real-time knowledge about production, on-time transportation of goods, to make correct supply chain decisions.

The supply chain-related systems play an important role in multiple factors of economic globalization. The supply chain-related system is used to minimize the complexity of the goods transportation cycle from origin to consumer, make industrial processes easier, and protect them from the evil habit intrusion and frauds in the system. But this system has some limitations when emerging decentralized ledger technology. However, mostly Present IoT-based traceability structures for food supply chains are based on centralized networks, but this creates a major space for unsolved issues and major concerns.

¹Karachi Institute of Economics & Technology, Karachi, Pakistan | danish2484@gmail.com

²Karachi Institute of Economics & Technology, Karachi, Pakistan | shahwaizhasan94@gmail.com

a) Problems in the supply chain system

As increasing technology in business processes, the supply chain process becomes more complex due to evolving customer requirements, competitors, geographically separated locations, operation, and advancement in the new business processes like e-commerce. In recent past years, technology likes e-commerce and mobile gadgets change the daily lives of people regarding purchasing from home, this is an entirely different way of business. So there is so much complexity arises in the traditional supply chain process. There are many issues to serve customers or other stakeholders with valuable transparency of provenance of goods. The traditional supply chain fails to provide risk management for cost reduction and market requirement changes. We summarize some of the major issues in the current supply chain system.

1 *Lack in traceability process*

Traceability is the major problem in the supply chain regarding customer service and planning in business operation. When limited trust exists among the participants, traceability is hard to implement a centralized system in interconnected network.

2 *Stakeholder distrust*

In supply chain management, trust is the main factor to build an effective supply chain network. The supply chain system does not provide a trustworthy system itself to strengthen the supply chain network so that distrust appears among the stakeholders and to make it a trustworthy mechanism, stakeholders take third parties services to verify the process and transactions, which is cause to increase the operational cost and reduce the efficiency of the process.

3 *Ambiguous transparency*

Transparency is the major factor of any business, a transparent supply chain cause to increase the integrity of the product and build trust among stakeholders. But mostly supply chain networks provide minimum transparency. So, sometimes very important information lost when data share from one stakeholder to another.

4 *Traditional methods of data sharing*

Traditional supply chain networks shared their data among other organization through paper work. Most of the time these paper documents travel with the goods, inefficient data sharing and misguided paper document cause to reach late shipment at their final destination.

5 *Compliance challenges*

With the passage of time business regulatory standards become strict to provide safe services to the customers. In the current supply chain system, it is hard to gather information from many stakeholders and compiles with new standards.

A Block Chain Tecnology

Blockchain technology is used to enhance customer services and increases the efficiency of the operations. It maintains an immutable and secure data ledger that is transparent and equally updates the information to each node present throughout the network. It is a very reputed technology adopted by many areas of economic and social systems. A blockchain-based network relies on distributed ledger technology, in which every transaction record in a block and each block linked to another block with a secure hash value of the previous block, transactional data become immutable because every transaction made by any node must accept and validate by the network consensus and will be updated in every node's ledger in the network. It is a synchronized decentralized network, which works without any central authority to validate the transactions.

There are two types of blockchain technology with respect to its transparency, security and verifiability. In which one is permissionless and the other one is permissioned blockchain. In permissionless blockchain anyone can join anonymously, it is also called the public blockchain. In this blockchain low confidence and trust exist among the network user. To reduce the fear of an untested network, miners are used to validating the transactions in the network.

In permissioned blockchain network, every user identifiable throughout the network and is part of a central agreement called consensus in blockchain technology. It is also called a private blockchain. No new user adds to this group (group is the collection of individual nodes who stores and access amount of information) until he grabs the majority voting of the group members to agree that this new user can be added to this group. These blockchain networks provide trust among the users and it is not required, costly miners (miners can be described as, who records the blockchain transaction and receive cost of this work).

a) Key characteristics of block chain technology Blockchain technology has many unique characteristics such as verifiability, transparency, and immutability of distributed ledger, some of the most important are as follows”

1 Efficient transaction recording

Blockchain provides a secure way to recording the transaction in efficient manner.

2 Distributed governance

Blockchain the network does not rely on any third party involvement or centralized authority. It has a distributed database that provides security and transparency. In the blockchain, when any transaction performs, the ledger update to all the participant present in the blockchain network.

3 Decentralized architecture

Blockchain used decentralized architecture to update the ledger that is why data stored at all nodes, due to this central infrastructural point failure is not possible so that an efficient and

robust network generate which have reliability, availability, and quality of data and information.

4 Data transparency

Blockchain technology provides high transparency to all stakeholders. All the actions made are available to every node to see it. That is why no chance to make an illegal transaction.

5 Immutable data

It is not allowed to change in any transaction which has already made and verifies by the consensus method of the network and has stored in the block.

6 Traceability

Traceability of the supply chain ensures environmental protection as we as backtracking of products demonstrated by a particular business case. Blockchain has record bulk of transaction data and keeps them in a network and the system also those transaction records are shared peer to peer to all the connected systems.

7 Enhanced data security

Blockchain technology provides a strong feature of security by using digital signature algorithms and cryptography. it ensures the security of data and builds confidence throughout the supply chain process. All data, that maintained in a blockchain relies on predefined consensus and required permission of the majority of network nodes present in the network. This public ledger is unchangeable in nature and provides audit-ability of all transactions performed. Thus the blockchain has the properties to enable decentralized traceability systems, and show all the steps when the transaction occurs in a supply chain.

a) Federated Block chain, a Future of Block chain.

A consortium, Federated and private blockchain are way similar in nature. In private or a permissioned blockchain every node has no access to make any functional behavior even if the node is authenticating the system. Only one person or a leader has determined permission to do anything in the system. This kind of blockchain is ideal for use by businesses that also have dealings with each other.

In Federated blockchain, only a few defined nodes are granted power. These blockchains are simpler and more elastic. A federated blockchain allows greater improvement and flexibility in the structure of blockchain systems. Organizations will now share information more easily and with less hesitation. Federated blockchain is built for particular groups or individuals. The assumption that unknown people are not permitted on the network, which eliminates the probability of 51 percent of peers [16].

Many of the industry claim that the federated blockchain provides higher speed, scalability, efficiency and overcomes

fraud, information breaching problems. The federated blockchain also has a great way to use

in the future, but the fact is that having just one kind of blockchain is not a wise idea, because each of these blockchains may offer more benefits depending upon the need and use in the industry they are used in.

8 Hyper ledger fabric

It is a private blockchain network which is open source and makes for enterprise applications. It was made up of Linux based foundation. It provides many features, like a public ledger, smart contract engines, and consensus protocols. These multi-functional features made up this to adopt widely in many business applications such as finance insurance supply chain health care and human resources.

B Smart Contracts

A smart contract is the set of actions that specify digitally with defined protocols. Smart contract term firstly proposed by Nick Szabo. This concept of smart contract introduced in the Ethereum blockchain network to provide verification and to improve contract performance. Before any transaction made in blockchain, the set of the contract defines the condition, obligation, rights, and concept between stakeholders. Set of predefine promises or contracts are stored and shared through- out the network and access to all node present in the network. Every transaction will perform under the net of a defined smart contract. Due to this more trust in the network developed which cause a reduction of risk, error, and fraud. Some of the smart contract benefits are:

1 Cost-saving

It cost saving due to reducing the time of processes and eliminate the intermediaries in the network.

2 Accurate:

It provides accurate and efficient mechanism of information storage through all agreement and defines conditions recorded in terms of computer code.

3 Speedy

It provides robust and efficient speed of transaction completion when define conditions are met with the contract.

4 Secure

Smart contracts are stored in the distributed ledger using an encryption mechanism to provide more security and distribute it to all nodes.

C Internet of Things (IOT)

Internet of things (IoT) is one of the popular technologies nowadays, it helps many industries to make their work easier. IoT is a device that is interconnected with digital objects with processes

and in the environment. You can gather knowledge, evaluate it, and take an intervention to support someone with a specific task or learn from a system by a unique digital identifier. These works are done without a human to human interaction and without a human to computer. Many of the researchers and developers are trying to implement many of the systems by adopting an IoT technology like RFID, and wireless network sensors or some open-source devices to monitor the real data related to the conditions in transporting of information and data.

1 IoT with digital Identifiers (RFID)

The digital Identifiers like RFID (Radio-Frequency Identification) tags, GPS, or a bar code. By these tags, assets are easily tracked and manageable. Many of the RFID systems are used in food-related companies to tracing food in whole supply chain management. It gives information about the producer, retailer, wholesaler, and consumer to the relevant management timely and effectively. This technology-related system is integrated with the supply chain systems to manage traceability.

D Hazard Analysis And Critrol Points (HAC- CAP)

The HACCP (Hazard analysis and critical control points) HACCP focused on risk control and avoidance, associated with food health. It easily connects with organization management like supply chain management (SCM) and food safety assurance. Fotopoulos et al. [19] have a literature review on food safety insurance schemes and the essential factors that influence the implementation of HACCP were reported. They reviewed 31 experiments in their study and found 32 variables that may influence the application of HACCP. By implementing HACCP in FSC with using (IoT), monitoring and traceability become much easy as compared with the traditional systems. HACCP method provides more efficiency and protection to all supply chain members and their related work.

In this section, we show an overview of the complete paper. In the first section, we highlight the overview of the technologies which is used in this paper or model. In the second problem-solving section, we provide number of research papers related to supply chain traceability systems with blockchain. We also proposed a food supply chain model that is beneficial in the entire factor when we use the food supply chain with blockchain for traceability.

Furthermore, our proposed model related to HACCP with Federated blockchain using IoT digital signature (RFID) for traceability solution for food supply chain management (FSCM).

2 Releted Work

In this modern era the supply chain-related process is more complex. So, to overcome this complexity many of the researchers and developers use blockchain technology to improve traceability, business model and goals, transform relationships, and enhance the performance of business activities.

In this section, we provide the focus path of research, regarding integrating the blockchain with a supply chain to slow down the complexity of the above factors. Furthermore, many of the

distinguished research articles and proposed models are selected and present below:

Supply chain provides various functions to enhance the business process such as traceability transparency security, among all stakeholders. Especially, in safety, a sensitive sector likes food, medicine. This lack of information flow and no trusted traceability of products cause to generate food contamination hazards, also increasing network of modern food supply chain (FSC) system is still a challenging to provide quality and integrity of traceability mechanism. So this paper provides an automated traceability system of food supply chain (FSC).

The food contamination scandals are a real hazard for public health. The food Supply chain helps to mitigate this issue but it has certain limitations that not fully resolve this issue.

This paper, describe the issues which is relevant to the implementation of blockchain technology in the FSC and analyze its opportunities to increase the safety of food and its waste reduction. To overcome this issue the retail giant Wal- Mart integrates a blockchain-based supply chain to make food safety easier, traceable timely effective, and transparent[1]. Due to this product visibility throughout the network is very difficult and traceability compromise. So, traditional methods of data sharing in the business process are very costly and unacceptable as compared to innovative supply-chain management. In paper [3], blockchain technology in supply chain functions to enhance their functions, like traceability, transparency, and integrity of the business process and discuss current situations, key features of blockchain, security, and challenges of deployment blockchain technology with the supply chain [3]. This paper proposed the COC (supply chain in the blockchain), the supply chain management system based on the hybrid DLT. It is a two-step block construction with a better model mechanism of security and performance and unauthorized access [4]. In paper [5] the researcher studies many of the past papers and identify that many authors and the projects have been done to improve the transparency, traceability, and other factors throughout the supply chain from suppliers. Information and Communication Technologies (ICT) based supply chain traceability (SCT) solutions have been implemented. GLOB- ID project using a cloud-based centralized system to merge legacy business information systems and increase SCT [5].In paper [7] the researcher analysis and supports the significance of this particular issue by addressing this topic's urgent needs in this authoritative journal. So, this paper provides the focus path to the OSCM researchers regarding integrating the blockchain. Furthermore, the twelve distinguished research article is selected and defined [7]. Nowadays, new technology is evolving day by day so the blockchain is one of them. But, this technology has some defective properties like scalability when we face a bulk amount of data in the real world. In paper [8] the researcher develops a food supply chain traceability network, based on HACCP (Hazard Identification and Essential Control Points), blockchain, and the Internet of Things, which will include an information network of accessibility, accountability, equality, efficiency, and protection for all the supply chain users also introduce a new concept BigchainDB (a database who offering a decentralization integrity and allow a large scale applications and many variety for supply chain) to fill the gap

in the decentralized systems at scale [8]. It seems that supply chains facing challenges in information sharing and trust. The main purpose of this research is to increase trust among all agricultural industries by using decentralized technology that is not dependent on trust [13].

At the above, we thoroughly discussed various functions of the food supply chain like safety, security, transparency, traceability mechanism, and authentication of traceability function by integrating blockchain technology. Although we have found many benefits of blockchain integration in the food supply chain. But some issues still need to work on it. Like safety and security of food, enhanced traceability mechanism which ensures to control, manage, prevent, mitigate, and contingency plan for food products from production to consumers hand in any means. Another issue that arises in the food supply chain due to the integration of blockchain technology is immutability and decentralized nature. Due to this nature, blockchain updates the ledger of the transaction to all nodes of the network. So, there are some food companies that do not want to share all the data and information through the network, somehow this information can be harmful to their goods or products. Competitors and other evil habits can use this information to fulfilling their unethical ways of benefits and for fraud.

HACCP is associated with food health and safety. It easily connects with the organization management system for food safety assurance. HACCP is a systematic approach to identify the hazards related to food safety and ensure to all the authorities, stakeholders, and customers about the quality work which is done in the food supply process. This assurance comes from the control point of the HACCP process. There are many control points in the HACCP for safety and to mitigate the risk factors in the food supply process. The control points are varying from food to food, as per food nature, the control points and monitoring plan changed. Because of this HACCP mechanism, the traceability of the system is far easier than before. After every event or process, the documentation and the data are updated in the system. So, in the end, the organization has a complete traceability record related to the current food process also, there are many types of research list which present the idea of HACCP with different technologies to make the traceability better and make the flow of the process fast to get effective output.

To overcome these issues we present a model in which we has use HACCP food safety management system with integration of federated blockchain using IoT (RFID) technology to enhance and improve the food safety mechanism throughout the network with greater transparency by providing leading and controlling rights of leader nodes of the federated blockchain network. Also remove the access of unwanted persons, who can capture important information unethically by using traditional blockchain technology. All of the data flow maintain and record by using IoT (RFID) technology.

3 Methodology

The safety of food and its reliability becomes a major problem in the world nowadays, due to increasing various ways of serving food to the customers. There are many intermediate stakeholders includes in the food supply chain, so that the FSC network becomes longer. This huge FSC network cause to arise many problems. Such as food fraud, illegal production of food, food contamination, that cause compromises people's health and loss of food industry. It also arises many technical issues, like transparent traceability of food and trustless environment among stakeholders, etc. food regulatory authorities, government, and concern department work a lot to overcome these issues. However, in this regard, we are going to contribute our methodology to solve these food safety and traceability issues.

In our system, we adopt Federated blockchain technology in the food supply chain system which provides some customized features to the user as their requirement or condition. Actually, a well-known use of blockchain technology is for digital currency transactions, which required a more transparent and detailed flow of information among the nodes of the network. But when block chain technology (BCT) integrates with the food supply chain, it makes the food supply chain transparently traceable from origin to customers with a complete flow of information. Somehow it is beneficial to make the FSC, traceable, transparent and trustworthy, among the nodes of the network, due to immutable distributed ledger update. But, besides this when BCT adopted by food industries, some privacy issues arises in FSC, because there is some private information that needs to hide from the other nodes present in the network. This private information can be used for fraud and competitors can be used this information to harm the reputation of the food companies. So, our methodology overcomes these issues by providing a Federated blockchain-based food supply chain system, In the Federated blockchain, Control is given only to few predetermined nodes. So, any particular transaction can make be restricted to non-leaders nodes.

Federated blockchain is partially decentralized. The consensus process is controlled by a pre-selected set of nodes in the network. These blockchains are faster and more scalable. Federated blockchain provides more transparent, trustworthy traceable information flow among the stakeholders. It provides the rights to the pre-selected set of nodes to customize the information regarding the federated member's need and they also customize intermediate stakeholder's appearance in the network.

Therefore, we divide our federated blockchain-based food traceability system into three streams. In which, 1st is Upstream which contains regulators, 2nd is Middle- stream which contains intermediate stakeholders and 3rd is Downstream which contains end-users. These streams build three pillars of federated members according to their control, rights, and usage. In Upstream, regulators are declared as leader nodes or full nodes, they can be food safety organizations. Leader nodes have complete rights to access blockchain information. All the information from origin to costumers is visible and traceable for them. They can manage and control, right of access of intermediate stakeholders and end-users, they can set different smart contracts for different types of intermediate stakeholders in the FSC network. They can insert data, search data, query data, and retrieve data at any time or any point from the blockchain. In the Middle- stream, intermediate stakeholders or middle participants are middle nodes, they can be suppliers, wholesalers or retailers, etc. Middle nodes have limited rights to access FSC information from the federated blockchain. Only one step above and one step down information are visible and traceable for them such as, where they got material and where they send after. They can insert data and queries of data from the FSC blockchain network. In Downstream, end-users are lower nodes, they can costumers. Lower nodes have only rights to get the query of data of food product from the FSC blockchain, regarding they bought from the retailer.

To select federated leader nodes of the FSC blockchain. We use PBFT to leader's election in the system among all the members in the blockchain. Elected nodes (Regulators) become registered leader nodes and get rights to login to the system as a leader or full node. They have the right to make different types of protocols and smart contracts for intermediate stakeholders

or end-user to access information from the FSC blocks chain. They have the right to get in or out of any intermediate federated members.

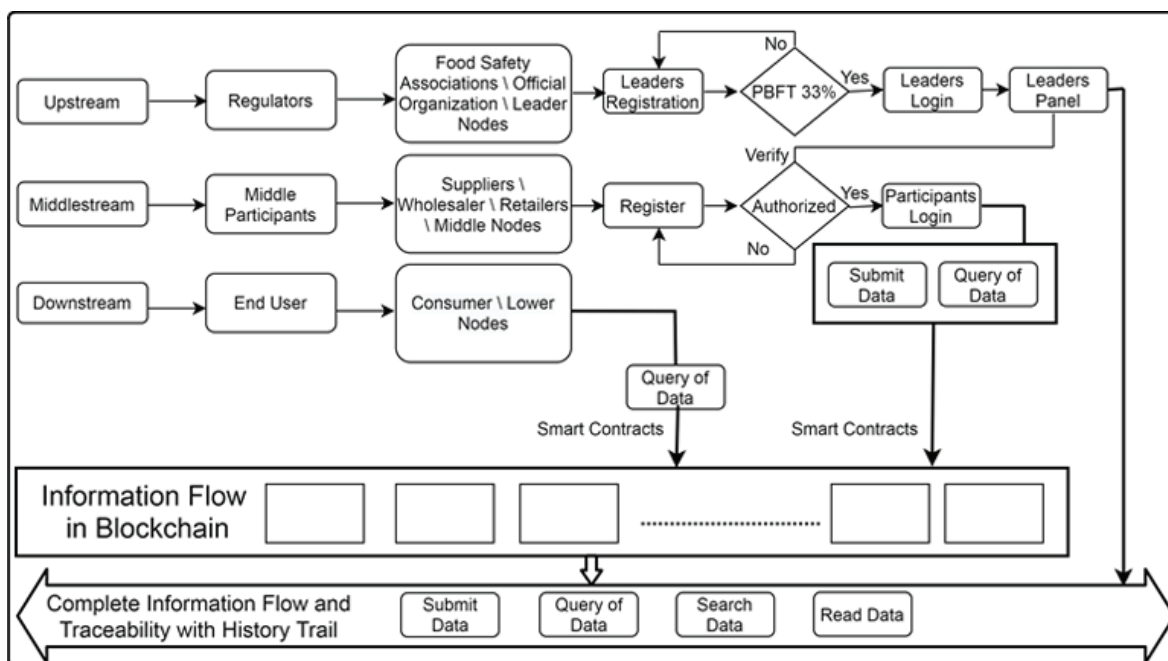


Figure 1: Federated Blockchain Based Food Supply Chain System

On the other hand, Federated intermediate stakeholders can register themselves in the system only when they are authorized by the leader nodes or smart contracts which they set for authorization. If they are authorized by the leader nodes, they have access to login in the system with limited rights. Access rights were assigned by smart contracts and the smart contracts set by the leader nodes with consensus. The registration, login, and access records store in the blockchain as proof. Middle nodes can register themselves with the permission of leader's nodes, they have the right to submit data regarding food information and get queries from previous data relevant to it by invoking smart contracts. The information they submit will be verified by all participants present in the network. There is no need for registration of lower nodes or customers. Customers have only the rights to access blockchain information for the query of data regarding they bought from the retailer. There are many smart contracts set for the query of data, insertion of data, and search for data to the regulators, intermediate stakeholders, and customers. So, in our Federated blockchain-based food supply chain system, we declare full nodes and medium nodes on the behalf of right, access, and control in the network. Lower node is not a part of this, because they do not have any kind of rights and control. Due to the full node, regulators have the right to see all information flow in the food supply chain without fear of privacy concerns regarding business competitors. Due to medium nodes, intermediate stakeholders stores and see their own information. The reason for setting full node and medium node in the FSC, that if regulators are the only nodes to insert data in the FCS can tamper with the original data and that information can be trustless for the intermediate stakeholders. So, to make transparent traceability mechanisms in FSC, Intermediate stakeholders or medium nodes have the right to store their own food information.

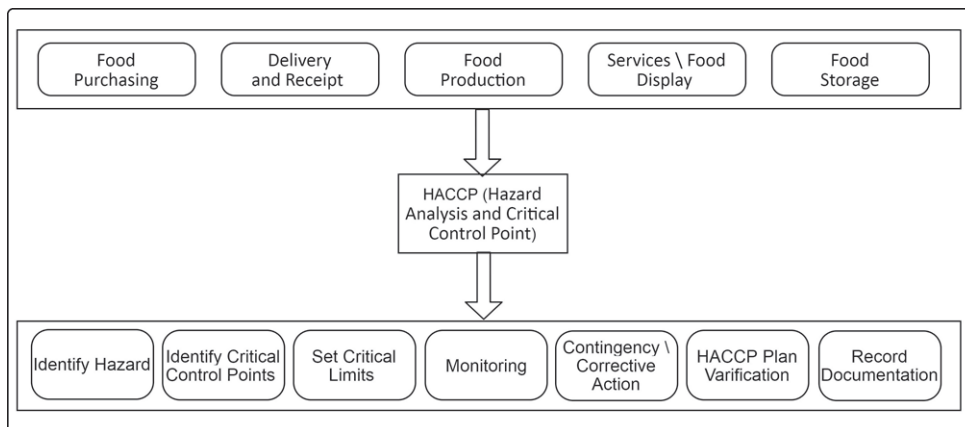


Figure 2: HACCP Based Food Supply Chain System

After proposed our Federated blockchain-based food supply chain system methodology, we can say that, the information visibility or data hiding rights, and controlled access of authorized members, regulators can achieve information privacy and controlled mechanism of transactions. Federated blockchain-based FSC minimizes the fear of theft of valuable information by bad cops and competitors. Our Federated blockchain-based food supply chain system provides effective information flow, customized transparent traceability, security of data, and trustworthy relation among stakeholders. When the federated blockchain mechanism is integrated with the HACCP so, the flow of the work goes more appropriate, secure, and traceable.

In the initial stage for the organization when the product or food is in the process of purchasing, the quantity, detail of the seller, detail of the receiver (this receiver is the part of the system and can authenticate by the RFID and insure by

the federated blockchain mechanism), detail of food before purchasing the food, is traced in the system and update the profile of the food product and the user, also all the product which are going to part of the system is assigned RFID tags to get more updates from time to time for further process.

The purchasing receipt is also updated in the system with a relevant receiver profile to ensure that there is no mishap or fraud in between this process. All this process is managed by the head of the sale and this managed by the main leader of the system who have a control of everything. By this surety, the further process flow moves to the next step. In food production and all the products are added to a system as a new product. Unique product information is added to the profile.

After that, the food service is initiated to check by the management to ensure the quality of the food, that which product should display or discard. The profile of the management is updated when task done by unique RFID tags by checking to the main lead head. The knowledge of purchased goods will be accessed automatically by setting up the appropriate IoT infrastructure in the warehousing center. Whereas, the product's real-time storage statistics, including amount, type, temperature, and the storage period, can be monitored by monitoring types of equipment and can track both, the product profile and the tag with wireless sensors. Inventory

information can be checked by the lead in the system by the RFID. In order to prevent waste and spoilage, managers may agree on the basis of the specific details for which items will be given priority to quickly transfer out of storage.

In between the mechanism from the production to storage, there are many processes going through to check the hazard of the food product. First check the nature of the food product to determine that what can go wrong in the process. Then ensure by the critical control points (CCPs) to identify key points where something can go wrong and its related control to mitigate these points. All the mechanism is updated time to time in a product and management profiles which are interlinked of the related tasks. Also, the main lead provides a manageable direction to the below management. Then, taking controls at CCPs to stop problems establish management in monitoring this process.

By this monitoring, the real-time sensor values are updated in the product profiles. If any harm is identified so take correct actions on the basis of the specific details. These controls and problem-solving strategies ensure and prove the HACCP Program works in the system. At last, all the basic information from start to end is reported as a document. RFID can be used for the basic information when any of the users purchase that food product. This is done because of the federated blockchain that can maintain all the supply chain data fully audit-able. However, users can also get the detail of the product by this traceability system.

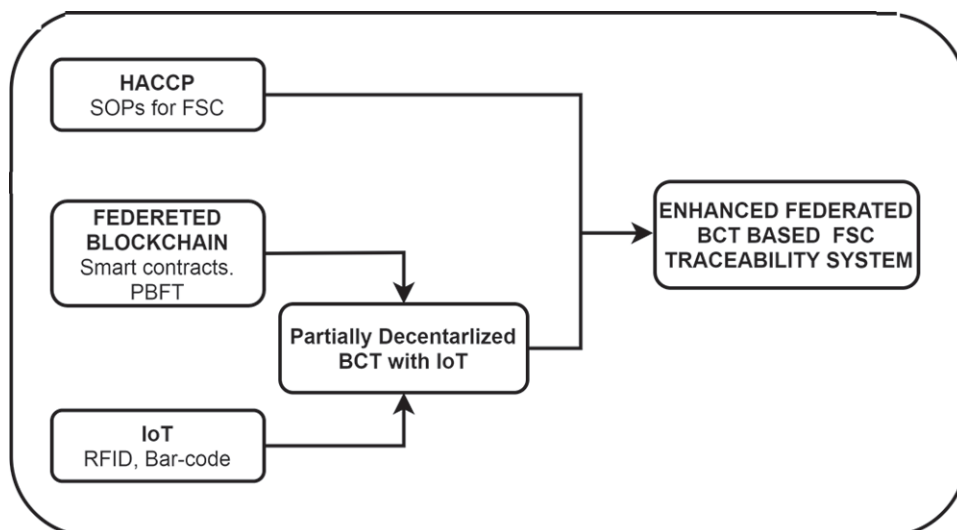


Figure 3: Enhanced HACCP Using Federated BCT Based FCS Traceability System

4 RESULTS

We have derived different abstract results as per other mentioned/listed/referenced papers from our proposed method with the comparison of traditional blockchain-based food supply chain system and HACCP based food supply chain system. The comparison table relies on many characteristics that are used to enhance the functionality and performance of the food supply chain system.

In our flowchart based model, the federated blockchain-based FSC model is comparatively enhanced and provides better features to make the food supply chain system traceable and reliable. we set different parameters like (high, medium, low) to check the effectiveness of our federated blockchain-based food supply chain model from other models.

A comparison table of our federated blockchain-based FSC model with the traditional blockchain FSC model is as under:

Performance Criteria	Performance Variables	Traditional HACCP with FSC	Block chain based FSC	Proposed system (Federated BC with HACCAP) For FSC
Trust	Accountability Immutability Verifiability	Medium Could be tampered Medium	High Nearly Impossible High	High Hard to tampered High
Trust Model	Trusted, Semi-trusted , Un-trusted	Semi-trusted	Trusted	Semi-trusted
Efficiency	Cost Speed Energy consumption Technology	Medium Cost Low Low Energy consumption Traditional Equipments	High Cost High High Energy consumption BCT + IoT	High Cost Very High High Energy consumption BCT + IoT
Access Permission	Read permission Write permission	Public	Private / Restricted	Semi - Private / Selected Nodes
Responsiveness	Customer complaints Response time	Medium Low	Low High	Low Very High
Traceability	Real time Food Tracking	Weak \ Slow Traceability	Strong \ Fast Traceability	Strong \ Fast Traceability
Consensus Mechanism	Poof of Stake \ PBFT Proof of authority Smart Contract	HACCP SOPs	Poof of Stake Proof of authority	PBFT \ Smart Contracts
Food quality	Process quality Product quality	Medium Medium	High High	Very High Very High
Context	Data transparency Security privacy	Low Low Low	High High Low	High High High
Safety	High Safety / Low Safety	Medium	High	Very High
Authority	Centralized / Decentralized	Centralized	Decentralized	Partial Decentralized
Automation	Automated / Semi- Automated	Semi-Automated	Digital Automated	Digital Automated
BFT Tolerance	<=33%	No	No	Yes
Participants	Permissioned / Permissionless	Anonymous	Identified / Trusted	Authorized by Leaders Nodes

Figure 4: Performance Comparison Table of the proposed block chain based FSC traceability model [6],[8],[9].

we observed that our flowchart based model resolves those issues who create hurdles to the adaption of blockchain in the food supply chain with respect to privacy relevant issues which are considered a major hurdle to integrate blockchain in the food traceability system.

5. Conclusion

In this paper we conclude, regarding the food supply chain system. There are many challenging issues which makes difficult to adaption of block chain in food supply chain. For instance many of the suppliers, manufactures coordinate and collaborate with big network of stake holders,

which involve directly or indirectly in FSC. Because of this, it is hard to update information in real-time, also it is hard to sustain confidentiality, transparency and the traceability of FSC. After integrating block chain technology in FSC many of these issues has been resolve but one of the major still remain to solve which is amount of data sharing or privacy concern. Many of the companies or stake holders wants to hide un- relevant information and don't want to share core private information with other stake holders but block chain based FSC failed to overcome with this issue. So, In view of this, we proposed a new partially decentralized FSC traceability model by enables the federated block chain technologies, HACCAP and integration of IoT. Moreover, we demonstrate the working of HACCAP system which give to the real-time information of food with supply chain also, in our Federated block chain based FSC model, full node access and authorization control in the hand of few predetermined nodes. Selection of full leaders nodes by PBFT. These nodes have rights to control action of half or middle nodes and lower nodes. The amount of data sharing can also be controlled by the leader's nodes. So, any particular transaction is restricted to certain nodes because federated block chain is partially decentralized.

Our system improved the traceability, efficiency, transparency, privacy and trust between the involved nodes like stakeholders. Our model highlights the food fraud, illegal production of food, food contamination, which cause to compromises people's health and loss of food industry. By using this digital automated structure of our system. We will able to send and retrieve the confidential information like transactions with real-time environment and partially distributed way also, the main federated node can continuously monitor the goods and transaction digitally. By using this system structure it's significantly reduce the complexity of the FSC system and build the customer confidence of the product.

References

- [1] Casino, Fran, Venetis Kanakaris, Thomas K. Dasaklis, Socrates Moschuris, and Nikolaos P. Rachaniotis. "Modeling food supply chain traceability based on blockchain technology." *IFAC-PapersOnLine* 52, no. 13 (2019): 2728-2733.
- [2] Kamath, Reshma. "Food traceability on blockchain: Wal- mart's pork and mango pilots with IBM." *The Journal of the British Blockchain Association* 1, no. 1 (2018): 3712.
- [3] Liu, Hairong, Xingwei Yang, Longin Jan Latecki, and Shuicheng Yan. "Dense neighborhoods on affinity graph." *International Journal of Computer Vision* 98, no. 1 (2012):65-82.
- [4] Xu, Lei, Lin Chen, Zhimin Gao, Yang Lu, and Weidong Shi. "Coc: Secure supply chain management system based on public ledger." In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1-6. IEEE, 2017.
- [5] Song, Ju Myung, Jongwook Sung, and Taeho Park. "Appli- cations of Blockchain to Improve Supply Chain Traceability." *Procedia Computer Science* 162 (2019): 119-122.
- [6] Caro, Miguel Pincheira, Muhammad Salek Ali, Massimo Vecchio, and Raffaele Giaffreda. "Blockchain-based trace- ability in Agri-Food supply chain management: A practical implementation." In *2018 IoT Vertical and Topical Summit on Agriculture-Tuscany (IOT Tuscany)*, pp. 1-4. IEEE, 2018.

- [7] Wamba, Samuel Fosso, and Maciel M. Queiroz. "Blockchain in the operations and supply chain management: Benefits, challenges and future research opportunities." (2020): 102064.
- [8] Tian, Feng. "A supply chain traceability system for food safety based on HACCP, blockchain Internet of things." In 2017 International conference on service systems and service management, pp. 1-6. IEEE, 2017.
- [9] Kamble, Sachin S., Angappa Gunasekaran, and Rohit Sharma. "Modeling the blockchain enabled traceability in agriculture supply chain." *International Journal of Information Management* 52 (2020): 101967.
- [10] Chang, Shuchih Ernest, Yi-Chian Chen, and Ming-Fang Lu. "Supply chain re-engineering using blockchain technology: A case of smart contract based tracking process." *Technological Forecasting and Social Change* 144 (2019): 1-11.
- [11] Abeyratne, Saveen A., and Radmehr P. Monfared. "Blockchain ready manufacturing supply chain using distributed ledger." *International Journal of Research in Engineering and Technology* 5, no. 9 (2016): 1-10.
- [12] Saberi, Sara, Mahtab Kouhizadeh, Joseph Sarkis, and Lejia Shen. "Blockchain technology and its relationships to sustainable supply chain management." *International Journal of Production Research* 57, no. 7 (2019): 2117-2135.
- [13] Wingreen, Stephen, and Ravishankar Sharma. "A BLOCKCHAIN TRACEABILITY INFORMATION SYSTEM FOR TRUST IMPROVEMENT IN AGRICULTURAL SUPPLY CHAIN." (2019).
- [14] Behnke, Kay, and M. F. W. H. A. Janssen. "Boundary conditions for traceability in food supply chains using blockchain technology." *International Journal of Information Management* 52 (2020): 101969.
- [15] Koirala, Ravi Chandra, Keshav Dahal, and Santiago Matallonga. "Supply Chain using Smart Contract: A Blockchain enabled model with Traceability and Ownership Management." In 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence), pp. 538-544. IEEE, 2019.
- [16] Dib, Omar, Kei-Leo Brousmiche, Antoine Durand, Eric Thea, and Elyes Ben Hamida. "Consortium blockchains: tOverview, applications and challenges." *International Journal On Advances in Telecommunications* 11, no. 12 (2018).
- [17] Tian, Feng. "An agri-food supply chain traceability system for China based on RFID blockchain technology." In 2016 13th international conference on service systems and service management (ICSSSM), pp. 1-6. IEEE, 2016.
- [18] Bocek, Thomas, Bruno B. Rodrigues, Tim Strasser, and Burkhard Stiller. "Blockchains everywhere-a use-case of blockchains in the pharma supply-chain." In 2017 IFIP/IEEE symposium on integrated network and service management (IM), pp. 772-777. IEEE, 2017.

- [19] Vilar, M.J., Rodriguez-Otero, J.L., Sanjua'n, M.L., Die'guez, F.J., Varela, M., Yusa, E., Implementation of HACCP to control the influence of milking equipment and cooling tank on the milk quality. *Trends in Food Science Technology*. 2012, 23(1), 4-12.
- [20] Fotopoulos, C., Kafetzopoulos, D., Critical factors for effective implementation of the HACCP system: a Pareto analysis. *British Food Journal*. 2011, 113(5), 578-597.

Improved User Authentication Process for Third-Party Identity Management in Distributed Environment

Kashif Nisar¹

Shamsuddeen Bala²

Abubakar Aminu Mu'azu³

Ibrahim A. Lawal⁴

Abstract

Third-party identity management user authentication process using single sign-on (SSO) in distributed computer networks requires modification as the process of authenticating user to log into relying party (RP) resources by either identity provider (IDP) or hybrid relying party (HRP) depend always on the authentication of user logins. In this research an algorithm is proposed to authenticate user only once by recording and encrypting user credential with one-way hashing algorithm (SHA2), this simplifies user subsequent logins into relying party by confirming user credentials without other authentication by IDP or HRP. Authentication time and response time continuous time plot of the proposed algorithm was plotted with respect to the arrival time of users in which we show the relationship of authentication time and response time with random arrival rate of users.

Keyword: Single Sign-On, Third-party, Identity management, Distributed networks

1 Introduction

Internet of everything is the current world of innovations, by making everything virtually available on the World Wide Web. The need to manage data easily and securely is the main interest as compared to managing memory space which is no longer an issue (Elgendy & Elragal, 2018). This makes it pretty much easier for people to access dazzling array of online resources at any point in time, anywhere around the globe with the few clicks of the mouse button and/or from their mobile devices (Wada & Tanaka, 2014). However, gaining access to many resources requires login credentials (username and password) and sometimes one time password (OTP) (David et al., 2008). For the obvious reason of security, it is improper to be using single password for accessing resources on line. Same login credentials for accessing different applications/resource provide the need for single sign-on (SSO) (B. Li et al., 2004).

Identity provider (IDP) authenticates and authorise user to have legal right to access relying party (RP) resources, RP is a party of which user access its resource after authentication by IDP (Vapen, Carlsson, Mahanti, & Shahmehri, 2016). However, some RPs are hybrid relying parties (HRPs) in that they act as both IDPs and RPs at the same time. This makes their third-party identity management to be authenticated and to authenticate other RPs. Yet, other RPs are been authenticated by HRPs and IDPs (Vapen et al., 2014). Distributed networks are designed

¹University Malaysia Sabah, Jalan UMS, Kota Kinabalu Saba h, Malaysia |jkashif@ums.edu.my

²Umaru Musa Yar'adua University Katsina, Nigeria |Shamsuddeen.bala@umyu.edu.ng

³Umaru Musa Yar'adua University Katsina, Nigeria |abubakar.muazu@umyu.edu.ng

⁴Bayero University Kano, Kano Nigeria |jialawan.it@buk.edu.ng

as integration of different components assembled together from different independent security domains (Radha & Reddy, 2012). Sensitive information access must be secured and private and the functionality of the system must be to the desired standard, with seamless operation process. Millions of RPs are using different IDPs for their SSO while others act as both IDP and RP (hybrid) at the same time (Vapen, Carlsson, Mahanti, & Shahmehri, 2016).

Internet access login is very essential in the contemporary World Wide Web, hence the need to have secure, well-functioning SSO configuration that has privacy to ascertain as well as the desired security for using one login to access different resources at any time while online, without malicious attack to the login credentials (J. Li et al., 2019)(Wassermann et al., 2019)(B. Li et al., 2004). Although many researchers have examined several aspects of SSO particularly in the areas of security, privacy, and functionality among others (e.g. Beer Mohamed, M. I., Hassan, M. F., Safdar, S., & Saleem, M. Q., 2019; Vapen, Carlsson, Mahanti, & Shahmehri, 2016; Heijmink, 2015; Science & Cao, 2014; David, Nascimento & Tonicelli, 2008 etc.), studies on authentication in Third-Party Identity Management (TIM) appear to be grossly inadequate. This research therefore aims to develop an algorithm geared towards redesigning the authentication process in TIM in order to eliminate the overhead in the authentication process for both IDPs and Hybrid Relying Parties (HRPs)

This paper therefore aims to develop an algorithm geared towards redesigning the authentication process in third-party identity management (TIM) in order to eliminate the overhead in the authentication process for both IDPs and (HRPs). Especially when user is authenticated for the first time there is no need for subsequent authentications, user credentials will be encrypted in one-way hashing algorithm (SHA2) (Beer Mohamed et al., 2019), user credentials will be used to validate subsequent logins by login directly into an RP with the stored credentials.

2 Review of Related Work

All the way through the design of our proposed algorithm previous related works were visited. Beer Mohamed, Hassan, Safdar, & Saleem (2019), proposed an adaptive security architectural model for federated identity management in cloud computing as Service Oriented Architecture (SOA) for software as a service (SaaS). Their architecture was implemented and tested in a large-scale identity enterprise computing environment, where they compared their model with a vendor security and security layer performance in which their model outperforms the vendor security and the security layer performance. In their model algorithm was incorporated to encrypt credential using MD5 hashing algorithm and SHA2 but we choose SHA2 alone due to some drawbacks of MD5 as presented by (hashedout).

In 2016, Vapen, Carlsson, Mahanti and Shahmehri used landscape overview of which sites act as SSO RPs and how different Classes of RPs select their IDPs. They collected data sets using both manual identification and large-scale crawling which they used to identify current state of third-party identity management landscape but class-based analysis was only used to characterize the third-party IDPs landscape. They apply the following method. First, during data collection, they identify RP-IDP relationships and other site characteristics for selected sample websites. Second, they classify the sampled site along four dimensions (primary services, popularity

segment, geographic region, byte/link volume). Third, they used hypothesis testing to identify website classes more likely to act as RPs. Use of popular sites as IDPs is dominant because these already have a large number of users with active accounts. In addition, in many cases, these sites may already access to large amounts of personal information that could help the RP improve their personalisation and service.

Jensen, Marsh, Dimitrakos, & Murayama (2015), adopted mathematical representation modelling and analysis of different requirements of federated identity management to develop a frame work that can formally express trust in federated identity management and how such expression can be used to analyse and evaluate trust qualitatively and quantitatively. In our work, we reduce the circuit of trust they applied in their modelling analysis as in section 3.

3 Architectural Model for User Authentication Process

The process of authenticating users in third-party identity management (TIM) needs modification, this leads to a new way of authenticating user once with a design of an improved authentication algorithm and development of smart security architectural model. The Smart security architectural model has three parts as in Figure 1 these three parts are; (1) user authentication services, (2) smart security engine, and (3) broker managing services.

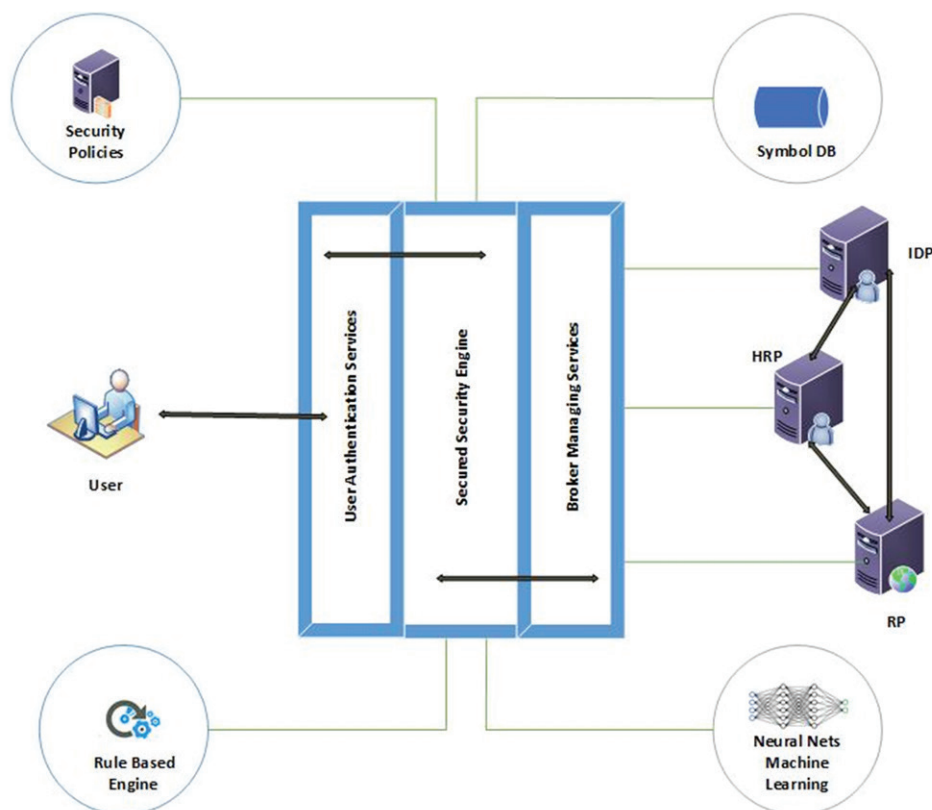


Figure 1: Proposed Model

User authentication services, this part of the Model is responsible for authenticating users, when they send their requests to login into an RP, user request is send to the IDP via broker managing

services and the smart security engine is to validate the user credentials before forwarding it to RP for resource access by user.

The smart security engine (SSE) serves as a central point between user authentication services and broker managing services. When user send a request to login into RP whereby user will be authenticated by IDP via SSE, then SSE will allocate the user with an access token and record user credentials together with access token in symbol table. SSE encrypts user details in one-way encryption format using secured hashing algorithm version II (SHA2) for subsequent logins without any other authentication. Therefore henceforth user will be able to login into any relying party authenticated once with the login details of the IDP that authenticates the user. Broker managing services is an inter-mediator between the SSE and the IDP, HRP, and RP. Broker accept user request via User Managing Services and interact with either RP and IDP, or RP and HRP for authentication.

Mathematical trust relationship of third-party identity management for identity providers (IDPs) and hybrid relaying parties (HRPs) are modelled as: $u_i \in u_{f_i}$, $RP_i \in RP_{f_i}$, $HRP_i \in HRP_{f_i}$, $IDP_i \in IDP_{f_i}$, $f_i \in TIM$, $i \in \{1, 2, \dots, n\}$. $u_i \xleftarrow{\frac{DT:f_i}{FT}} RP_i$ user (u_i) has direct trust relationship (DT) with the Relaying Party (RP_i), and is fully trusted (FT) by the (RP_i); $u_i \xleftarrow{\frac{DT:f_i}{FT}} HRP_i$ user (u_i) has direct trust relationship (DT) with the HRP_i , and is fully trusted (FT) by the RP_i ; $u_i \xleftarrow{\frac{DT:f_i}{FT}} HRP_i$ Relaying Party (u_i) has direct trust relationship (DT) with the HRP_i , and is fully trusted (FT) by the HRP_i ; $RP_i \xleftarrow{\frac{DT:f_i}{FT}} IDP_i$ Relaying Party (RP_i) has direct trust relationship (DT) with the IDP_i , and is fully trusted (FT) by the as inspired by (Jensen et al., 2015).

Thus, the circuit of trust (COT) is reduced to direct trust relationship (DT) only and the trust level is reduced to fully trusted (FT) only. In this research other level of trust, semi trusted (ST), not trusted (NT), and restricted but trusted (RT) were eliminated due the nature of the proposed architectural model, for the level of trust indirect trust (IT) is also eliminated according to what (Beer Mohamed et al., 2019) suggested. Trust relationship between User, RP, HRP, and IDP is expressed as:

$$\left(u_i \xleftarrow{\frac{DT:f_i}{FT}} RP_i\right) \left(RP_i \xleftarrow{\frac{DT:f_i}{FT}} IDP_i\right) = \left(u_i \xleftarrow{\frac{DT:f_i}{FT}} RP_i IDP_i\right) \quad (1)$$

$$\left(u_i \xleftarrow{\frac{DT:f_i}{FT}} HRP_i\right) \left(RP_i \xleftarrow{\frac{DT:f_i}{FT}} HRP_i\right) = \left(u_i \xleftarrow{\frac{DT:f_i}{FT}} RP_i HRP_i\right) \quad (2)$$

A Proposed Algorithm User Authentication Process

The proposed algorithm is designed in such a way that, when user want to access RP resources it will start with authentication step, i.e. step 1 if user is not registered by SSE user will not be authenticated. Therefore user should go to step2, registration step user will be registered in a symbol table and will be given an access token and directed to step 3, encryption step in this step user credential and access token will be encrypted using one-way hashing algorithm (SHA2)

from then user details cannot be accessed unless when user want to login back for subsequent logins SSE will verify user details when the user entered his credentials to confirm the user access token for allowing user to access RP resources or HRP resources as the algorithm is designed to allow IDP and HRP to authenticate user only once i.e. for the first time, and IDP can also authenticate user to access HRP resources. Subsequent logins of user after authentication will be handled by checking user details encrypted if the access token of user is valid for the authenticating party i.e. IDP or HRP user is granted to login as illustrated in Figure 2.

```

1  Hybrid_Algorithm: SecuredSecurityForThirdPartyIdentityManagementUsingSSE
2  Input:- Authentication: User access request from Identity Prover
3  Output:- Response from Relying Party
4  Variables:- U : User; Rp: Relying Party; PKI: Public Key Infrastructure;
5  IDP: Identity Provider; f: Third-Party Identity Management;  $\lambda$  : Arrival rate
6  AT: Access Token; CR: User credentials; HRP: Hybrid Relying Party,
7  HashAlgorithm: Applied Hashing Algorithm (SHA2)
8  broker: Internal resource mediator at SSE;
9  linkageService: Boolean;
10 trustType: (Direct Trust (DT));
11 trustLevel: (Fully Trusted(FT));
12
13 Initialize
14 Step 1: /* Authentication user by IDP */
15   For i = 1 to n
16      $\lambda = 1$ 
17     If  $U_i \in U_n$  and  $IDP_i \in IDP_n$  and  $HRP_i \in HRP_n$  then
18        $U_i \rightarrow$  Authenticate with SSE
19        $U_i \rightarrow$  Authenticated with HRP
20        $U_i \rightarrow$  Authenticate with IDP
21        $U_i \rightarrow$  Authenticated
22     Step 3 /* If user is authenticated goto step 3 */
23   Else
24     Step 2 /* If user is not authenticated goto step 2 */
25   EndIf
26 Step 2: /* User registration with SSE */
27    $U_i \rightarrow$  Registration with SSE
28   If  $U_i \in U_n$  and  $IDP_i \in IDP_n$  and  $Rp_i \in Rp_n$  and  $HRP_i \in HRP_n$  then
29      $U_i \leftarrow$  SSE [CR, Rp, HRP, IDP] /*DT trust type*/
30      $U_i \rightarrow$  Registered with SSE
31     Step 1 /* If user is registered goto step 1 */
32   Else
33     throwException "User can't be registered"
34   EndIf
35 Step 3: /* Encrypting user details */
36   If  $U_i \rightarrow$  Authenticated then
37      $U_i \rightarrow$  Encryption with SSE ] /*FT level of trust*/
38     SSE  $\rightarrow$  generate [AT]
39     SSE  $\rightarrow$  broker with request [ $U_i$ , AT, CR, Rp, HRP, IDP, linkageService]
40     SSE  $\rightarrow$  hashAlgorithm [AT, CR]
41     broker  $\rightarrow$  acceptrequest [AT, CR, digitalSign]
42     broker  $\rightarrow$  forward request Rp, using PKI
43     Rp  $\rightarrow$  verifiedLogin
44     If  $HRP_i \in IDP_n$  then
45       Rp  $\rightarrow$  verifiedLogin
46     Else
47       HRP  $\rightarrow$  verifiedLogin
48     EndIf
49   Else
50     throwException "Access Denied"
51   EndIf
52 End

```

Figure 2: Proposed Algorithm

4 Implementation

The proposed algorithm in this research was implemented, where an exponential distribution was invoked for optimal performance of the algorithm operation. Random arrival time ($t = \lambda$) of users was used and authentication time (T_a) was compared against response time (T_r) for different arrival time to compare T_a of IDP and HRP against T_r of RP for users. The aim of reducing multiple authentication of user during login into RP resources by IDP and HRP for

random users in order to know how they arrive and the relationship of the authentication time of the IDP or HRP and the response time of the RP. Random real values were generated for the authentication time for random users and their corresponding response time.

The arrival of users during authentication is random, there is need to use Exponential Distribution as inspired by (Haq et al., 2019)(Gupta et al., 2010).

$$\text{Arrival pattern will be: } 0 \leq t_{(0)} < t_{(1)} < t_{(2)} < \dots < t_{(n)} \tag{1}$$

The stating time for queuing is assume to start at $t = 0$. The random variables are expressed as:

$$\tau_k = t_k - t_{(k-1)}, \quad (k = 1, 2, 3, \dots) \quad \tau_k = t_k - t_{(k-1)}, \quad (k = 1, 2, 3, \dots) \tag{2}$$

$$\text{Therefore the exponential pattern is: } A[t] = 1 - e^{(-\lambda t)} \tag{3}$$

The result in respect with $\lambda = 1 \text{ ms}$ is shown in Table 1. The result is presented graphically as in Figure 3 result was generated from random real values from 0 to 1.

Table 1: Proposed Algorithm Result

Users	Response Time (Seconds)	Authentication Time (Seconds)	λ (Mili-Seconds)
User1	0.0047835	0.0512164	1
User2	0.0318328	0.4579892	1
User3	0.0344461	0.646313	1
User4	0.0357117	0.793975	1
User5	0.0364413	0.807531	1

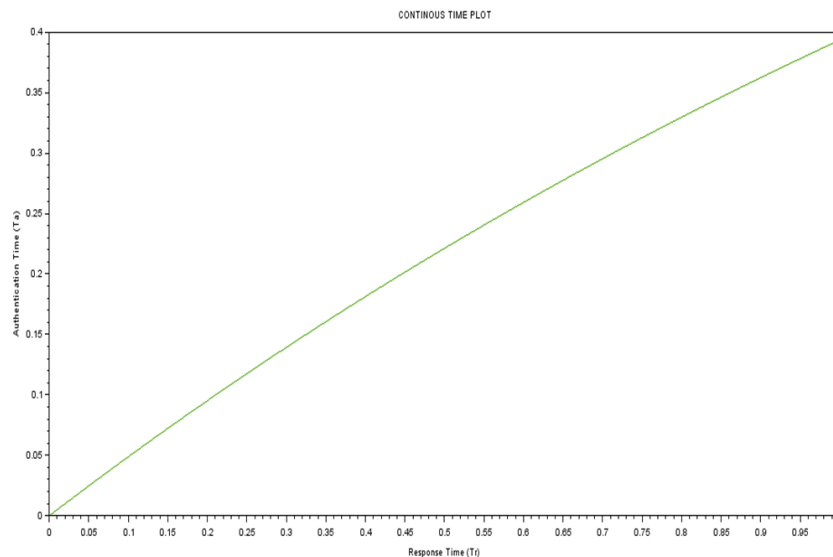


Figure 3: Proposed Algorithm I Result with $\lambda = 1 \text{ ms}$

To know the difference of result plotted for authentication time against response time with respect to $\lambda = 0.5$ ms and $\lambda = 0.1$ ms for the proposed algorithms, their results is compared in figure 4.3 to showcase their difference. Therefore the arrival rate determines the authentication time of the Identity Providers (IDP) and Hybrid Relying Party (HRP) relationship with response time of Relying Parties (RP). Hence the high the arrival rate a better performance is recorded, whereas the lower the arrival rate a lower performance is recorded

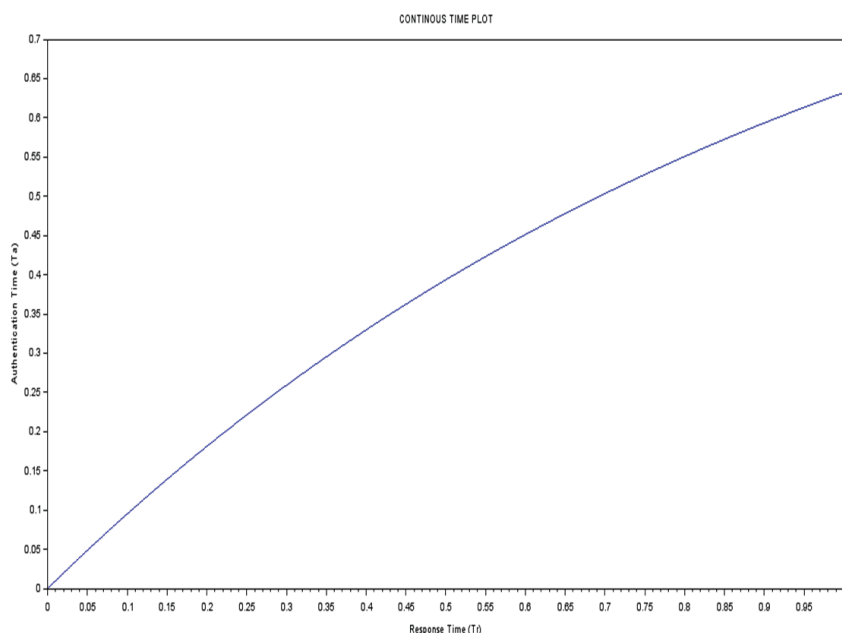


Figure 4: Proposed Algorithm II Result with $\lambda = 0.5$ ms

5 Conclusion

User authentication for Third-Party Identity Management is a phenomenon whereby user needs authentication from IDP and HRP for each login into RP resources. This paper have presented improved authentication process for third-party identity management in which the overhead of authenticating user at each login to access RP resources by IDP and HRP is reduced to single authentication as in the proposed algorithm. The user credentials are been encrypted using one-way hashing algorithm (SHA2) as depicted in the proposed model.

References

- [1] Armando, A., Carbone, R., Compagna, L., Cuellar, J., & Tobarra, L. (2008). Formal analysis of SAML 2.0 web browser single sign-on: Breaking the SAML-based single sign-on for google apps. Proceedings of the ACM Conference on Computer and Communications Security, 1–9. <https://doi.org/10.1145/1456396.1456397>
- [2] Beer Mohamed, M. I., Hassan, M. F., Safdar, S., & Saleem, M. Q. (2019). Adaptive security architectural model for protecting identity federation in service oriented computing. Journal of King Saud University - Computer and Information Sciences, xxx. <https://doi.org/10.1016/j.jksuci.2019.03.004>

- [3] David, B. M., Nascimento, A. C. a, & Tonicelli, R. (2008). A Framework for Secure
- [4] Elgendy, N., & Elragal, A. (2018). Big Data Analytics : A Literature Review Paper Big Data Analytics : A Literature Review Paper. September 2014, 214–227. <https://doi.org/10.1007/978-3-319-08976-8>
- [5] Gupta, A. K., Zeng, W. Bin, & Wu, Y. (2010). Probability and statistical models:
- [6] Haq, M. A. ul, Usman, R. M., Hashmi, S., & Al-Omeri, A. I. (2019). The Marshall-Olkin length-biased exponential distribution and its applications. *Journal of King Saud University - Science*, 31(2), 246–251. <https://doi.org/10.1016/j.jksus.2017.09.006>
- [7] Jensen, C. D., Marsh, S., Dimitrakos, T., & Murayama, Y. (2015). Trust management IX: 9th IFIP WG 11.11 international conference, IFIPTM 2015 Hamburg, Germany, may 26-28, 2015 proceedings. *IFIP Advances in Information and Communication Technology*, 454(December 2016). <https://doi.org/10.1007/978-3-319-18491-3>
- [8] Li, B., Ge, S., Wo, T. Y., & Ma, D. F. (2004). Research and implementation of single sign-on mechanism for ASP pattern. *Grid and Cooperative Computing Gcc 2004, Proceedings*, 3251(2001), 161–166.
- [9] Vapen, A., Carlsson, N., Mahanti, A., & Shahmehri, N. (2016). A look at the third-party identity management landscape. *IEEE Internet Computing*, 20(2), 18–25. <https://doi.org/10.1109/MIC.2016.38>
- [10] Vapen, A., Carlsson, N., Mahanti, A., & Shahmehri, N. (2014). Third-party identity management usage on the web. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8362 LNCS, 151–162. https://doi.org/10.1007/978-3-319-04918-2_15

Predicting the Visibility of the First Crescent

Tafseer Ahmed¹

Abstract

This study presents an application of machine learning to predict whether the first crescent of the lunar month will be visible to naked eye on a given date. The study presents a dataset of successful and unsuccessful attempts to find the first crescent at the start of the lunar month. Previously, this problem was solved by analytically deriving the equations for visibility parameter(s) and manually fixing threshold values. However, we applied supervised machine learning on the independent variables of the problem, and the system learnt about the criteria of classification. The system gives precision of 0.88 and recall of 0.87 and hence it treats both false positives and false negatives equally well.

Keyword: crescent visibility, astronomy, supervised learning, feature engineering, ensemble

1 Introduction

Humans are interested in the problem of first sighting of the crescent from ancient times. There were many civilizations that used the lunar or lunisolar calendar, and the first crescent after the sunset marked the beginning of the new lunar month. The following paragraphs explain some fundamental phenomena and terms of astronomy. The details of these terms and phenomena can be found at [1] & [2].

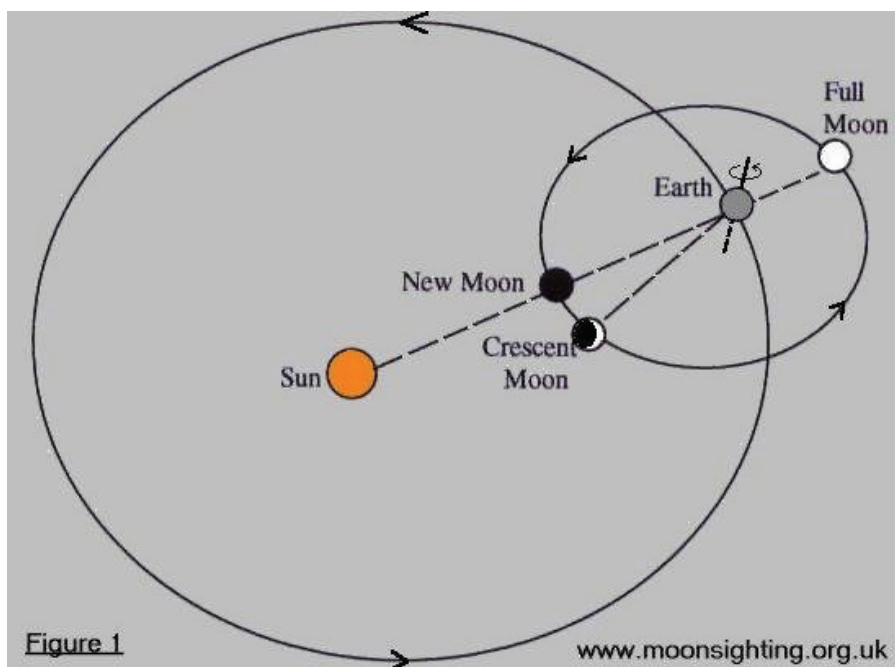


Figure 1: New Moon and Crescent Formation

¹Mohammad Ali Jinnah University | tafseer.ahmed@jinnah.edu

The formation of the crescent is explained in figure 1. The moon rotates around the sun and its relative position with the sun and earth gives the phase of the moon. When the moon is opposite to the sun with respect to earth, we see the full moon. Afterwards, its anticlockwise movement gives waning phases and it becomes invisible after 14 days, as its face towards the earth does not receive light from the sun and it becomes dark.. In the last days of the month, it is visible before the sunrise at the eastern horizon.

When the moon comes between the sun and the earth, or alternatively speaking when sun and moon meet in the sky, we call it sun-moon conjunction or simply conjunction. It is also said that the new moon is born at the conjunction. However, this new moon is not visible to the human eye. The first crescent or the waxing crescent becomes visible above the western horizon after the sunset when it has become sufficiently luminous (by reflecting the light of the sun).

The astronomers know precisely when the (sun-moon) conjunction occurs. The charts are available for the conjunction time of each new moon e.g sun moon conjunction time of each lunar month can be seen at [3]. However, crescent visibility information is not known precisely, because it depends on many factors including human vision, atmospheric conditions, moon's phase and moon's altitude over the horizon. Hence, there exist many models that try to predict the possibility of sighting of the first crescent after the new moon's birth e.g. [3] and [4].

The astronomers focus on this problem for scientific as well social reasons. Many cultures and religions use lunar and lunisolar calendars. They start the new lunar month with the new moon. Many of these use astronomical data to mark the new moon / crescent. The muslim hijri calendar is also a lunar calendar, and the new month start with sight of the crescent [5]. Some countries e.g. Turkey (and its followers) and organizations use the astronomical calculation [4], in place of actual sighting. There are many other calendars on the basis of different visibility or presence criteria of new moon e.g. [6] and [7]. The Ministry of Science and Technology, Pakistan has also issued a criterion and a calendar [8].

However, many communities and countries rely on actual sighting of the crescent (inside or outside the country) to start the new month. Pakistan's authority for moonsighting (Ruet e Hilal Committee) uses crescent sighting models to eliminate improbable claims. In this context, we modeled this problem using machine learning. It is considered as a classification problem having the classes: visible = yes and visible = no. There exists many other works for this prediction, however they are based on astronomical models. According to our knowledge, it is the first attempt to create the crescent visibility model using machine learning.

The remaining paper is organized as follows. Section 2 presents earlier work on the astronomical models of crescent visibility prediction. We mention some limitations of these models, and then present our model in section 3. The section 3 also describes the experiment to evaluate our model. The results of the experiment are presented and discussed in section 4. Section 5 gives description of the further work that can be done to improve the model.

2 Literature Review

The prediction of the first crescent of the month has been an area of interest from the ancient time. Babylonians [9] devised a method to predict visibility of the first moon. Their method involved two features: the elongation and moonset lag time. Table 2 shows that the results of babylonian method are comparable to many modern day methods. However, as the evaluation results show that there were many errors in the prediction. Later astronomers tried to improve the prediction model by adding other factors that affect the prediction of crescent visibility.

In 1911, Maunder [10] proposed that there is a region of brightness around the point of sunset. If the moon is present in that region, then the crescent will not be visible. He proposed the following formula for ARCV (arc of vision). It is calculated by the following equation using the difference in azimuth (DAZ) of the sun and the moon.

$$\text{ARCV} = 11 - |\text{DAZ}|/20 - (\text{DAZ}^2)/100$$

The equation is obtained by learning parameters using observed values of DAZ and ARCV. The crescent is only visible, if its altitude is less than ARCV. Altitude and Azimuth are the coordinates of the objects in the sky [1] [2], and these are explained by the following figure.

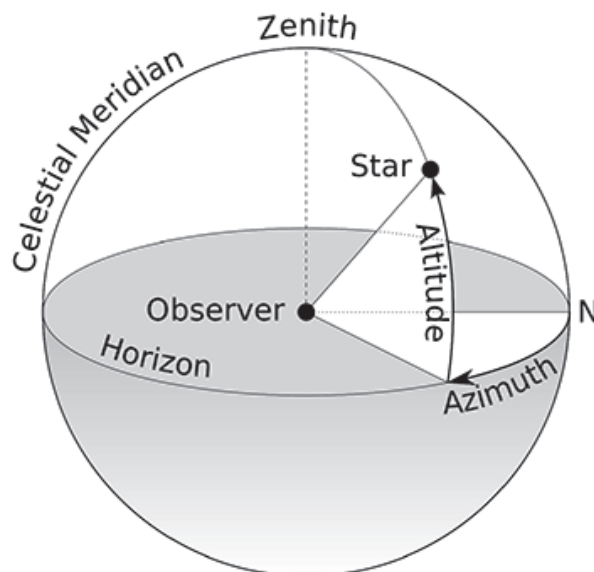


Figure 2: Altitude and Azimuth¹

In 1930, Schoch gave an alternate equation to calculate ARCV based using the observational data [11]. So, we have alternate methods to calculate ARCV. However, the parameters of both of these equations are derived using a small dataset. In section 3, we collect a larger dataset and propose a method that is able to use the purpose of ARCV without calculating it. The next generation of researchers focused on other factors of crescent visibility. Yallop [12] introduced other derived variables in the calculation. As this paper does not deal with astronomy, so we do not present the equations involved and the reasons for these equations. The calculations can be summarized as:

$$\text{ARCV} = f(\text{DAZ}), \text{ARCL} = f(\text{ARCV}, \text{DAZ}), W' = f(\text{ARCL}), q = f(\text{ARCV}, W') \quad (1)$$

Yallop collected 296 observations, and he recorded results of visibility by eye as well as by optical aid (binocular or telescope). By inspecting the observations and corresponding q-values, he presented the following rules.

Table 1: Yallop Visibilitaty Rules

Range	Remarks
$q > +0.216$	Easily visible
$+0.216 \geq q > -0.014$	Visible under perfect conditions
$-0.014 \geq q > -0.160$	May need optical aid to find crescent
$-0.160 \geq q > -0.232$	Will need optical aid to find crescent
$-0.232 \geq q > -0.293$	Not visible with a telescope
$-0.293 \geq q$	Not visible

Odeh [13] collected 737 observations to develop a better system of prediction. He derived an equation for the single parameter V, such as $V = f(\text{ARCV}, W')$. On the basis of different values of V, he defined the following four zones for crescent visibility.

- Zone A: visible by naked eyes
- Zone B: visible by optical aid, and it could be seen by naked eyes
- Zone C: visible by optical aid only
- Zone D: not visible even by optical aid

Qureshi [14] worked on the Odeh's dataset, however he removed some observations from this dataset. These removed observations were made in the morning. He derived new equations and then manually defined regions on the basis of threshold values. He compared the results of his method with other methods. The following table is created by using data presented in his evaluation results.

Table 2: Evaluation of different Methods (src: [14])

Criterion	Precision	Recall	F-score	Accuracy
Babylonian	0.63	0.96	0.76	0.75
s-value	0.61	0.94	0.74	0.73
q-value	0.69	0.92	0.79	0.79
Maunder	0.8	0.76	0.78	0.82
Fotheringam	0.87	0.54	0.67	0.79

The above data clearly depicts the precision-recall tradeoff in different methods. The methods with high recall (i.e. low number of false negatives) have low precision (i.e high number of false positives) and vice versa. In this case, a false positive means that the moon is predicted as visible, but it was not actually visible, whereas false negative means that the crescent is predicted as not visible, but it was actually visible. We need a method that minimizes both of these errors. To accomplish this task, we decided to apply machine learning on a bigger dataset. The following section gives the description of our method and its result.

3 Crescent Visibility Prediction

The literature discussed in the above section tells that the first crescent visibility is predicted by deriving different equations on the basis of astronomy based variables, and then manually selecting a cut-off number for the decision of the visibility.

We choose machine learning as an alternate way to model this problem. The advantage of using machine learning is that we do not need to make a theory of crescent visibility and derive different equations. In the presence of training examples, the first crescent visibility can be modeled as a classification task. The classification algorithm can itself learn the importance (e.g. weight) of different astronomy based features. However, for the application of classification algorithms, we need a bigger dataset. Moreover, we need to select (and create) features for the algorithms. The details of these two tasks are presented in section 3A and 3B.

A Dataset Collection

The currently available datasets are not big. Odeh presented 737 observations [13]. Quereshi [14] removed some irrelevant observations from this list and he worked on 463 observations. A recent work has used only 254 observations, of which only 81 are of positive sighting [15]. These examples are not enough to train a machine learning algorithm, hence we searched for more observations.

There are some websites that record the (successful or unsuccessful) attempts to find the first crescent for the islamic month. One of the websites is managed by Islamic Crescent Observation Project (ICOP). The website has records of observations by different professionals and amateurs for 21 lunar years (1419AH to 1441AH) [3]. Every month (of muslim hijri calendar) has a page associated with it. The page has the visibility map and conjunction date and time for the new moon of that month. The details of observations for that month's first crescent are presented on this page.

The earlier years have a free text description of the observations in which the location of the observer, sky conditions and result of the observation are described. However, from 1431AH, we find that the observations are presented in a regular format and each of the following are mentioned for each observation. We term the features of table 3 as raw features. The hijri month and conjunction date are added to the raw features of each observation.

Table 3: ICOP Observation Features

Observation Feature	Possible Values
Time of Observation	After Sunset, ...
Observer Name	
Location	City, State and Country
Sky Condition	Clear, Partially Cloudy, Total Cloudy
Atmosphere	Superb, Clear, Hazy, Very Hazy
Visible By	Eye, Binocular, Telescope, CCD image

The observation results are added as four different features corresponding to Visible by Eye, Visible by Binocular, Visible by Telescope, and Visible by CCD image. The features have binary values, and more than one features can have yes values, as given in the icop observations.

This raw feature dataset has 2592 observations. In this dataset, there were 459 observations in which the sky condition is totally cloudy. As we cannot see the crescent behind the clouds, we removed all the observations having total cloudy sky condition. Similarly, 611 observations were removed that have partly cloudy and not visible (by eye, binocular, telescope, or CCD image) features.

After this filtering, we added astronomy based features to each observation. We termed the resulting feature-set as rich feature-set. The new features added are longitude and latitude of the location, moonset time, sunset time, altitude and azimuth of sun and moon, and moon phase. These features are obtained by using python libraries geopy [16] and pyEphem [17]. The rich feature-set have independent variables related to the first crescent visibility problem. The features used in the machine learning system are described in the next subsection.

B Feature Engineering

An important decision for machine learning is the selection and engineering of the relevant features. We have seen in the literature review that researchers have created equations for ARCV and ARCL etc. on the basis of the variables that are part of our rich feature-set. The weights of some of these equations e.g. (ACRV equation) are learnt using regression. However, this weight learning is performed on a smaller dataset.

We created a rule that only the independent variables and the difference of independent variables are used in our feature-set for machine learning. The contribution of any dependent variables, if relevant, will be devised by the machine learning algorithm. Hence we have the following features in our Machine Learning feature-set.

Table 4: Crescent Features for Machine Learning

Feature	Calculation Method
Age of Moon	conjunction – sunset
Sun Moon Lag	moonset – sunset
Altitude Diff.	moonAltitude - sunAltitude
Azimuth Diff.	moonAzimuth - sunAzimuth
MoonPhase	given by pyEphem
Atmosphere	Superb = 1, Clear = 0.85, Hazy=0.6, Very Hazy = 0.35

The value of these features are normalized to a range of 0-1. The atmosphere condition is represented by a number having high value, if the atmosphere is clear. The output variable for this experiment is the binary feature visible by eye. However, in any further work, the other visible by features or their combination can also be used as output variable.

C Experiment

The dataset described in section 3A is transformed into the machine learning features described in table 4. The first 80% of this data (approximately first 8 lunar years) is used as the training data. The remaining data (approximately last 18 months) is used as the test data. We did not split the data randomly into training and test sets, as our current experiment setting is more challenging. The test data is completely unseen. The classifiers have not seen the observation and result of adjacent city or country for the same lunar month, so any overfitted model will not get help in the test.

We used four supervised learning algorithms. Three of these are classification algorithms namely Logistic Regression (LR), Support Vector Machine (SVM), Random Forest Regressor (RFR), and Neural Network (NN). The Neural Network has one hidden layer. All of these algorithms are used by using scikit-learn library [18].

As Random Forest Regressor (RFR) [19] is a regression based technique, it accepts and returns a number as the output. Hence, for RFR, we use visible = yes as 1.0 and visible = no as 0. Then, if the test example gives a result of 0.5 or greater, we consider it as visible = yes. The results of these four supervised learning algorithms are discussed in the next section.

4 Results and Discussion

As described in section 3.3, we applied four supervised learning algorithms on the dataset. The results of applying these models for predicting 304 observations of the test data (last 18 months from 5-1440 AH to 10-1441 AH) are given in table 5.

The fourth row of table 5 gives the evaluation results of the classifier on the basis of Random Forest Regressor (RFR). When the output value is equal or greater than 0.5, then we classify the observation as visible = yes. We used some other values of this threshold (0.5) to improve

the recall of the system. We want to reduce the count of false negatives i.e. the cases in which crescent is predicted to be not visible, but it becomes visible to the human eye. Hence, we tried different settings to get a system, given in the fifth row, with better recall.

Table 5: Result

	Precision	Recall	F-score	Accuracy
LR	0.87	0.82	0.84	0.83
SVM	0.88	0.83	0.86	0.84
NN	0.85	0.87	0.86	0.84
RFR (≥ 0.5)	0.88	0.88	0.88	0.86
RFR(≥ 0.33)	0.83	0.94	0.88	0.86
≥ 1 positive	0.86	0.9	0.88	0.86
≥ 2 positive	0.87	0.88	0.88	0.86
≥ 3 positive	0.88	0.81	0.84	0.83

Similarly, we used ensembles of the four algorithms and created systems that declare visible = yes when at least one, at least two or at least three algorithms give the positive result. The last three rows present results of these ensembles. The ≥ 1 positive gives a good precision, and the highest recall value (among the ensembles). We propose to use it as the best system that balances both the precision and recall.

5 Conclusion and Future Work

We developed a system for predicting first crescent visibility using machine learning algorithms. We created this system by using observations of 11 lunar years. According to our knowledge, it is the first attempt of applying machine learning on first crescent visibility prediction². It shows that the usage of simple features (age of moon, sun_moon_lag, altitude_difference, azimuth_difference, moon_phase, atmospheric_ondition) can give a good prediction for modeling this astronomy based phenomenon. This method is different from the method used by other researchers who created derived features by using astronomical techniques/models.

We will share the rich feature set of these observations on some dataset sharing platform e.g. Kaggle. It will enable other researchers to apply their models to create an improved system. The system can be improved by adding more observation examples. There is more data available on ICOP [1] and moonsighting [2] website. This data is needed to be manually inserted into the raw dataset. Moreover, a major crowdsourcing drive of crescent visibility observation is required, as more data (and specially data from different regions of the world will improve the system.) The system has only one feature for the atmospheric conditions, and it is a subjective feature having

²There are some works on searching and detecting the crescent in the sky using computer vision techniques e.g. [20] and [21]. However, our work is different as it predicts the visibility several days or years before the actual time of crescent sighting.

values superb, clear, hazy and very hazy. In its place, we need empirical features e.g. temperature and humidity etc. as done by [22]. We need to add these into this and future datasets.

References

- [1] [1] I. Ridpath, "Oxford Dictionary of Astronomy", 2nd edition, Oxford University Press, 2012.
- [2] W. M. Smart, and R. M. Green. "Textbook on Spherical Astronomy", revised edition. Cambridge University Press, 1977.
- [3] <http://www.icoproject.org/res.html?l=en>
- [4] <https://www.moonsighting.com/>
- [5] M. Ilyas, "Lunar Crescent Visibility Criterion and Islamic Calendar", Quarterly Journal of the Royal Astronomical Society, Vol. 35: 425-461, 1994.
- [6] Maskufa, and S. Hidayatulla, "Global Hijriyah Calendar as Challenges Fikih Astronomy", International Conference on Law and Justice (ICLJ 2017), Indonesia, 2017.
- [7] O. Zainon, H. R. Ali and M. F. Abu Hussin, "Comparing the New Moon Visibility Criteria for International Islamic Calendar Concept", 6th International Conference on Space Science and Communication (IconSpace): 144-149, Johor Bahru, Malaysia, 2019.
- [8] <http://pakmoonsighting.pk/Introduction.aspx>
- [9] L. J. Fatoohi, F. R. Stephenson, S. S. Al-Dargazelli, The Babylonian First Visibility of the Lunar Crescent: Data and Criterion, Journal for the History of Astronomy 30 (1): 51-72, 1999.
- [10] M. Maunder, "On the smallest visible phase of moon", Journal of the British Astronomical Association, XXI:355-362, 1911.
- [11] C. Schoch, "Tafel fur Neulicht", Ergaenzungsheft zu den Astronomischen Nachrichten, 8(2): B17, 1930.
- [12] B. D. Yallop, "A method of predicting the first sighting of new moon", NAO Technical Note No. 69, HM Nautical Almanac Office, Royal Greenwich Observatory, Cambridge, UK, 1997.
- [13] M. S. Odeh, "New criterion for lunar crescent visibility", Experimental Astronomy 18: 39-64, Springer, 2004.
- [14] M. S. Qureshi, "On the comparative study of mathematical models for earliest visibility of the crescent moon and their modification", Ph.D. Thesis, University of Karachi, 2007.
- [15] N. Ahmad et al., "A New Crescent Moon Visibility Criteria using Circular Regression Model: A Case Study of Teluk Kemang, Malaysia", Sains Malaysiana 49(4)(2020): 859-870, 2020.
- [16] <https://pypi.org/project/geopy/>
- [17] B. C. Rhodes, "PyEphem: Astronomical Ephemeris for Python", Astrophysics Source Code Library, ascl:1112, 2011.

- [18] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python", the Journal of machine Learning research 12: 2825-2830, 2011.
- [19] T. F. Cootes, et al., "Robust and accurate shape model fitting using random forest regression voting", European Conference on Computer Vision. Springer, 2012.
- [20] M. Fakhar, et. al., "Lunar Crescent Detection Based on Image Processing Algorithms", Earth, Moon, and Planets, 114.1-2: 17-34, 2014.
- [21] K. Alhammadi, et al., "Moon Crescent Tracker", International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2019.
- [22] B. E. Schaefer, "Visibility of the lunar crescent", Quarterly Journal of the Royal Astronomical Society 29:511-523, 1988.

Call for Papers/Authors Guideline

KIET Journal of Computing & Information Sciences (KJCIS) is biannual publication of College of Computing & Information Sciences, Karachi Institute of Economics and Technologies. It is published in January and July every year. We are lucky to have on board prominent and scholarly academicians as part of Advisory Committee and reviewers.

KJCIS is a multi-disciplinary journal covering viewpoints/ researches / opinions relevant to the non exhaustive list of the topics including data mining, big data, machine learning, artificial intelligence, mobile applications, computer networks, cryptography & information security, mobile and wireless communication, adhoc & body area networks, software engineering, speech & pattern recognition, evolutionary computation, semantic web & its application, data base technologies & its applications, Internet of Things (IoT), computer vision, distributed computing, grid and cloud computing.

The authors may submit manuscripts abiding to following rules:-

- Certify that the paper is original and is not under consideration for publication in any other journal. Please mention so, in case it has been submitted elsewhere.
- Adhere to normal rules of business or research writing. Font style be 12 points and the length of the paper can vary between 3000 to 5000 words.
- Illustrations/tables or figures should be numbered consecutively in Arabic numerals and should be inserted appropriately within the text.
- The title page of the manuscript should contain the Title, the Name(s), email address and institutional affiliation, an abstract of not more than 200 words should be included. A footnote on the same sheet should give a short profile of the author(s).
- Full reference and /or websites link, should be given in accordance with the APA citation style. These will be listed as separate section at the end of the paper in bibliographic style. References should not exceed 50.
- All manuscripts would be subjected to tests of plagiarism before being peer reviewed.
- All manuscripts go through double blind peer review process .
- Electronic submission would only be accepted at kjcis@pafkiet.edu.pk
- All successful authors will be remunerated adequately.
- The Journal does not have any article processing and publication charges.

Submission is voluntary and all contributors will find a respectable acknowledgment on their opinion and effort from our team of editors. Submission of a paper will be held to imply that it contains original unpublished work. In case the paper has been forwarded for publication

elsewhere, kindly apprise in time if the paper has been accepted elsewhere. Manuscripts may be submitted before September and May to get published in Jan & July issues respectively. We encourage you to submit your manuscripts at kjcis@pafkiet.edu.pk

Editorial Board KJCIS
College of Computing & Information Sciences
Karachi Institute of Economics and Technology



Karachi Institute of Economics and Technology

Korangi Creek, Karachi-75190, Pakistan

Tel: (9221) 3509114-7, 34532182, 34543280 Fax: (92221) 35009118

Email: kjcis@pafkiet.edu.pk

<http://kjcis.pafkiet.edu.pk>